2021 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT) | 978-1-6654-4293-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/JEEIT53412.2021.9634099

Distance Based Joint Probability Density Estimation For Unsupervised Outlier Detection

Atiq ur Rehman
ICT Division, College of Science and Engineering,
Hamad Bin Khalifa University,
Doha, Qatar.
atrehman2@hbku.edu.qa.

Samir Brahim Belhaouari
ICT Division, College of Science and Engineering,
Hamad Bin Khalifa University,
Doha, Qatar.
sbelhaouari@hbku.edu.qa

***Abstract*— Outlier detection is a vital preprocessing step in data mining and it holds a great importance for Machine Learning (ML) algorithms. If a ML model is learned without removing the outliers from the data, the outliers present in the data can influence the prediction accuracy of a ML model and the outcome of such a model can be misleading. Keeping in view the importance of outliers detection, this paper proposes an unsupervised outlier detection mechanism. The proposed outlier detection mechanism is based on the Joint Probability Density Estimation (JPDE) with an integration of a Distance Measure (DM). The proposed approach has an advantage of utilizing only a single dimensional distance vector to compute the outliers in a dataset. This enables the proposed algorithm to find the outliers from a high dimensional dataset with low computational complexity. Furthermore, three different approaches based on JPDE-DM are proposed and evaluated using some complex benchmark synthetic datasets.**

***Keywords*—Anomaly detection, Data mining, Joint Probability Density Estimation, Outlier detection, Unsupervised learning.**

I. Introduction

Outliers are data points that have characteristics that differ from the rest of the data. Due to the fact that the outliers can sometimes provide critical insights regarding the data and also the fact that their existence can degrade the prediction accuracy of Machine Learning (ML) models, the outlier detection techniques are widely being used in different application domains [1]–[4].

A straightforward approach to develop an outlier detection technique is to create a model for all the data and to distinguish the outliers from the inliers based on deviation from the normal profiles. The already developed methods can be broadly classified as: (i) statistical approaches [5], [6], (ii) clustering approaches [7] (iii) regression approaches [8] (iv) information theoretic approaches [9], [10], (v) proximity based approaches [11], and (vi) ensemble approaches [12], among others. However, the selection of appropriate outlier detection method is difficult due to the unknown data distribution in real scenarios. Although, the non-parametric techniques are able to perform well without a priori knowledge regarding the data distribution but these techniques require a lot of data and resources.

Proximity based outlier detection methods do not require any training or any assumptions regarding the data. However, with increasing data dimensions the effectiveness of these methods is decreased. In order to solve the aforementioned issues, this paper proposes an unsupervised outlier detection method. The proposed method is based on the proximity based statistical outlier detection concept, where the problem of high data dimensionality is resolved by converting the data into a single dimension distance vector. The Joint Probability Density Estimation (JPDE) of the distance vector is computed and some novel parameters for the estimation of outliers in a dataset are proposed.

Three different approaches are proposed to identify the outliers and the proposed approaches are evaluated using some complex synthetic benchmark datasets. The datasets used for evaluation are contaminated with different noise distributions to create a complex structure of the data. The proposed approaches are found effective to identify the outliers in an unsupervised manner.

The rest of the article is organized as: Section II (Proposed Approaches), Section III (Results and Discussion) and Section (IV) Conclusion.

II. Proposed Methodology

The proposed outlier detection method computes a single dimensional distance vector d_k from the actual multi-dimensional data. The distance vector is a one-dimensional vector containing the distance between each observation and its nearest k^{th} neighbor. The transformation of multi-dimensional data in a single dimensional distance vector is represented as:

$$d_k\colon \mathbb{R}^N \to \mathbb{R} \qquad (1)$$

The computed distance vector d_k is used to estimate the joint distribution function's parameters which are further utilized to detect the outliers. Three different unsupervised statistical outlier detection approaches are presented here which are as follows:

Approach 1:

