International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
IJISAE
ISSN:2147-6799 www.ijisae.org Original Research Paper

ULODF: An Unsupervised Learning based Outlier Detection Framework in High Dimensional Data

C. Jayaramulu [1], Bondu Venkateswarlu [2]

Submitted: 04/02/2024 **Revised:** 12/03/2024 **Accepted:** 18/03/2024

Abstract: Outliers play crucial role in applications like disease diagnosis, fraud detection techniques and cyber security to mention few. Unsupervised learning techniques like clustering are widely used, in the area of machine learning, towards outlier detection. However, most of the existing methods did not consider dual tasking benefits of using clustering that not only renders quality clusters but also identifies outliers effectively. We proposed a framework named Unsupervised Learning based Outlier Detection Framework (UL-ODF). An algorithm named Novel Outlier Detection Method in High Dimensional Data (NODM-HDD) is defined. The algorithm has mechanisms to improve compactness of clusters made besides determining outliers. The algorithm exploits an enhanced version of K-Means clustering technique. A prototype is built to validate the utility of the framework and the underlying algorithm. Different benchmark datasets and metrics are used in the empirical study. The experimental results revealed that the NODM-HDD shows better performance over the state of the art.

Keywords *–Outlier Detection, Unsupervised Learning, Outlier Detection Framework, Clustering High Dimensional Data*

1. Introduction

Outliers play crucial role in applications like disease diagnosis, fraud detection techniques and cyber security to mention few. Unsupervised learning techniques like clustering are widely used, in the area of machine learning, towards outlier detection. However, most of the existing methods did not consider dual tasking benefits of using clustering that not only renders quality clusters but also identifies outliers effectively. ML based outlier detection methods such as [1], [2], [3], [4], [5], [6], [7], [8], [9], [10] are found in the literature. However, they used different application domains such as traffic, networks, Wireless Sensor Network (WSN), Internet of Things (IoT) etc. Evolutionary approaches for outlier detection are investigated in [3]. Generative Adversarial Network (GAN) is used for outlier detection as studied in [11] and [12]. Ensemble approaches for improving accuracy are found in [13] and [14]. Clustering based approaches are discussed in[15], [16] [17] and [18],in [17] Shared Neighbor based clustering on text is discussed, and in [18] C-means a soft computing based clustering algorithm on images is applied. From the literature, it is understood that there are various approaches for outlier detection. However, we believe that unsupervised approaches with optimization provide better performance for outlier detection. We also found that Holoentropy metric based approach helps in detecting outliers while performing clustering to have dual benefits. In this paper we proposed an approach that exploits clustering in a novel way. A framework known as Unsupervised Learning based Outlier Detection Framework (UL-ODF) is proposed. An algorithm named Novel Outlier Detection Method in High Dimensional Data (NODM-HDD) is defined. The algorithm has mechanisms to improve compactness of clusters made besides determining outliers. The algorithm exploits an enhanced version of K-Means clustering technique. A prototype is built to validate the utility of the proposed framework and the underlying algorithm. Different benchmark datasets and metrics are used in the empirical study. The experimental results revealed that the NODM-HDD shows better performance over the state of the art. Our contributions are as follows.

[1]*Research Scholar, Dayananda Sagar University, Bangalore, India*
[2]*Associate Professor, Department of CSE Dayananda Sagar University, Bangalore,India*
1jayaramuluc.res-soe-cse@dsu.edu.in
2bonduvenkat-cse@dsu.edu.in

1. A framework known as Unsupervised Learning based Outlier Detection Framework (UL-ODF) is proposed and implemented.

2. An algorithm named Novel Outlier Detection Method in High Dimensional Data (NODM-HDD) is defined.
3. A prototype is built to validate the utility of the proposed framework and the underlying algorithm.

Other sections in the paper are as follows. Section 2 make a review different methods of outlier detection that provides required gaps on the research. Section 3 presents the proposed framework while section 4 provides evaluation methodology. Section 5 provides results of empirical study while Section 6 concludes the work.

