Received: June 10, 2020. Revised: July 13, 2020.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Outlier Detection Method based on Adaptive Clustering Method and Density Peak

Neelampalli Jayanthi[1]* Burra Vijaya Babu[2] Nandam Sambasiva Rao[3]

[1,2]*Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, India*
[3]*Department of Computer Science and Engineering, Vardhaman College of Engineering Hyderabad, India*
* Corresponding author's Email: jneelampallioffice@gmail.com

Abstract: The outlier detection technique is widely used in the data analysis for the clustering of data. Many techniques have been applied in the outlier detection to increase the efficiency of the data analysis. The Local Projection based Outlier Detection (LPOD) method effectively identifies neighbouring values of data, but this has the drawback of random selection of the cluster centre that affects the overall clustering performance of the system. In this study, the Adaptive Clustering by Fast Search and Find of Density Peak (ACFSFDP) is proposed to select the clustering centre and density peak. This ACFSFDP method is implemented with the min-max algorithm to find the number of categories that measured the local density and distance information. The density and distance are used to select the cluster centre, but density is not calculated on the existing distance based clustering techniques. The ACFSFDP method calculates cluster centre based on the density and distance during the clustering process, whereas the existing techniques randomly select the data centre. The results indicated that the ACFSFDP method is provided effective outlier detection compared with existing Clustering by Fast Search and Find of Density Peak (CFSFDP) methods. The ACFSFDP is tested on two datasets Pen-digits and waveform datasets. The experiment results proved that Area Under Curve (AUC) of the ACFSFDP is 99.08% on the Pen-Digit dataset, while the existing distance classifier method k-Nearest Neighbour has achieved 68.7% of AUC.

Keywords: Adaptive clustering by fast search and find of density peak, Local projection based outlier detection, Outlier detection, Pen-digits dataset, Waveform dataset.

1. Introduction

Detection of outliers plays a major role in various applications and it is considered as one of the key pillars of data mining technology [1]. Nowadays, the researchers have conducted several studies for extending its applications in several areas like clustering, data mining, etc. In a given dataset, the outliers are separated from the normal data points using outlier detection algorithms and these traditional approaches are mainly relying on robust statistics [2]. The nature of the dataset is defined by the outliers, where a certain interesting point in data didn't conform to the natural behaviour of the dataset [3]. The pattern-based or coupling based methods are used as the most unsupervised methods in outlier detection for categorical data. The outline measure is a pattern frequency that is employed by the pattern based methods to search the normal or outlying patterns. [4]. According to the researchers, the most popular techniques are parametric and non-parametric density based methods to detect outliers. Stochastic (generative) density is learned from data by these methods, then the data points with outliers as low probabilities are identified [5]. For instance, anomaly detection problems are solved by developing Gaussian Mixture Models (GMM).

The broad spectrum of application fields and the definitions of precise conceptual complexity have been revealed by several related terminologies, those are deviants, anomalies, discordant, novelties, contaminants, exceptions and abnormalities, where all these terms are used to refer the outliers [6]. In

general, first three methods Local Outlier Factor (LOF), kNN and distribution hypothesis have been developed to identify the degree of isolation or outliers. Moreover, the various clustering technologies have been implemented to identify the outliers, i.e. the points that do not belong to any cluster [7]. The parameters for outlier factors are obtained using LOF method, which is an important outlier detection approach. The identified outlier factors are represented by the degree of abnormality of the model parameters [8]. Many existing methods have been applied to the outlier method to increase the efficiency of the analysis, but these methods have the limitation of selecting the cluster centre (i.e. random selection of the data centres that leads to high computation time) [9, 10]. In this study, the ACFSFDP method is proposed to increase the performance of the outlier detection by proper selection of data based on the min-max algorithm. This algorithm is used to identify the number of categories that are used to measure the distance information. Local density and distance information are measured to select the cluster centre and perform the clustering process. The existing methods have used the cluster centre based on random selection of clusters during clustering process that leads to false positive cluster. The proposed ACFSFDP method uses the density based cluster centre that leads to adapt the neighbourhood size and data instances for cluster analysis. The results proved that the ACFSFDP method has achieved higher performance compared to the existing methods kNN and CFSFDP in outlier detection due to density based cluster centre.

The paper is organised as the literature review is presented in the Section 2, the explanation on the proposed methodology is presented in the Section 3, experimental result is discussed in the Section 4 and conclusion is drawn in the Section 5.

2. Literature works

M. Bai, X. Wang, J. Xin, and G. Wang [11] designed a Distributed Local outlier factor Computing (DLC) and Grid-Based Partition algorithm as (GBP) to detect outliers. This method processed the large datasets with less processing time compared to existing techniques. In large-scale data, this method consumed less number of resources. However, the network resources of existing techniques were highly overhead, which indicates that the performance of existing techniques is limited. In addition, while computing the distance between two tuples, the dimensionality increases in this developed method that lead to high consumption of time and reduction in the performance.

R. J. Campello, D. Moulavi, A. Zimek and J. Sander [12] identified the presence of outliers by implementing the Global-Local Outlier Scores from Hierarchies (GLOSH). A complete density-based clustering hierarchy was developed to find solutions for an infinite range of density thresholds. This made the GLOSH as robust as other existing techniques like LOF, kNN, local distance-based outlier detection (LDOF) approach and Local Outlier Probabilities (LoOP) in the outlier detection process. However, the results indicate that this method is not capable of providing better density estimations that lead to poor computing estimations.