The real valued random variables with unknown distributions are often estimated by a Normal distribution [13], [14] and an independent and identically normal distributed function's joint probability density is given as:

$$f(x_1, \dots, x_N) = \frac{1}{\left(\sigma\sqrt{2\pi}\right)^N} e^{\sum_{i=1}^{N} -\frac{1}{2}\left(\frac{x_i-\mu_i}{\sigma}\right)^2} \qquad (2)$$

where, μ represents the random variable's mean, N is dimension size of the data and σ is the standard deviation. In the proposed approach σ is modeled differently based on the distance vector d_k. Assume that $D(x, y)$ is a two-dimensional dataset, the outliers in this data can be identified by defining a normal distribution-based separation threshold T, such that:

$$\begin{cases} Z = \sum_{i=1}^{n} f(x_i, y_i)\,, \\ T = \alpha\, max(Z). \end{cases} \tag{3}$$

where, Z represents the joint probability distribution function, n represents the total number of observations, α is the significance value, which may be used to regulate the percentage of data that should be discarded as outliers and $f(x, y)$ is defined as:

$$f(x,y) = \frac{1}{2\pi\zeta^{I}} e^{-\left(\frac{(x-x_i)^2+(y-y_i)^2}{2\zeta^2}\right)}; i = 1,2,\ldots,n.; I = 0,1,2. \tag{4}$$

where, ζ is computed as:

$$\zeta = \beta Q3_{d_k}. \tag{5}$$

where, $Q3_{d_k}$ is the third quartile of the distance vector d_k and β is a user defined constant value. The points under the defined threshold T in (3) are marked as outliers and the points those lie above the defined T are considered the normal data points.

Approach 2:

To further weaken the position of outliers in the form of amplitude and support of the function, a better function $f(x, y)$ can be constructed wherein it is easy to detect the outliers. As a result, a new method is developed to detect outliers using the same threshold as used before in (3), by using the following function:

$$f(x,y) = \frac{\zeta^2}{\pi} e^{-\left(\frac{(x-x_i)^2+(y-y_i)^2}{\zeta^2}\right)}; i = 1,2,\ldots,N. \tag{6}$$

where, ζ is computed as:

$$\zeta = \frac{\gamma}{(1+d_k)^2} \tag{7}$$

where, γ is a user-controlled constant that adjusts the gaussian distribution's smoothness. The notion is illustrated in Fig 1, where (a) depicts the standard gaussian approach's impact on compression and (b) demonstrates how suggested approach 2 affects compression.

The two approaches defined above are based on a single distribution and are expected to perform well on the datasets for which a single gaussian distribution may be estimated. However, if a dataset can be approximated well by using multiple gaussian distribution, the above defined approaches will fail to detect the outliers. Therefore, a better idea for such datasets is to utilize a model based on multiple gaussian distribution. A useful concept using multiple gaussian distribution is given below:

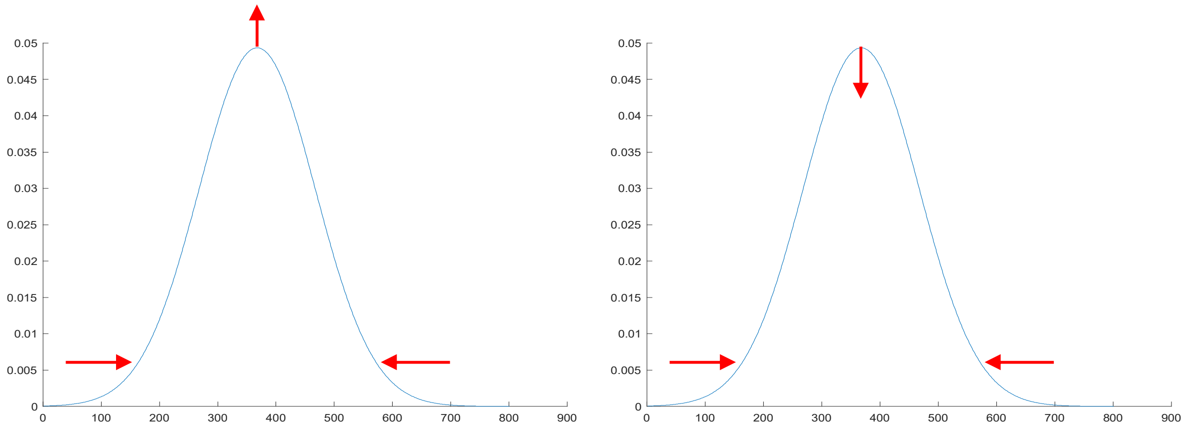
Approach 3:

Outliers are the points that fall on the extreme tails of the distribution, as shown in Fig. 2 (a), in cases when a single gaussian distribution may be used to approximate a dataset. However, if a finer approximation of a dataset is possible using multiple gaussian distribution, the outlier points located at the intersection of two gaussians are missed out if the same phenomenon of outlier detection is used as used in a single distribution scenario. Therefore, in order to detect the outliers for multiple gaussian distribution, a Rejection Area (RA) is required to incorporate the intersection to different gaussians. This will detect the outliers located at the extreme tails and the intersection of different gaussians both. The RA for multiple gaussian distribution is computed and defined as:

$$\begin{cases} RA = \{\vec{x} : f(\vec{x}) \leq Cv\}, \\ f(x \in RA) = \tau. \end{cases} \tag{8}$$

where, τ is the significance level and the RA is the area that lies below a critical value Cv which is a threshold to detect the outliers. The same concept of outlier detection using multiple gaussians is explained in Fig 2 (b), where Cv marks the threshold below which the points are classified as outliers, and a single dimensional data is calculated using two gaussians.

The sorted values of the vector $d_k$can be used to identify the optimum number of gaussians that best approximate the joint probability distribution for a particular dataset. For example, the graph of sorted



(a)

(b)

Fig 1. (a) Compression effect of the conventional gaussian method. (b) Compression in the x and y axes as a result of proposed approach 2.

values of the vector $d_k$is given in Fig 3, and the optimal value of number of gaussians may be determined by choosing the value where the curve takes off sharply.
Each calculated gaussian refers to a region, and the mean and variance values for each gaussian within a region may be calculated as follows:

$$\mu_i = \frac{\sum_{i \in R_j} x_i}{n_i}, j = 1, 2, \dots., m \tag{9}$$

$$Var(x) = \frac{\sum_{i \in R_j} (x_i - \mu_i)^2}{n_i - 1}, j = 1, 2, \dots, m \tag{10}$$

where R_j is the j^{th} region, the number of gaussians estimated is given by m, and n_i is the number of data points in a region, respectively. The estimation of combined multiple gaussians is done as:

$$x \sim \sum_{i=1}^{m} \alpha_i N(\mu_i, C_i) \tag{11}$$

where,

$$\alpha_i = \frac{Card(R_i)}{N} \text{ and } \sum \alpha_i = 1. \tag{12}$$

To find the regions, lets create an application S_U that sorts each given sequence $U_i, i = 1, \dots., N$ in the order $U_{S_U(1)} \leq U_{S_U(2)} \leq \cdots \leq U_{S_U(N)}$. The sorted data may be written as $\vec{X}_{S_U(i)}$, for each given data $\vec{X}$, and suppose that $\overrightarrow{\Delta X}_{S_U(i)}$ denotes the difference between two successive entries of $\vec{X}_{S_U(i)}$. Similarly, $\overrightarrow{\Delta X}_{S_{\Delta X}(S_U(i))}$ may be used to indicate the sorted difference. To create the regions, the entries are successively grouped together until $\Delta X_{S_{\Delta X}(S_U(i))} \leq \overrightarrow{\Delta X}_{S_{\Delta X}(S_U(N-m+1))}$; Once the stated condition is no longer true, begin grouping the rest of entries into new regions until all of the entries have been allocated to one.

In order to make the proposed approach computationally efficient, the given sequence U_i is taken as the distance vector $\boldsymbol{d_k}$ rather than the actual data. Since only a 1-dimensional distance vector is used instead of all the dimensions in the real data, the computational cost is reduced. As an example, the estimation of multiple gaussians using the proposed approach is demonstrated in Fig. 4, where the ground truth data (red color) is estimated by the proposed model (blue color) with two gaussians.

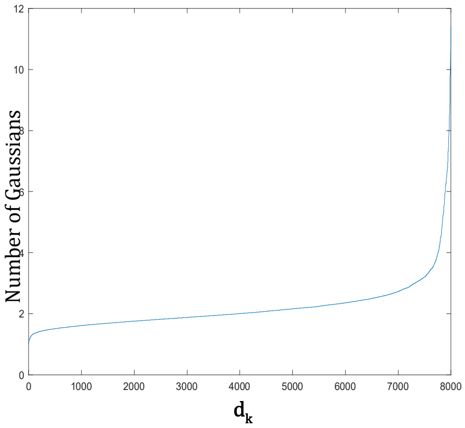


Fig 3. The sorted values of the vector $\boldsymbol{d_k}$ are shown.