2. Related Work

This section reviews literature on various methods of outlier detection. Dwivedi *et al*. [1] explored WSNfor possibilities in outlier detection application. They proposed outlier detection method in WSN based on MLapproaches like Bayesian Belief Network. However, their method depends on training samples for ground truth. Jiang *et al*. [2] also used ML techniques for outlier detection but focused on Internet of Things (IoT) use cases. Deng *et al*. [3] used one class Support Vector Machine (SVM) along with Genetic Algorithm (GA)for detecting outliers. Liu *et al*. [11] explored GAN based method for outlier detection where generator and discriminator components play a non-cooperative game for efficient detection of outliers. Rayana *et al*. [13] proposed an ensemble method for outlier detection with many base detectors such as kNN. Malini and Pushpa [19] focused on kNN based outlier detection for credit card fraud detection. Liu *et al*. [20] proposed a method known as Local Projection Score (LPS) and used it for outlier detection. Domingues *et al*. [21] reviewed different outlier methods while Nesa *et al*. [22] and Nguyen *et al*. [23] used machine learning techniques for detecting outliers among traffic incidents.

Bondu Venkateswarlu *et al* attempted and used for the algorithm of Clustering is an unattended classification and is a process of partitioning a group data object from one set into several classes. This can be done by applying various equations and steps regarding the distance algorithm, namely the Euclidean Distance [24]. Zhao *et al*. [28] proposed a multi-view outlier detection technique to support computer vision tasks. Chakraborty *et al*. [14] combined ensemble learning and deep learning to have better detection of outliers. Christy *et al*. [15] explored clustering for outlier detection. Souza and Amazonas [26] proposed an outlier detection method in the context of big data processing. Munoz-Organero *et al*. [27] combined outlier detection and deep learning to solve problems associated with traffic in cities. Maniruzzaman *et al*. [28] used outliers for diabetes risk stratification while Ren *et al*. [29] used outlier detection for realising intrusion detection system (IDS). Erkus and Purutçuoglu [30] proposed a method based on frequency domain. Althaf Hussin Basha *et al* proposed and used Partial Least Square approach regression analysis technique for to detect the outliers. They has been used Laser dataset to find out the outliers and also the Mahalanobis distance, Jackknife distance and T2 distance were calculated for finding the outliers [31]. Inoue *et al*. [4] focused on finding anomaly detection in water treatment plant using outlier detection. Outliers to detect financial frauds [5], [32], fraud detection in medical care [6], detection of Trojan attacks [33], IDS [34], indoor localization [35], GAN [12], monitoring wind turbine condition [7] and detection of positive active power [38] are contributions found in the literature. Ayesha *et al* [37] reviewed different deep learning models which can also be used for detecting outliers. It is understood that there are various approaches for outlier detection. Furthermore, Unsupervised K-Means technique is used for clustering high dimensional microarray data [51]. However, we believe that unsupervised approaches with optimization provide better performance for outlier detection. We also found that Holoentropy metric based approach helps in detecting outliers while performing clustering to have dual benefits. In this paper we proposed an approach that exploits clustering in a novel way.

3. Unsupervised Learning Based Outlier Detection Framework

This section presents the proposed framework for outlier detection and its underlying algorithm and its functionality.

3.1 The Framework

An outlier detection framework named Novel Outlier Detection Method in High Dimensional Data (NODM-HDD) is proposed. Its novelty lies in its underlying mechanisms in dealing with dual tasks of efficient clustering and thereby isolating outliers effectively. The framework results into a set of clusters and some points as outliers that are of much value in deriving business intelligence (BI). The framework is illustrated in Figure 1. It takes high dimensional data as input and results in highly compact clusters and correctly identified outliers. The given data is taken as input and its feature space is extracted. Afterwards, the feature space is divided into many partitions. Here a basic strategy such as K-Means is used for partitioning. Since basic partitions are further processed, simple K-Means is found sufficient. There are benefits in dividing data into initial partitions with corresponding matrix. First, the matrix can show information pertaining to cluster belonging which is crucial for outlier detection context. Second, the binary space in the form of matrix is much easier to detect outliers provided categorical features. As explored in [38] we used Holoentropy metric for outlier detection that is suitable for the work in this paper. Holoentropy is the summation of entropies obtained from all attributes. It is based on information theory and handles well when there is categorical data.