B. Tang and H. He [13] developed an outlier detection method based on local kernel density estimation (KDE). The object's local outlier was measured using KDE with density distribution at the location of the object. The KDE method considered the reverse nearest and shared nearest neighbour of an object for density distribution estimation. The KDE method was tested on both synthetic and real-time datasets and results showed that the KDE method achieved higher performance. However, the KDE method works effectively only on standard dataset and unable to adopt in real-life applications.

M. A. Rahman, K. L. M. Ang and K. P. Seng [14] developed parameter based on the two neighbouring clustering method of independent density. A unique closest neighbourhood and unique neighbourhood set were the two methods used in clustering method. The first method used the dataset with no explicit outliers for simulation, but the other method provided higher performance even when the explicit outlier was present. However, the computational complexity of the developed algorithms reduced due to clusters' arbitrary shape and it is unable to handle complex datasets.

A. Abid, A. Kachouri, and A. Mahfoudhi [15] implemented a Density-Based Spatial Clustering of Applications with Noise Outlier Detection (DBSCANOD) method. The minimum accepted cluster with radius value were computed by this method to make the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) process for data clustering. A decision was made about each point by this method and concluded that the point was either normal or abnormal. However, in the detection of outliers amelioration is highly imposed, especially in the very near outlier points.

H. Liu, X. Li, J. Li, and S. Zhang [16] proposed k of kNN with Local Projection Score (LPS) to analyse the neighbourhood of the data using low-rank approximation. The local neighbourhood information was used to find the data is outlier or not. The experimental result showed that the proposed method

achieved higher performance compared to existing methods. However, the efficiency of the method relies only on kNN and it is affected by distance formulation of kNN.

S. Wang, W. Hua, H. Liu and L. Jiao [17] identified the number of classes for polarimetric synthetic aperture radar (PolSAR) images by implementing an unsupervised classification method. The complementary information from Yamaguchi decomposition was used to cluster the PolSAR images. Next, the appropriate category number was chosen by CFSFDP and PolSAR images were classified by complex K-Wishart function. The three real datasets San Francisco area, Xi'an area and Flevoland area of PolSAR images were selected to test the performance of CFSFDP in terms of time cost and accuracy. However, there was misclassification of some urban area into forest area due to improper cluster formation.

J. He and N. Xiong [18] identified a high density outliers by developing the Decision Graph Based Outlier Detection (DGOD) method. Moreover, the techniques LPS, LOF and Angle-based Outlier Factor (ABOF) were implemented in this study to detect the outliers. For each samples, the score for decision graph was calculated initially and the samples were ranked based on the score values. At last, the outliers were classified by returning the samples with top-k score values. The experiments were conducted on real-world and synthetic datasets to test the efficiency of developed techniques. To get the density information, pairwise distances between all samples needed to identify by DGOD, but these distance computation affected the performance of DGOD method.

From the analysis of existing techniques, it is found out that the major issues are dimensionality increased between two data that leads to high computation time and those techniques are based only on random selection of cluster centre during clustering process. In order to address the issues, the proposed ACFSFDP method uses the average of n_samples by choosing distance cutoff in ACFSFDP (i.e. density estimations during clustering), which is explained in the next section. In addition, the existing techniques CFSFDP [17] and LPS [18] are implemented on the Pen-digits and Waveform dataset to validate the effectiveness of proposed ACFSFDP.

3. Method

In order to analyse the high dimensional data effectively, the outlier detection method is used. LPS provides the effective analysis in the deviation of an observation of its neighbourhood. However,

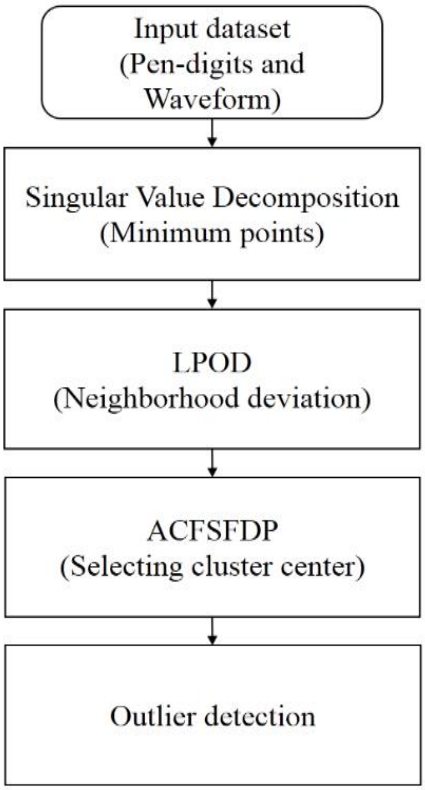


Figure. 1 Proposed ACFSFDP in outlier detection

clustering efficiency has to be improved in the LPS.

In this study, the ACFSFDP is proposed for the outlier detection technique to improve the performance of the data analysis. The proposed ACFSFDP uses the density peak for the cluster analysis and max-min algorithm is used to determine the cluster centre. The Fig. 1 presents the process of proposed ACFSFDP.

3.1 Outlier detection

In linear algebra, the most important and fundamental concept is the matrix rank. The leading entries are referred as matrix rank that corresponds to linearly independent columns or rows of the matrix. The total number of non-zero singular values in the matrix is considered as rank. The columns of a large matrix are arranged as high-dimensional data, i.e. $\in Rn \times m$, where the number of observations and variables (or features) is defined as n and m, respectively. The Eq. (1) presents the decomposition of D by considering the singular value decomposition technique.

$$D = USV^T \tag{1}$$

Where, left and right singular vectors are defined as $U \in R^{n \times r}$ and $V \in R^{m \times r}$. The diagonal matrix is

$S \in R^{r \times r}$, that consists of singular values of D as $S = diag\{\sigma_1, \sigma_2, \ldots, \sigma_r, 0, 0, \ldots, 0\}$ and it is arranged in decreasing order as $\sigma_1 \geq \ldots \geq \sigma_r > 0$. Finally, r is the rank of D (i.e. $ank(D) = r$) and $r \leq min\{n, m\}$.