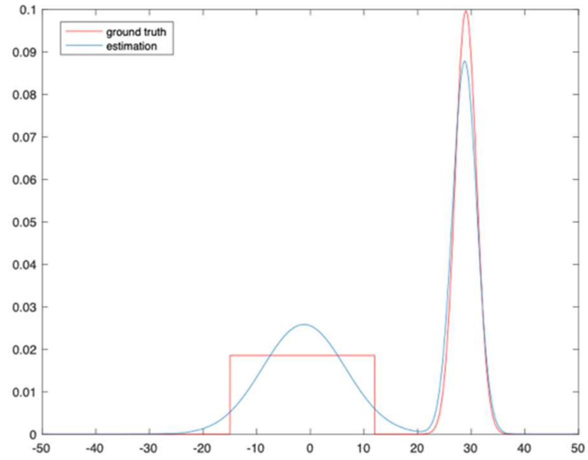


Fig. 4: Estimation using the proposed approach 3 with two gaussians.

```
Pseudocode for Approach 3:
Input: Data $D \in \mathbb{R}^N$, $k, \tau$.
Step 1: Compute the distance vector $d_k: \mathbb{R}^N \rightarrow \mathbb{R}$ by considering $k$ number of neighbors.
Step 2: Sort $d_k$ and estimate the optimum number of gaussians $O_g$ by considering a sharp rise in $d_k$.
```

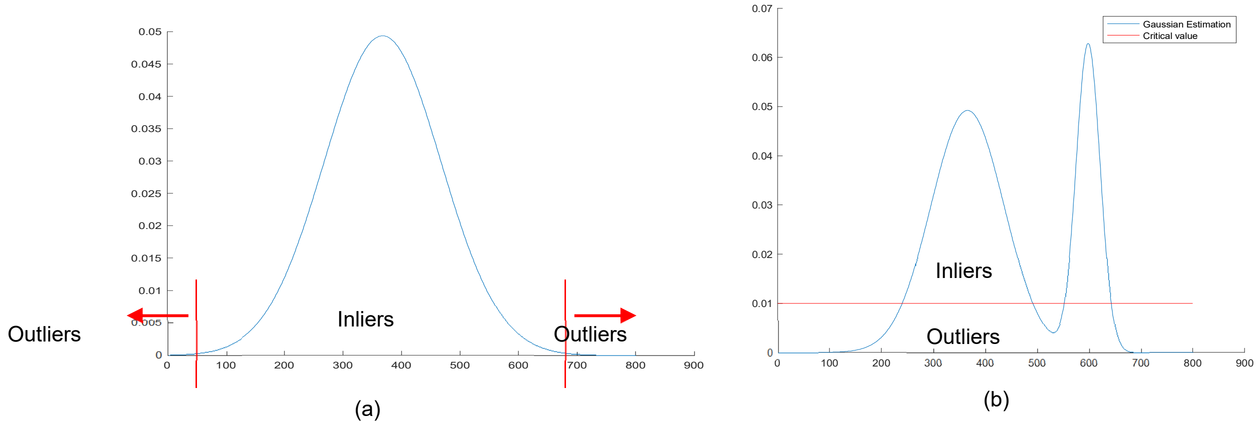
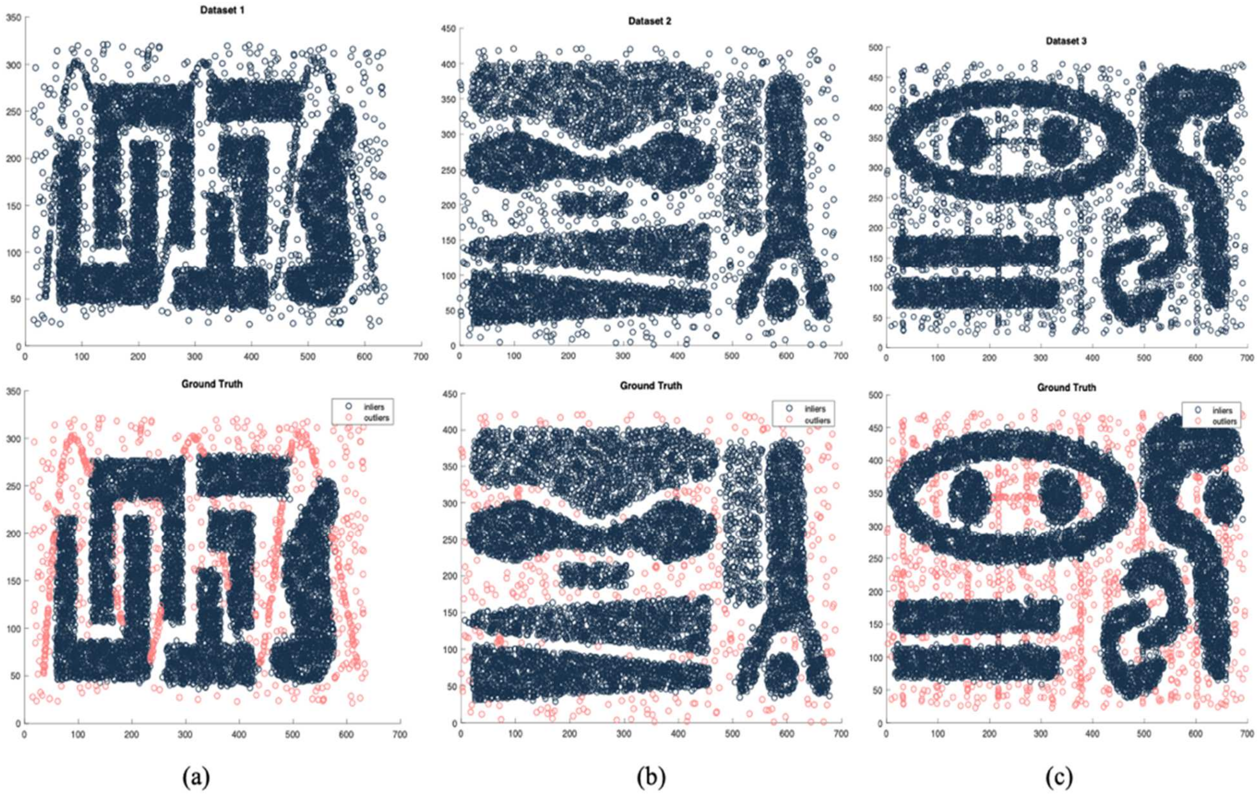