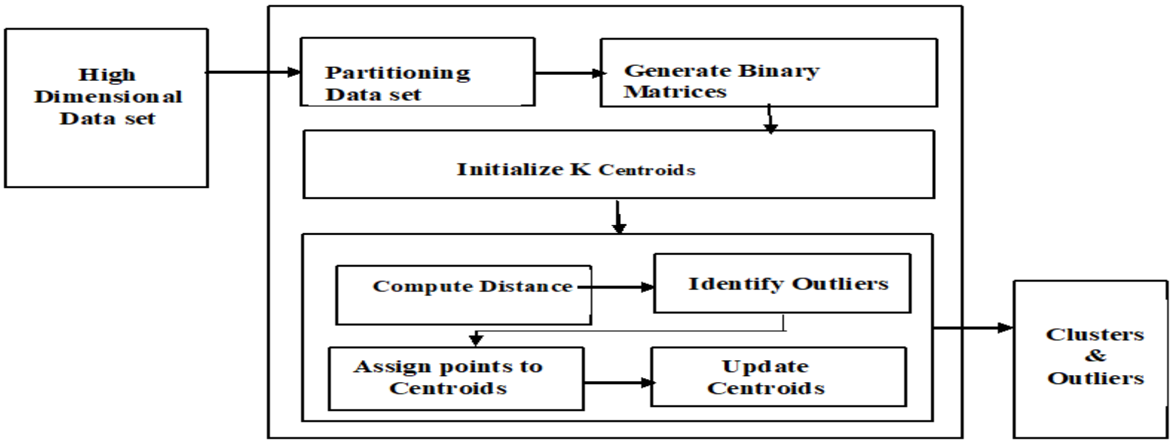

Fig 1: Overview of the proposed outlier detection framework

Two binary matrices are derived from the partitioned data. One is derived initially while the other is the optimized binary matrix. These matrices are used in order to initialize K centroids. After this step, an iterative process is involved in order to compute distance, identify outliers, assign points to centroids and update centroids. This process continues until convergence. Finally, the process results in compact clusters and correctly identified outliers. While discovering points to be included in clusters, the framework simultaneously discovers outliers that are isolated from clusters. Thus the framework reflects a clustering mechanism which is non-exhaustive where some data points are not assigned any cluster labels. Such points are outliers that are used to made well informed decisions in different real world applications.

3.2 NODM-HDD Algorithm

An algorithm named Novel Outlier Detection Method in High Dimensional Data (NODM-HDD). The algorithm has mechanisms to improve compactness of clusters made besides determining outliers. The algorithm exploits an enhanced version of K-Means clustering technique. A prototype is built to validate the utility of the proposed framework and the underlying algorithm. Different benchmark datasets are used in the empirical study. Different metrics are used to evaluate the proposed algorithm. The experimental results showed that the NODM-HDD shows better performance over the state of the art.