The sparsity of matrix is calculated by an effective tool called rank, where high sparse of matrix is represented by lower rank. The data are often generated in the real-world applications from low dimensional spaces, where the corresponding data matrices rank are relatively low. But, these data matrices are obtained from the high ranks due to the presence of noises that are raised from different aspects. Therefore, it is important to eliminate the noises from high-dimensional data matrices and achieved the matrices with low rank for analysing the data.

The representation and recover of data within low-dimensional subspaces from high-dimensional spaces are carried out using low-rank approximation, which is a versatile technique. The aim of this technique is to reduce the matrix discrepancy between a high-dimensional data matrix and its reduced matrix, i.e. achieving a low rank matrix $\bar{D}$ for D. In general, the matrix $\bar{D}$ within a low-dimensional subspace is identified using the mathematical model of low-rank approximation, hence the following constraint is reduced in Eq. (2).

$$\min_{\bar{D}} \left|\left|D - \bar{D}\right|\right|_F \quad s.t. rank(\bar{D}) \leq t \tag{2}$$

Where, D is a data matrix and $\bar{D}$ represents the reduced matrix and t denotes the time. The Frobenius norm (F) of X is $\left|\left|X\right|\right|_F = \sqrt{\Sigma_i \Sigma_j x_{ij}^2}$. It is considered and known as a combinatorial NP hard for the above optimization problem. Therefore, constraints are relaxed to make the minimization problem trackable and identify the feasible solution. In Eq. (3), the convex optimization problem is calculated from Eq. (2).

$$\min_{\bar{D}} \left|\left|D - \bar{D}\right|\right|_F \quad s.t. \left|\left|\bar{D}\right|\right|_* \tag{3}$$

Where, nuclear norm or trace norm is represented as $\left|\left|\bar{D}\right|\right|_*$ that can be defined as the summation of top t singular values of $\bar{D}$, i.e. $\left|\left|\bar{D}\right|\right|_* = \Sigma_{i=1}^{t} \sigma_i$. The method selects the top t singular values for minimizing the noise effects that increase the robustness of this research work.

The various effective solutions like augmented Lagrange multipliers, iterative thresholding, alternating direction and accelerated proximal gradient methods are considered for the optimization problem of Eq. (3). Here, nuclear norm minimization problems are efficiently solved by considering the technique of Singular Value Thresholding (SVT). It is important that the Eq. (3) has the same solution for the following optimization problem as present in Eq. (4).

$$\min_{\bar{D}} \frac{1}{2} \left|\left|D - \bar{D}\right|\right|_F^2 + \lambda \left|\left|\bar{D}\right|\right|_* \tag{4}$$

Where, λ denotes the weighted values or threshold values.

3.2 Local projection-based outlier detection

According to the above analysis, an outlier detection method called Local Projection based Outlier Detection (LPOD) is applied, where it consists of two steps; the calculation of LPSs is the first step and the identification of outliers based on scores is the second step. The divergence degree of neighbourhood after projected into low-dimensional subspace is the core idea of LPOD. Initially, the kNN is used to obtain the x neighbors within the former stage and projected into a low dimensional subspace. The process of estimating the anomalous score $lps(x)$ of x is carried out, when the singular values are available.

Within the data collection D, m features are used to represent the number of n observations, where $O(kn^2)$ is the time complexity of conventional kNN algorithm. The kNN efficiency is improved to $O(knlogn)$, when the k-d tree searching technique is considered [16]. The cost for the projection and optimization problems is considered as $O(max(km^2, k^2m))$ time. In general, m is higher than k and $O(max(kn^3, knm^2))$ is the time complexity of the proposed method. When compared with other outlier detection algorithms, LPOD finishes quickly in this experiment.

The appropriate values are assigned using two parameters in LPOD algorithm, where the number of desirable nearest neighbour is defined as the first parameters, i.e. $k\ (1 \leq k \leq n)$. Suppose if the value of k is too low, then kNN is very sensitive to noise. The process of identifying the outliers is difficult, because the probability density function of $lps(x)$ is flat, when k is high. The cross validation is used as an empirical solution to identify the value of k [16]. In the next section, the simulation results are

presented, which shows that assigning a proper value are ranging from five to ten for k.

3.3 Adaptive clustering by fast search and find of density peaks

Two indexes distance as d_i and local density as ρ_i must be computed for each data point i with higher density. Consider $S = \{x_1, x_2, \dots, x_n\}$ is the dataset that needs to be clustered, $I_s = \{1,2, \dots, n\}$ is the corresponding indicator set and the $d_{ij} = dist(x_i x_j)$ is the distance between two points such as x_i and x_j. In the dataset S, two qualities include distanced as δ_i and local density should be computed for every point x_i.

There are two kinds of methods Cut-off kernel and Gaussian kernel those are used to identify the local density as ρ_i. The mathematical expression of these models is presented in Eqs. (5) and (6).

Cut-off kernel:

$$\rho_i = \Sigma_j x\left(d_{ij} - d_c\right) \tag{5}$$

if $x < 0$ otherwise, specific the cut-off distance as d_c.

Gaussian kernel:

$$\rho_i = \Sigma_j e^{-\left(\frac{d_{ij}}{d_c}\right)^2} \tag{6}$$

From the above two equations, it is clear that the discrete value is represented by Cut-off kernel and continuous value is represented by Gaussian kernel. But, there are conflict occur between probability values, i.e. same local density values are presented in various data points. Hence, local density values are identified using the Gaussian kernel method in this study.