Fig 2. A Gaussian estimate example with crucial value labeling for outlier identification. (a) Outliers are points that fall outside of the red borders. (b) Outliers are defined as points below the crucial value.

Step 3: Compute the difference vector ∇d_k of d_k which contains the differences of two consecutive elements of the sorted d_k.
Step 4: Define the regions sequentially using ∇d_k by grouping the elements until $\nabla d_k(i) \leq \nabla d_k(N - m + 1)$, where m is the number of elements in a region. Once the stated condition is no longer true, begin grouping the rest of entries into new regions until all of the entries have been allocated to one.
Step 5: Compute the combined probability density estimation of regions found in Step 4 using O_g.
Step 6: Compute the Critical Value $Cv = \tau \times max(pdf)$.
Step 7: Mark points below Cv as outliers.

III. Results and Discussion

The performance evaluation of the proposed JPDE-DM approaches is done using some complex synthetic benchmark datasets. The datasets used for evaluation are heavily contaminated with different noise distributions which makes it difficult for the outlier detection methods to differentiate between the inliers and the outliers. The proposed approaches are able to identify both, the single outlying data points as well as outlying noisy clusters. The Receiver Operating Characteristics (ROC) are a performance statistic that is calculated as the Area Under Curve (AUC) to evaluate the performance of proposed approaches [15]. The datasets used for evaluation with their ground truth values can be visually seen in Fig. 5. The computed ROC-AUC values using the three different approaches to detect the outliers from these datasets are given in Table I. The visual outlier detection results for the proposed approaches reported in Table I are given in Fig 6, Fig 7 and Fig 8 for approach 1 (dataset 1), approach 2 (dataset 2) and approach 3 (dataset 3), respectively.



(a) (b) (c)

Fig. 5: Synthetic datasets and ground truth values used for evaluation. (a) Dataset 1, (b) Dataset 2 and (c) Dataset 3.