Algorithm: Novel Outlier Detection Method in High Dimensional Data
Input: High dimensional dataset D
Output: Clusters C and outliers O

Start
Initialize centroids vector V
P←CreateBasicPartitions(D)
[B1, B2] ← CreateBInaryMatrices(P)

```
V←InitializeCentroids(B1, B2)
For each point p1 in B1
For each point p2 in B2
find distance between p1 and p2
    Find suitable centroid
    Find outliers (where distance is large) //apply Holoentropy metric
    Isolate outlier point
    Update O
    Assign other point to centroid
    Compute arithmetic mean
    Update centroids
    Update C
  End For
End For //convergence
Print C
Print O
Evaluate Performance
End
```

Algorithm 1: Novel Outlier Detection Method in High Dimensional Data

As presented in Algorithm 1, it takes given dataset D as input and generates clusters (C) and outliers (O). Step 3 creates basic partitions from D using simple K-Means. In Step 4, two binary matrices are computed while Step 5 gives initialized centroids. Steps 6 through Step 17, there is an iterative process that helps in creating clusters and isolate outliers. There is an inner iterative process from Step 7 through Step 17. Each point in B1 and B2 are compared in order to compute distance, compute centroids, find outliers using the specified metric, isolate outliers, update outlier vector O, assign non-outliers to centroids, compute arithmetic mean in order to update centroids and clusters C. After convergence, the algorithm returns both set of clusters and set of outliers. Evaluation is made as per the procedure discussed in Section 4 and empirical results are provided in Section 5.

4. Evaluation Methodology

Evaluation of the proposed algorithm is made using different metrics such as Normalized Mutual Information (NMI), normalized rand index, Jaccard, F-measure and execution time. For cluster validity, both NMI and normalized rand index are widely used. The former measures mutual information (MI) that is obtained by comparing ground truth and resultant clusters besides normalizing the outcome. NMI is computed as in Eq. 1.

$$NMI = \frac{\sum_{i,j} n_{ij} \log \frac{n.n_{ij}}{n_{i+}.n_{+j}}}{\sqrt{\left(\sum_i n_i + \log \frac{n_{i+}}{n}\right)\left(\sum_j n_j + \log \frac{n_{j+}}{n}\right)}}, \quad (1)$$

Normalized rand index is used to measure accuracy of clusters made with respect to ground truth. It is computed as in Eq. 2.

$$R_n = \frac{\sum_{i,j}\left(n_{ij_2}\right) - \sum_i (n_{i+_2}).\sum_j\left(n_{+j_2}\right)/(n_2)}{\frac{\sum_i\left(n_{i+_2}\right)}{2} + \sum_j \frac{\left(n_{+j_2}\right)}{2} - \sum_i (n_{i+_2}).\sum_j (n_{i+_2})/(n2)}, \quad (2)$$

Where n_{ij} is known as co-occurrence number while n_{i+} and n_{+j} denote i^{th} cluster size and j^{th} cluster size respectively. The outlier detection performance is evaluated using Jaccard index and F-Measure as expressed in Eq. 3 and Eq. 4 respectively.

$$Jaccard = \frac{|o \cap o^*|}{|o \cup o^*|}, \quad (3)$$

$$F - measute = 2 * \frac{precition.reccall}{precition + recall}, \quad (4)$$

Both O and O* are denoting predicted outliers and the ground truth respectively in Jaccard index while F-measure is the harmonic mean of two measures namely precision and recall.

5. Experimental Results

Experiments are made with different benchmark high-dimensional datasets available. These datasets are obtained from UCI repository. They are also used by the researchers of [39]. In order to evaluate the proposed algorithm thoroughly, different datasets with diversity are used for empirical study. The datasets are diversified in terms of number of instances, type, number of features, number of clusters and number of outliers.