It is important to identify the number of categories of the data set, so this study uses Max-Min algorithm. According to the Euclidean distance, Max-Min algorithm has a specific implementation steps that are represented in the following points:

1 The parameter θ is selected, i.e. $0 < \theta < 1$ and the first sampling point is randomly chosen as $Z_1 = x_1$;
2 Identify the new cluster centre as follows:
   A. From all other samples, evaluate the distance D_{i1} to Z_1;
   B. Take the second sample points Z_2 as x_k, if $D_{k1} = \max\{D_{i1}\}$;
   C. From all samples to the sample points Z_1 and Z_2, identify the distance D_{i1} and D_{i2};
   D. Select the x_m as the third sample Z_3; if the distance between Z_1 and Z_2 is $D_m > \theta * D_{12}, D_{12}$, where $D_m = \max\{\min\{D_{i1}, D_{i2}\}\}, i = 1,2,..,n$;
   E. Evaluate $D_j = \max\{\min(D_{i1}, D_{i2}, D_{i3})\}, i = 1,2, \dots n,$ if $D_j > \theta * D_{12}$ and Z_3 exist, then set Z_4 as forth sample point. Repeat the steps until the maximum and minimum distance is not greater than $\theta * D_{12}$ to find the sampling points.
3 The number of categories is also called as the number of sampling points, which is used to represent the output.

1) Cluster center selection

The identification of cluster centres is an important step, where c_i indicates i^{th} data with initial as $c_i = 0$ for all data points. The γ value is obtained using the ρ and δ values, which are used to select the cluster center. Each data point's γ value is used to select the clustering centres, where the following statements explain its process.

1 For every point, identify the distance as δ and the local density as ρ.
2 According to orders of magnitude, the effects of ρ and δ values are eliminated by normalizing these values. Then, the γ value are identified as $\gamma_i = \rho_i \delta_i$;
3 In descending order, the obtained values of γ need to arrange. The cluster centers are selected as the first k data points those have largest γ value. The min-max algorithm is used to obtain the k categories, where the cluster centers corresponds to $c_i = 1,2, \dots k$.

4. Result and discussion

Outlier detection measures the deviation from the normal behaviour and is used in the data analysis and clustering techniques. Many existing methods have been used the outlier detection techniques as the clustering method, but it has the limitation of random selection of cluster centre that affects the performance. The proposed ACFSFDP uses the density peak to select the cluster center based on a min-max algorithm. The two datasets Pen-Digits and waveforms are used for the performance analysis of the proposed ACFSFDP method in outlier detection [19, 20]. The metrics such as Accuracy, precision, recall, F-measure and AUC are used to measure the performance of the proposed method, which is described as follows.

The mathematical Eq. (7) shows the formula of precision, which is used to measure the true positive data, where a portion of positive data is identified by recall for a given cluster is defined in Eq. (8).

$$Precision = \frac{TP}{TP+FP} \times 100 \quad (7)$$

$$Recall = \frac{TP}{TP+FN} \times 100 \quad (8)$$

In order to determine the outliers, the measurement of statistical variability and random errors are used to define the overall accuracy, which is presented in Eq. (9).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \quad (9)$$

where, True Positive is represented as TP, True Negative is presented as TN, False Positive is described as FP and False Negative is depicted as FN in the three Eqs. (7)-(9).

Wherever Times is specified, Times Roman of Times New Roman may be used. If neither is available on your word processor, please use the font closest in appearance to Times. Avoid using bit-mapped fonts if possible. True-Type 1 fonts are preferred.

The mathematical Eq. (10) provides the score calculated by considering the recall and precision of the test and then, F-measure is defined.

$$F - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (10)$$

This study has used the two dataset Pin-digits and waveform [12] for simulations. The Table 1 presents the performance of the proposed ACFSFDP in terms of accuracy, precision, recall and F-score. The four metrics are calculated to analyse the efficiency of the proposed ACFSFDP method. The proposed method achieves more than 98% accuracy in the Pen-digits dataset and more than 75% accuracy in the waveform dataset. The proposed method uses the min-max algorithm to find the number of categories. The local density and distance information are used to select the cluster centre. The proposed method gains significant performance due to this advantage. The ACFSFDP method achieve higher efficiency in the Pen-digits datasets and considerable performance in the waveform dataset. The waveform dataset attributes contain noise and 19 attributes out of 40 attributes are all noise attributes. The ACFSFDP method achieves higher performance in both datasets in terms of the clustering functions.

Table 1. The proposed ACFSFDP method in various metrics

	Adaptive Clustering by fast search and find of density peaks (ACFSFDP)				
Datasets	**Iterations**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1score (%)**
Pen-Digits	K=4	99.79	96.26	96.26	96.26
	K=8	99.83	96.26	96.26	96.26
	K=12	99.77	96.34	96.34	96.34
	K=16	99.80	96.17	96.17	96.17
	K=20	99.75	96.26	96.26	96.26
	K=24	99.80	96.26	96.26	96.26
	K=28	99.74	96.75	96.75	96.75
	K=32	99.75	96.58	96.58	96.58
	K=36	99.78	96.26	96.26	96.26
	K=40	99.79	96.26	96.26	96.26
Waveform	K=4	76.16	58.25	58.25	58.25
	K=8	76.18	58.27	58.27	58.27
	K=12	76.16	58.25	58.25	58.25
	K=16	76.16	58.25	58.25	58.25
	K=20	76.18	58.27	58.27	58.27
	K=24	76.18	58.27	58.27	58.27
	K=28	76.16	58.25	58.25	58.25
	K=32	76.16	58.25	58.25	58.25
	K=36	76.16	58.25	58.25	58.25
	K=40	76.16	58.25	58.25	58.25