TABLE I
AUC Values Using Different Parameters For All the Proposed Approaches

	Approach 1	
Dataset	$\alpha = 0.1, \beta = 3, k = 1, I = 1$	$\alpha = 0.3, \beta = 3, k = 1, I = 1$
1	0.6439	0.9332
2	0.7971	0.9528
3	0.7341	0.9829
	Approach 2	
Dataset	$\alpha = 0.1, \gamma = 2, k = 1$	$\alpha = 0.25, \gamma = 5, k = 1$
1	0.8181	0.9323
2	0.9191	0.9456
3	0.9122	0.9711
	Approach 3	
Dataset	$\alpha = 0.4, k = 18, G = 6$	$\alpha = 0.4, k = 18, G = 10$
1	0.8616	0.8686
2	0.9287	0.9305
3	0.9635	0.9623

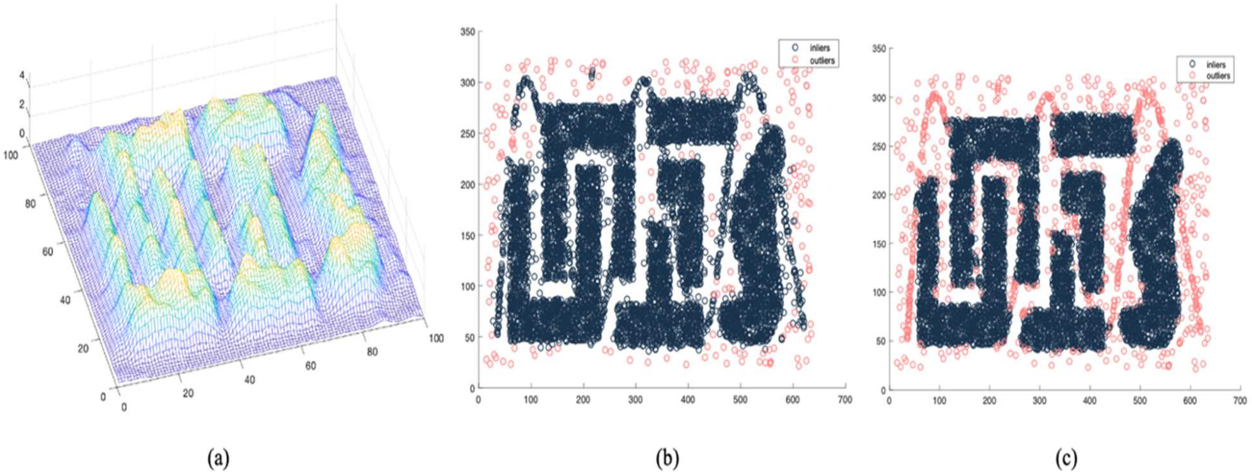


Fig. 6: Outlier detection results for dataset 1 using approach 1. (a) joint PDF estimation. (b) Results using $\alpha = 0.1, \beta = 3, k = 1, I = 1$ (c) Results using $\alpha = 0.3, \beta = 3, k = 1, I = 1$.

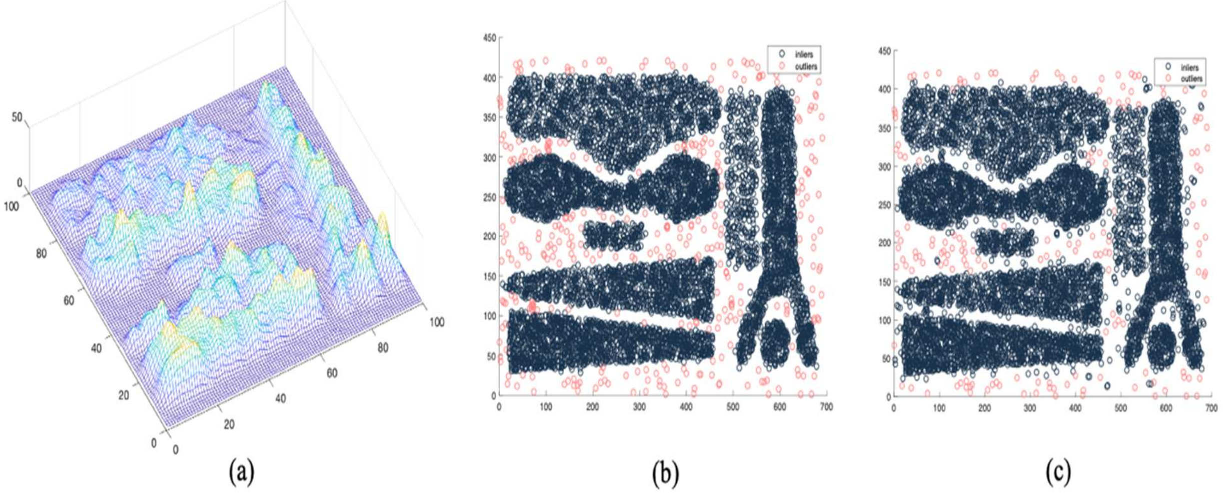


Fig. 7: Outlier detection results for dataset 2 using approach 2. (a) joint PDF estimation. (b) Results using $\alpha = 0.1, \gamma = 2, k = 1$ (c) Results using $\alpha = 0.01, \gamma = 5, k = 1$.