5.1 Datasets

Ecoli dataset is of gene type containing 336 instances, 7 features, 5 clusters and 9 outliers. Yeast dataset is of gene type containing 1484 instances, 8 features, 4 clusters and 185 outliers. Caltech dataset is of image type containing 1415 instances, 4096 features, 4 clusters and 67 outliers. Sun09 dataset is of image type containing 3282 instances, 4096 features, 3 clusters and 50 outliers. Fbis dataset is of text type containing 2463 instances, 2000 features, 10 clusters and 332 outliers. Klb dataset is of text type containing 2340 instances, 21839 features, 5 clusters and 60 outliers.Re0 dataset is of text type containing 1504 instances, 2886 features, 5 clusters and 218 outliers. Re1 dataset is of text type containing 414 instances, 6129 features, 6 clusters and 527 outliers.Tr11 dataset is of text type containing 2463 instances, 2000 features, 4 clusters and 87 outliers. Tr23 dataset is of text type containing 204 instances, 5832 features, 3 clusters and 32 outliers.Wap dataset is of text type containing 1560 instances, 8460 features, 10 clusters and 251 outliers. Glass dataset is of UCI type containing 214 instances, 9 features, 3 clusters and 39 outliers. Shuttle dataset is of UCI type containing 58000 instances, 9 features, 3 clusters and 244 outliers. Kddcup dataset is of UCI type containing 494021 instances, 38 features, 3 clusters and 54499 outliers. Thus there is high diversity among the aforementioned datasets

5.2 Results

The proposed algorithm is evaluated using the datasets aforementioned and the results are compared with different outlier techniques such as K-Means, LOF [40], COF [41], LDOF [42], FABOD [43], iForest [44], OPCA [45], TONMF [46], MICR [47], Linear Regression [48] and K-Means [49,50]. The results are observed in terms of Normalized Mutual Information (NMI), normalized rand index and F-Measure.

Dataset	**Normalized Mutual Information (NMI)**		
	K-Means	**K-Means--**	**NODM-HDD**
Ecoli	65.1151	64.2442	64.98492
Yeast	20.7007	17.3473	62.14208
caltech	79.1291	77.1771	89.81973
sun09	19.8999	12.1822	22.69267
Fbis	12.1922	33.7337	55.03498
k1b	53.003	50.2202	55.20515
re0	20.2202	18.0781	34.91488
re1	19.6797	21.5115	83.77369
tr11	10.3003	21.8618	62.69263
tr23	7.89789	12.6927	26.05603
Wap	43.4034	33.2032	50.83078
Glass	37.2873	37.2973	39.85982
shuttle	23.5736	26.1862	36.18615
kddcup	1.46146	77.2972	86.80672

Table 1: Performance in terms of NMI

As presented in Table 1, the performance of the proposed algorithm named NODM-HDD is compared with that of existing algorithms in terms of NMI against different datasets.

Dataset	**Normalized Rand Index**		
	K-Means	**K-Means--**	**NODM-HDD**

ecoli	68.0335	63.1389	70.63126
yeast	15.1654	13.8213	20.17033
caltech	63.3194	78.4346	89.69829
sun09	18.8765	10.8324	22.2666
fbis	1.13339	12.688	40.80204
k1b	44.122	44.3527	42.13603
re0	11.695	13.3198	25.66677
re1	4.16245	5.4162	23.3699
tr11	0.52156	8.65589	59.6785
tr23	-3.2096	4.34299	22.11615
wap	14.383	12.698	36.74992
glass	23.6006	25.6367	26.65974
shuttle	40.9726	33.5403	60.47087
kddcup	0.04012	81.4536	95.04428

Table 2: Performance in terms of normalized rand index

As presented in Table 2, the performance of the proposed algorithm named NODM-HDD is compared with that of existing algorithms in terms of normalized rand index against different datasets.

Dataset	**Jaccard Measure**		
	K-Means	**K-Means--**	**NODM-HDD**
ecoli	4.37308	58.71562	51.27336
yeast	6.26875	20.58156	52.07576
caltech	19.73904	45.94743	98.87574
sun09	1.93579	3.72113	2.49747
fbis	0.09027	5.37608	26.08803
k1b	0	0	21.41405
re0	5.57668	9.5285	29.7891
re1	0.54162	17.14127	29.60856
tr11	0	10.38105	37.20127
tr23	0	6.91067	15.05503
wap	1.11333	10.32087	23.37993
glass	13.68092	32.37684	35.64662
shuttle	0	5.40617	6.52953
kddcup	0.01003	18.34487	16.65983

Table 3: Performance in terms of Jaccard measure

As presented in Table 3, the performance of the proposed algorithm named NODM-HDD is compared with that of existing algorithms in terms of Jaccard measure against different datasets.