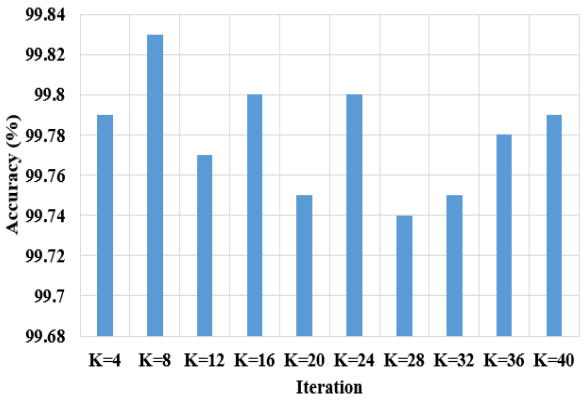


Figure. 2 Accuracy of the proposed method in various iteration

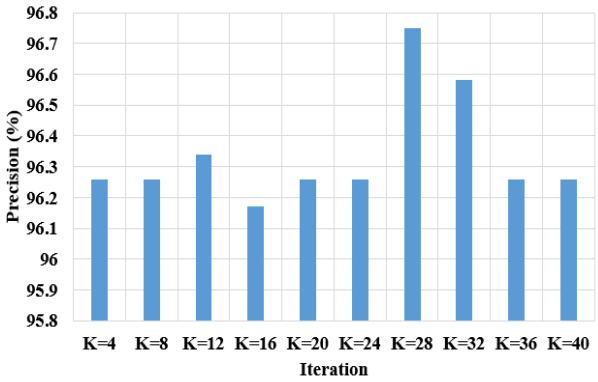


Figure. 3 Precision of proposed method in various iterations

The ACFSFDP method is tested on Pen-Digits datasets and accuracy for various iterations is presented in Fig. 2. The accuracy of ACFSFDP is high for 8 iterations and 24 iterations. In the 40th iteration, the ACFSFDP method is able to analyse all the attributes for the outlier detection. The ACFSFDP method is able to adapt to the data instance due to the selection of cluster centre based on the local density and distance information.

The ACFSFDP method's precision value is measured for various iterations on Pen-Digits datasets, as presented in the Fig. 3. The ACFSFDP method is analysed for 40 iterations and has achieved 96.26 % precision. The results indicate that ACFSFDP method has achieved higher clustering performance.

The AUC metrics provide the capacity of the model to distinguish between the classes in the clustering or classification process. The AUC of the ACFSFDP method for two datasets Pen-Digits and waveforms are presented in the Fig. 4. The proposed ACFSFDP method gain higher AUC for the Pen-digits dataset. For waveform dataset, the ACFSFDP method achieves 80 AUC. The ACFSFDP method has the advantages of using the local density and distance information to select the cluster centre. The AUC is measured in the various Neighbours sizes in the data and presented in the Fig. 5. For various neighbour size, the proposed ACFSFDP method achieves the similar AUC for the dataset.

The AUC is measured for the proposed ACFSFDP method, existing methods LPS [18] and CFSFDP on Pen-digits dataset, as presented in Fig. 6. The AUC of the proposed ACFSFDP is more than 98 and for the LPS method, AUC is around 60. The ACFSFDP method uses the local density and distance information to select the cluster centre that improves the method's efficiency by adaptive to data instances. The LPS method achieves less than 60 AUC for the clustering process. As result shows that the ACFSFDP has higher capacity to distinguish various classes in the data. The AUC is measured for the various neighbour size that shows the proposed method has a similar AUC for the various neighbours' size.

The computational time of the proposed ACFSFDP method and existing LPS [18] and CFSFDP [17] methods are presented in the Table 2. The ACFSFDP method has lower computation time (i.e. 19.56s for Pen-digits and 3.96s for Waveform), where LPS method [18] has the highest computation time (i.e. 36.09s for Pen-Digits and 9.69s for Waveform dataset). The ACFSFDP method has the capacity to adapt to the data and find the cluster centre for clustering process. Therefore, this method is suitable for the real-time application.