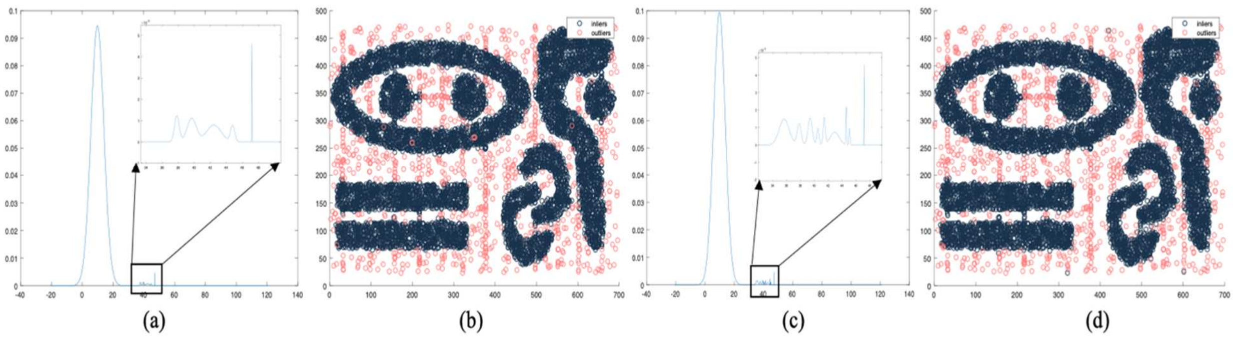


Fig. 8: Outlier detection results for dataset 3 using approach 3. (a) Multiple gaussian estimation with $G = 6, \alpha = 0.4, k = 15$. (b) Results for parameters given in (a), (c) Multiple gaussian estimation with $G = 10, \alpha = 0.4, k = 15$. (d) Results for parameters given in (c).

TABLE II
COMPARISON BASED ON AUC WITH EXISTING STATE-OF-THE-ART APPROACHES

	State-of -the-Art					**Proposed**		
Dataset	KNN	ABOD	FastABOD	COF	LOF	Approach 1	Approach 2	Approach 3
1	0.6222	0.7692	0.7431	0.9081	0.9240	**0.9655**	0.9524	0.9574
2	0.7281	0.8903	0.8951	0.8839	0.9435	0.9025	0.9030	**0.9520**
3	0.5527	0.8526	0.8368	0.9047	0.9527	**0.9829**	0.9779	0.9799

The best results achieved using the three proposed approaches are compared with some of the existing state-of-the-art methods and the results are reported in Table II. The methods used for comparison include kNN [16], Local Outlier Factor (LOF) [17], Connectivity based Outlier Factor (COF) [18] Angle-Based Outlier Detection (ABOD) and Fast Angle-Based Outlier Detection (FastABOD) [19]. It can be seen from the comparison given in Table II that the proposed approaches have performed better than the existing approaches to identify outliers with different distributions. A more comprehensive performance analysis of the unsupervised outlier detection schemes is provided in [20].

IV. Conclusion

Three different unsupervised outlier detection approaches based on the JPDE using a distance vector are proposed in this paper. For the first two approaches, the distance vector is used to estimate the parameters of the JPDE function. Whereas, for the third approach the distance vector itself is utilized to identify the outliers from the underlying data. The three proposed unsupervised outlier detection approaches are found efficient in detecting both, the single outlying data points as well as outlying noisy clusters. The usefulness of the proposed approaches is demonstrated by evaluating some complex synthetic benchmark datasets and the results are compared with some existing state-of-the-art approaches. The proposed approaches are found better in terms of the AUC values which is a benchmark evaluation metric for outlier detection algorithms. The future work includes the expansion of the proposed algorithm to higher dimensional datasets and to the test the performance of proposed approaches on some real datasets.