Dataset	**F-Measure**		
	K-Means	**K-Means--**	**NODM-HDD**
ecoli	8.23463	76.40854	67.65235
yeast	11.8254	33.71083	68.56508
caltech	31.5644	64.40263	99.58787

sun09	3.79134	7.17145	4.87458
fbis	0.17051	10.21054	41.4239
k1b	0	0	35.28554
re0	10.5516	17.40205	45.91734
re1	1.09327	29.29763	45.71674
tr11	0	18.81628	54.34254
tr23	0	12.95876	26.26857
wap	2.17651	20.34084	37.9134
glass	22.9888	49.70868	52.37666
shuttle	0	10.25066	12.32687
kddcup	0.02006	31.67474	8.49541

Table 4: Performance in terms of F-Measure

As presented in Table 4, the performance of the proposed algorithm named NODM-HDD is compared with that of existing algorithms in terms of F-Measure against different datasets.

Outlier Detection Method	**Execution Time (seconds)**				
	sun09	**k1b**	**wap**	**shuttle**	**kddcup**
K-means	1.12336	4.56365	1.25375	0.22066	0.62186
LOF	65.3555	150.831	26.8904	11.9658	0
COF	79.7385	154.482	30.2705	181.994	0
LDOF	278.082	2646.89	906.14	247.611	0
FABOD	569.172	5389.88	1816.71	496.916	0
IForest	12.5877	12.9186	8.55559	165.916	1459.78
OPCA	0.4012	6.19854	1.75525	0.3009	2.51753
TONMF	7.89361	31.8553	7.69301	1.18354	18.2245
K-means--	3.57068	65.4758	12.7682	0.33099	5.99794
NODM-HDD	2.31693	0.15045	0.19057	0.57171	2.89867

Table 5: Performance in terms of execution time

As presented in Table 5, the performance of the proposed algorithm named NODM-HDD is compared with that of existing algorithms in terms of execution time against different datasets.

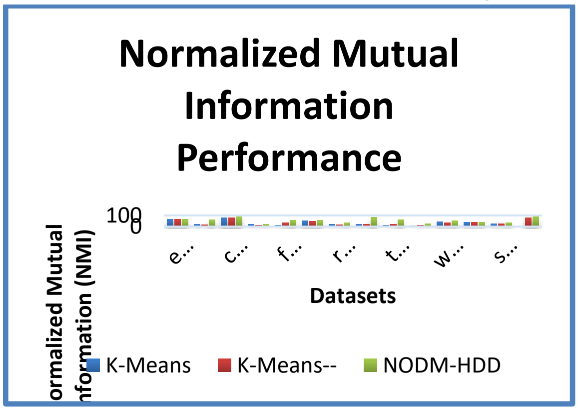

Fig 2: Performance evaluation in terms of NMI against different datasets

As presented in Figure 2, the performance of the proposed algorithm is compared with existing methods against different datasets. The important observation is made with NMI measure. Higher in NMI shows better performance. The horizontal axis shows the benchmark datasets used in empirical study while NMI measure is shown in vertical axis. An important observation is that there is different performance based on the dataset and its characteristics. Another observation is that the NMI performance of the proposed algorithm NOMD-HDD is better than that of existing methods for all the datasets consistently.