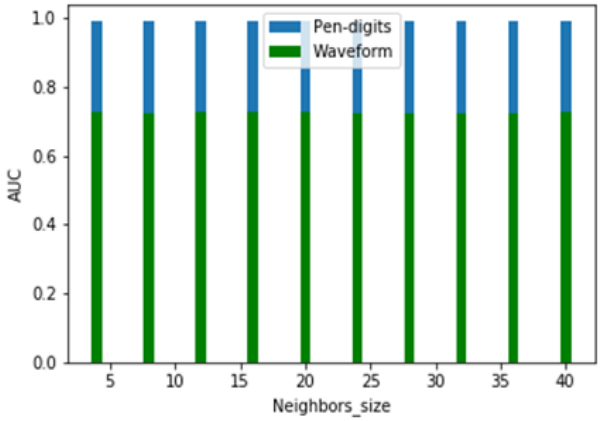


Figure. 4 AUC of the proposed method for two datasets

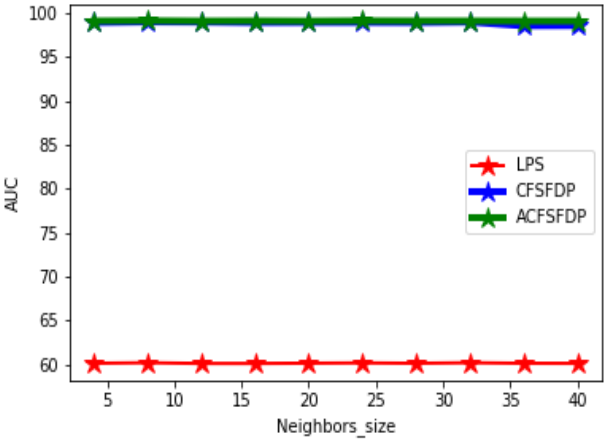


Figure. 5 The AUC of the proposed and existing method in pen-digits dataset

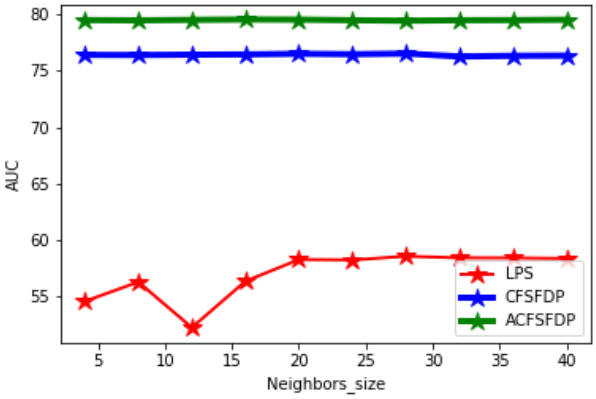


Figure. 6 AUC metrics for the various methods in waveform dataset

Table 2. Computational time of the proposed method

Dataset	Computational time (s)		
	LPS [18]	CFSFDP [17]	ACFSFDP
Waveform	9.69	4.57	3.96
Pen-Digits	36.09	25.16	19.56

The proposed ACFSFDP has the AUC of more than 80, while existing methods LPS has the AUC of

50. The proposed ACFSFDP method has more capacity to distinguish between the different classes of data. For the various neighbours' size, the AUC of the proposed and existing is similar.

4.1 Performance analysis of proposed method with existing techniques on pen-digit and waveform dataset

In this section, the effectiveness of proposed ACFSFDP is compared with the existing technique LPS [18] and traditional CFSFDP method [17]. Table 3 presents the validated results of the proposed ACFSFDP method with LPS and CFSFDP in terms of accuracy, precision, recall of different numbers of iterations.

From the above Table 3, the results point out that the proposed ACFSFDP method has achieved better results in the case of Pen-Digits datasets in terms of accuracy, precision and recall. The existing LPS method achieved only the average accuracy of 90.55% and average precision, recall and f-score of 89.15%, this is because the efficiency depends on the kNN method. But, the ACFSFDP method achieved the average accuracy of 99.78% and average precision, recall and f-score of 96.34%, which is slightly better than traditional CFSFDP method [17]. The reason is that the proposed ACFSFDP uses the density based cluster centre, where the random centre is used by the traditional CFSFDP method [17] during clustering process. Table 4 presents the experimental results of the proposed ACFSFDP method with the existing techniques LPS [18] and CFSFDP [17] on Waveform Dataset in terms of accuracy, precision and recall.

The proposed method has gained low accuracy in the case of the Waveform Dataset, because the outliers are high in this dataset compared to the Pen-Digit Dataset. For instance, the ACFSFDP method achieved the average accuracy of 76.16%, where the LPS method achieved only the average accuracy of 66.90%. In addition, the average recall of proposed ACFSFDP is only 58.25%, where the traditional CFSFDP achieved the average recall is 56.22%. The Fig. 7 shows results of the cluster outlier detection for two datasets.

The performance of the ACFSFDP method is compared with the existing techniques CFSFDP [17] and LPS [18] in terms of F-score and AUC. Table 5 shows the comparative results of proposed ACFSFDP based on Pen-Digits dataset.

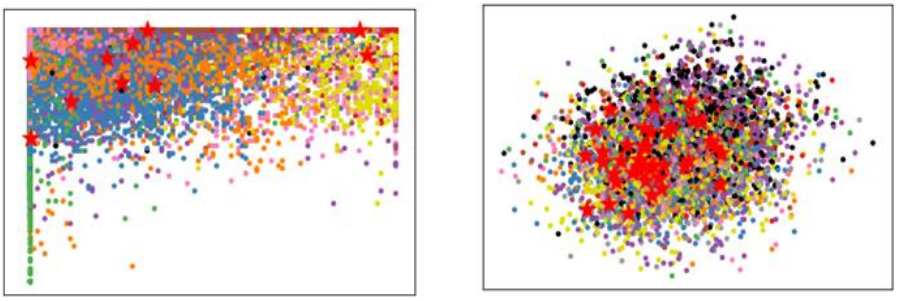
The existing techniques LPS [18] and CFSFDP [17] achieved nearly 60% to 63% of AUC and 88% to 90% of F-score on Pen-Digits dataset, because LPS technique is unable to estimate the outliers effectively. The CFSFDP has been implemented on the standard dataset and identified the outliers and increased the AUC to 98.85%, but it depends on the selection of the random centre in clustering process. In order to address the issues, adaptive technique is included in the existing CFSFDP as ACFSFDP based on density centre and achieved 99.08% of AUC. Likewise, the

Table 3. Analysis of ACFSFDP method on pen-digits dataset in terms of accuracy, precision and recall