References

[1] A. Boukerche, L. Zheng, and O. Alfandi, "Outlier Detection: Methods, Models, and Classification," *ACM Comput. Surv.*, vol. 53, no. 3, pp. 1–37, 2020.

[2] H. Wang, M. J. Bah, and M. Hammad, "Progress in Outlier Detection Techniques: A Survey," *IEEE Access*, vol. 7, pp. 107964–108000, 2019.

[3] Y. Liu *et al.*, "Generative Adversarial Active Learning for Unsupervised Outlier Detection," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 8, pp. 1517–1528, 2020.

[4] A. U. Rehman, S. B. Belhaouari, M. Ijaz, A. Bermak, and M. Hamdi, "Multi-Classifier Tree with Transient Features for Drift Compensation in Electronic Nose," *IEEE Sens. J.*, vol. 21, no. 5, pp. 6564–6574, 2021.

[5] C. Barreyre, L. Boussouf, B. Cabon, B. Laurent, and J.-M. Loubes, "Statistical Methods for Outlier Detection in Space Telemetries," *Sp. Oper. Inspiring Humankind's Futur.*, no. Springer, Cham, pp. 513–547, 2019.

[6] P. D. Domański, "Study on Statistical Outlier Detection and Labelling," *Int. J. Autom. Comput.*, 2020.

[7] G. Mishra, S. Agarwal, P. K. Jain, and R. Pamula, "Outlier Detection Using Subset Formation of Clustering Based Method," in *Advances in Intelligent Systems and Computing*, 2019, vol. 870, pp. 521–528.

[8] J. Fan, Q. Zhang, J. Zhu, M. Zhang, Z. Yang, and H. Cao, "Robust deep auto-encoding Gaussian process regression for unsupervised anomaly detection," *Neurocomputing*, vol. 376, pp. 180–190, 2020.

[9] Y. Dong, S. B. Hopkins, and J. Li, "Quantum entropy scoring for fast robust mean estimation and improved outlier detection," in *Advances in Neural Information Processing Systems*, 2019, vol. 32, pp. 6067–6077.

[10] R. Lehmann and M. Lösler, "Multiple Outlier Detection: Hypothesis Tests versus Model Selection by Information Criteria," *J. Surv. Eng.*, vol. 142, no. 4, p. 04016017, 2016.

[11] M. Radovanović, A. Nanopoulos, and M. Ivanović, "Reverse nearest neighbors in unsupervised distance-based outlier detection," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 5, pp. 1369–1382, 2015.

[12] J. Chen, S. Sathe, C. Aggarwal, and D. Turaga, "Outlier detection with autoencoder ensembles," in *Proceedings of the 17th SIAM International Conference on Data Mining, SDM 2017*, 2017, pp. 90–98.

[13] N. Distribution, "Encyclopedia.com: https://www.encyclopedia.com/social-sciences/applied-and-social-sciences-magazines/distribution-normal," *Gale Encyclopedia of Psychology*. .

[14] G. Casella and R. L. Berger, "Statistical Inference," *(2nd ed.). Duxbury. ISBN 978-0-534-24312-8.*, 2001.

[15] G. O. Campos *et al.*, "On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study," *Data Minning Knowl. Discov.*, vol. 30, pp. 891–927, 2016.

[16] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," in *In: Proceedings of the ACM international conference on management of data (SIGMOD), Dallas*, 2000, pp. 427–438.

[17] M. Breunig, H. Kriegel, R. Ng, and J. Sander, "LOF: identifying density-based local outliers," in *In: Proceedings of the ACM international conference on management of data (SIGMOD), Dallas*, 2000, pp. 93–104.

[18] J. Tang, Z. Chen, A. Fu, and D. Cheung, "Enhancing effectiveness of outlier detections for low density patterns," in *In: Proceedings of the 6th Pacific-Asia conference on knowledge discovery and data mining (PAKDD), Taipei*, 2002, pp. 535–548.

[19] H. Kriegel, M. Schubert, and A. Zimek, "Angle-based outlier detection in high-dimensional data," in *In: Proceedings of the 14th ACM international conference on knowledge discovery and data mining (SIGKDD), Las Vegas*, 2008, pp. 444–452.

[20] A. ur Rehman and S. B. Belhaouari, "Unsupervised outlier detection in multidimensional data," *J. Big Data*, vol. 8, no. 1, 2021.