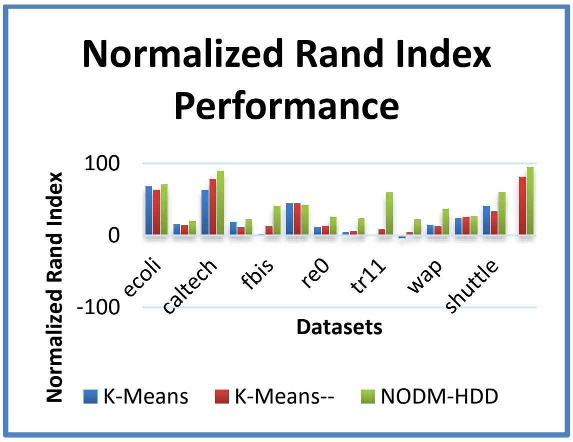

Fig 3: Performance evaluation in terms of normalized rand index against different datasets

As presented in Figure 3, the performance of the proposed algorithm is compared with existing methods against different datasets. The important observation is made with normalized rand index measure. Higher in randomized rand index shows better performance. The horizontal axis shows the benchmark datasets used in empirical study while normalized rand index measure is shown in vertical axis. An important observation is that there is different performance based on the dataset and its characteristics. Another observation is that the performance of the proposed algorithm NOMD-HDD is better than that of existing methods for most of the datasets except Klb.

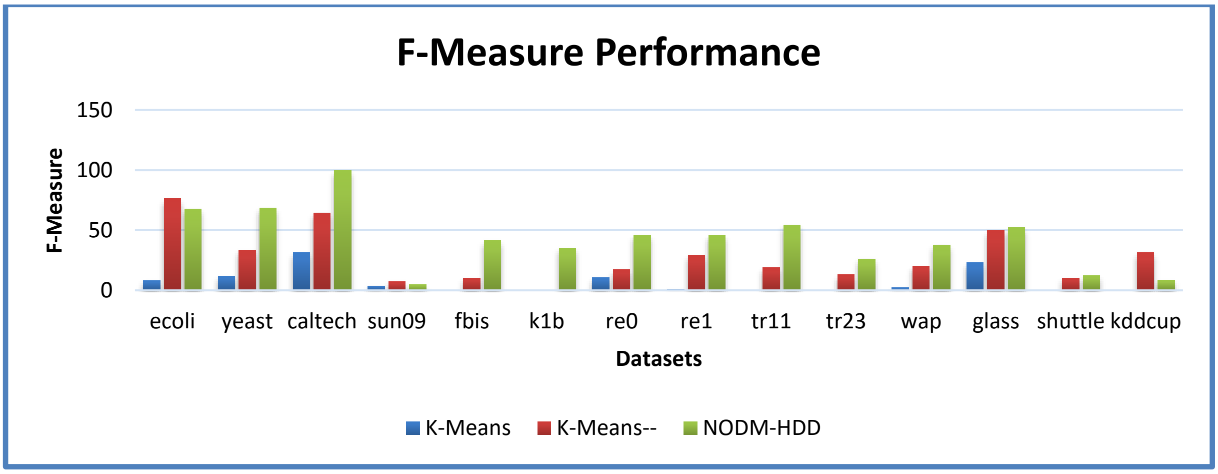

Fig 4: Performance evaluation in terms of Jaccard measure against different datasets

As presented in Figure 4, the performance of the proposed algorithm is compared with existing methods against different datasets. The important observation is made with Jaccard measure. Higher in Jaccard measure shows better performance. The horizontal axis shows the benchmark datasets used in empirical study while Jaccard measure is shown in vertical axis. An important observation is that there is different performance based on the dataset and its characteristics. Another observation is that the performance of the proposed algorithm NOMD-HDD is better than that of existing methods for most of the datasets except Kddcup, Ecoli and Sun09.