Number of Iterations	Pen-Digits								
	Accuracy (%)			Precision (%)			Recall (%)		
	LPS [18]	CFSFDP [17]	ACFSFDP	LPS [18]	CFSFDP [17]	ACFSFDP	LPS [18]	CFSFDP [17]	ACFSFDP
K=4	90.5	96.58	99.79	89.0	95.35	96.26	89.0	95.35	96.26
K=8	90.6	98.89	99.83	89.1	95.48	96.26	89.1	95.48	96.26
K=12	90.5	98.56	99.77	89.1	95.33	96.34	89.1	95.33	96.34
K=16	90.5	98.66	99.80	89.1	9542	96.17	89.1	9542	96.17
K=20	90.5	98.61	99.75	89.1	95.46	96.26	89.1	95.46	96.26
K=24	90.6	98.67	99.80	89.1	95.28	96.26	89.1	95.28	96.26
K=28	90.5	98.69	99.74	89.2	95.34	96.75	89.2	95.34	96.75
K=32	90.5	98.66	99.75	89.2	95.36	96.58	89.2	95.36	96.58
K=36	90.5	98.65	99.78	89.1	95.35	96.26	89.1	95.35	96.26
K=40	90.5	98.69	99.79	89.2	95.35	96.26	89.2	95.35	96.26
Average Values	90.55	98.46	99.78	89.15	95.40	96.34	89.15	95.40	96.34

Table 4. Analysis of ACFSFDP method on waveform dataset by means of accuracy, precision and recall

Number of Iterations	Waveform								
	Accuracy (%)			Precision (%)			Recall (%)		
	LPS [18]	CFSFDP [17]	ACFSFDP	LPS [18]	CFSFDP [17]	ACFSFDP	LPS [18]	CFSFDP [17]	ACFSFDP
K=4	66.9	72.11	76.16	34.6	56.21	58.25	34.6	56.21	58.25
K=8	66.9	72.14	76.18	34.6	56.23	58.27	34.6	56.23	58.27
K=12	66.9	72.15	76.16	34.5	56.23	58.25	34.5	56.23	58.25
K=16	66.9	72.18	76.16	34.6	56.21	58.25	34.6	56.21	58.25
K=20	66.9	72.18	76.18	34.4	56.21	58.27	34.4	56.21	58.27
K=24	66.9	72.15	76.18	34.6	56.24	58.27	34.6	56.24	58.27
K=28	66.9	72.17	76.16	34.6	56.23	58.25	34.6	56.23	58.25
K=32	66.9	72.19	76.16	34.6	56.21	58.25	34.6	56.21	58.25
K=36	66.9	72.18	76.16	34.6	56.22	58.25	34.6	56.22	58.25
K=40	66.9	72.18	76.16	34.6	56.22	58.25	34.6	56.22	58.25
Average Values	66.90	72.16	76.16	34.62	56.22	58.25	34.62	56.22	58.25



(a) (b)

Figure .7 Outlier detection results of two datasets

Table 5. Comparative analysis of proposed method on pen-digits in terms of F-score and AUC analysis

Number of iterations	Pen-Digits					
	F-score (%)			AUC (%)		
	LPS [18]	CFSFDP [17]	ACFSFDP	LPS [18]	CFSFDP [17]	ACFSFDP
K=4	89.0	95.35	96.26	60.12	98.91	99.09
K=8	89.1	95.48	96.26	60.18	98.98	99.11
K=12	89.1	95.33	96.34	60.10	98.96	99.08
K=16	89.1	9542	96.17	60.10	98.91	99.09
K=20	89.1	95.46	96.26	60.13	98.92	99.07
K=24	89.1	95.28	96.26	60.16	98.92	99.10
K=28	89.2	95.34	96.75	60.12	98.91	99.09
K=32	89.2	95.36	96.58	60.18	98.96	99.09
K=36	89.1	95.35	96.26	60.12	98.54	99.08
K=40	89.2	95.35	96.26	60.11	98.56	99.08
Average Values	89.15	95.40	96.34	60.13	98.85	99.08

CFSFDP achieved average F-score of 95.40%, where the ACFSFDP achieved 96.34% of average F-score on Pen-Digits Dataset. Table 6 presents the comparative analysis of ACFSFDP with traditional techniques namely LPS [18] and CFSFDP [17] based on Waveform Dataset.

From the comparative analysis, it is concluded that the ACFSFDP method has achieved better AUC

Table 6. Comparative analysis of proposed method on waveform dataset in terms of F-score and AUC analysis

Number of iterations	Waveform					
	F-score (%)			AUC (%)		
	LPS [18]	CFSFDP [17]	ACFSFDP	LPS [18]	CFSFDP [17]	ACFSFDP
K=4	34.6	56.21	58.25	54.56	76.42	79.51
K=8	34.6	56.23	58.27	56.28	76.41	79.49
K=12	34.5	56.23	58.25	52.24	76.44	79.53
K=16	34.6	56.21	58.25	56.34	76.46	79.56
K=20	34.4	56.21	58.27	58.27	76.52	79.55
K=24	34.6	56.24	58.27	58.23	76.48	79.50
K=28	34.6	56.23	58.25	58.56	76.54	79.45
K=32	34.6	56.21	58.25	58.42	76.28	79.50
K=36	34.6	56.22	58.25	58.41	76.34	79.51
K=40	34.6	56.22	58.25	58.36	76.36	79.54
Average Values	34.62	56.22	58.25	56.96	76.42	79.51

values 79.51% compared to the other existing techniques LPS and CFSFDP on Waveform dataset. Moreover, the average F-score of ACFSFDP achieved only 58.25%, where LPS achieved only 34.62% of average f-score. However, the proposed method achieved poor performance on Waveform than Pen-digit dataset. The reason is that number of outliers' presents in the Waveform dataset. Therefore, the effectiveness of the proposed method on waveform dataset need to improve in terms of accuracy, precision, recall and f-score as a future study.

5. Conclusion

The outlier detection method is used in the data analysis and in the machine learning method. Though many techniques have been proposed on the outlier detection, the effectiveness of the method is still needs to be improved. LPOD method used the kNN and low rank approximation for the clustering purposes. Other existing methods used the random choosing of clusters centre that affected the performance of the clustering technique. In this study, ACFSFDP technique was proposed to select the clustering centre based on the density peak. The ACFSFDP method processed the max-min algorithm based on the number of clusters to select the cluster centre. The experimental result proved that the ACFSFDP method achieved higher performance compared to other existing methods LPS, kNN, GLOSH and CFSFDP. The AUC of the ACFSFDP method was more than 75 and the existing method had less than 60 AUC, because ACFSFDP method used local density and distance information to select cluster centre. The proposed method gained accuracy of 99.79 % on Pen-digits dataset and 76.16% on Waveform dataset. In future work, effective distance metrics will be used to increase the performance of the outlier detection.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

The paper conceptualization, methodology, have been done by 3rd author. The software, validation, formal analysis, investigation, resources, data curation, writing-original draft preparation, writing-review and editing, visualization, have been done by 1st author. The supervision and project administration, have been done by 2nd author.

References

[1] C. Wang, Z. Liu, H. Gao, and Y. Fu, "VOS: A new outlier detection model using virtual graph", *Knowledge-Based Systems,* Vol. 185, pp. 104907, 2019.

[2] J. K. Dutta and B. Banerjee, "Improved outlier detection using sparse coding-based methods", *Pattern Recognition Letters*, Vol. 122, pp. 99-105, 2019.

[3] D. Chakraborty, V. Narayanan, and A. Ghosh, "Integration of deep feature extraction and ensemble learning for outlier detection", *Pattern Recognition*, Vol. 89, pp. 161-171, 2019.

[4] L. Cheng, Y. Wang, and X. Ma, "A Neural Probabilistic outlier detection method for categorical data", *Neurocomputing*, Vol. 365, pp. 325-335, 2019.

[5] H. Estiri and S.N. Murphy, "Semi-supervised encoding for outlier detection in clinical observation data", *Computer Methods and*

Programs in Biomedicine, Vol. 181, pp. 104830, 2019.

[6] R. D. Rodrigues, L. Zhao, Q. Zheng, and J. Zhang, "A tourist walk approach for internal and external outlier detection", *Neurocomputing*, 2019.

[7] Y. F. Wang, Y. Jiong, G. P. Su, and Y. R. Qian, "A new outlier detection method based on OPTICS", *Sustainable Cities and Society*, Vol. 45, pp. 197-212, 2019.

[8] Z. Chen, K. Xu, J. Wei, and G. Dong, "Voltage fault detection for lithium-ion battery pack using local outlier factor", *Measurement*, Vol. 146, pp. 544-556, 2019.

[9] B. Wang, Z. Mao, and K. Huang, "A prediction and outlier detection scheme of molten steel temperature in ladle furnace", *Chemical Engineering Research and Design*, Vol. 138, pp. 229-247, 2018.

[10] L. You, Q. Peng, Z. Xiong, D. He, M. Qiu, and X. Zhang, "Integrating aspect analysis and local outlier factor for intelligent review spam detection", *Future Generation Computer Systems*, Vol. 102, pp. 163-172, 2020.

[11] M. Bai, X. Wang, J. Xin, and G. Wang, "An efficient algorithm for distributed density-based outlier detection on big data", *Neurocomputing*, Vol. 181, pp. 19-28, 2016.

[12] R. J. Campello, D. Moulavi, A. Zimek, and J. Sander, "Hierarchical density estimates for data clustering, visualization and outlier detection", *ACM Transactions on Knowledge Discovery from Data (TKDD)*, Vol. 10, No. 1, pp. 1-51, 2015.

[13] B. Tang and H. He, "A local density-based approach for outlier detection", *Neurocomputing*, Vol. 241, pp. 171-180, 2017.

[14] M. A. Rahman, K. L. M. Ang, and K. P. Seng, "Unique Neighborhood Set Parameter Independent Density-Based Clustering with Outlier Detection", *IEEE Access*, Vol. 6, pp. 44707-44717, 2018.

[15] A. Abid, A. Kachouri, and A. Mahfoudhi, "Outlier detection for wireless sensor networks using density-based clustering approach", *IET Wireless Sensor Systems*, Vol. 7, No. 4, pp. 83-90, 2017.

[16] H. Liu, X. Li, J. Li and S. Zhang, "Efficient outlier detection for high-dimensional data", *IEEE Transactions on Systems, Man and Cybernetics: Systems*, Vol. 48, No. 12, pp. 2451-2461, 2017.

[17] S. Wang, W. Hua, H. Liu, and L. Jiao, "Unsupervised classification for polarimetric SAR images based on the improved CFSFDP algorithm", *International Journal of Remote Sensing*, Vol. 40, No. 8, pp. 3154-3178, 2019.

[18] J. He and N. Xiong, "An effective information detection method for social big data", *Multimedia Tools and Applications*, Vol. 77, No. 9, pp. 11277-11305, 2018.

[19] F. Alimoglu and E. Alpaydin, "Methods of combining multiple classifiers based on different representations for pen-based handwritten digit recognition", In: *Proc. of the Fifth Turkish Artificial Intelligence and Artificial Neural Networks Symposium (TAINN 96)*, 1996.

[20] L. Breiman, J. H. Friedman, A. Olshen, and J. Stone, "Classification and Regression Trees", *CRC press*, 1984.