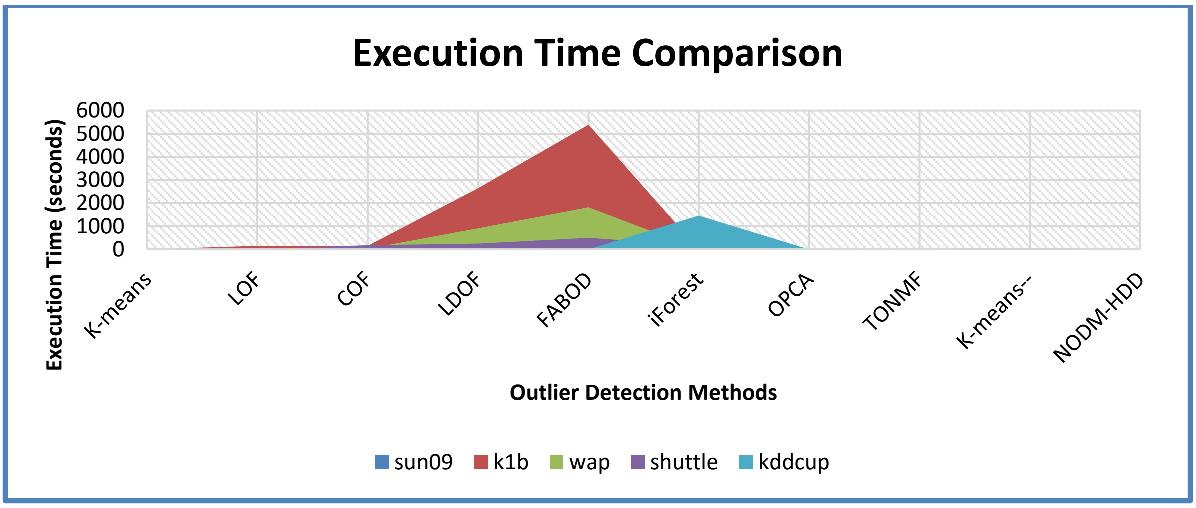

Fig 5: Performance evaluation in terms of F-Measure against different datasets

As presented in Figure 5, the performance of the proposed algorithm is compared with existing methods against different datasets. The important observation is made with F-measure. Higher in F-measure shows better performance. The horizontal axis shows the benchmark datasets used in empirical study while F-measure is shown in vertical axis. An important observation is that there is different performance based on the dataset and its characteristics. Another observation is that the performance of the proposed algorithm NOMD-HDD is better than that of existing methods for most of the datasets except Kddcup, Ecoli and Sun09.

6. Conclusion And Future Work

Outlier detection is an indispensable task that can be reused as part of real world applications. However, most of the existing methods dealt with unsupervised methods for clustering and outlier detection separately. There is need for having an integrated approach that leverages cluster performance and lead to outlier detection. In this paper we proposed a framework known as Unsupervised Learning based Outlier Detection Framework (UL-ODF). An algorithm named Novel Outlier Detection Method in High Dimensional Data (NODM-HDD). The algorithm has mechanisms to improve compactness of clusters made besides determining outliers. The algorithm exploits an enhanced version of K-Means clustering technique. A prototype is built to validate the utility of the proposed framework and the underlying algorithm. Different benchmark datasets are used in the empirical study. Different metrics are used to evaluate the proposed algorithm. The experimental results revealed that the NODM-HDD shows better performance over the state of the art in terms of clustering performance and outlier detection. However, our work is based on only unsupervised learning based approach. It lacks the advantages of a supervised method taking benefits of ground truth from unsupervised method. Therefore, in our future work, we exploit both supervised and unsupervised learning techniques by defining a hybrid algorithm to detect outliers in high dimensional data.

7. Statements And Declarations

Competing Interests and Funding:

The authors did not receive support from any organization for the submitted work.

References

[1] Kumar Dwivedi, R., Pandey, S., & Kumar, R. (2018). A Study on Machine Learning Approaches for Outlier Detection in Wireless Sensor Network. 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence). P1-4.

[2] Jiang, J., Han, G., liu, L., Shu, L., &Guizani, M. (2020). Outlier Detection Approaches Based on Machine Learning in the Internet-of-Things. IEEE Wireless Communications, 27(3), 53–59.

[3] Deng, X., Jiang, P., Peng, X., &Mi, C. (2018). An Intelligent Outlier Detection Method with One Class Support Tucker Machine and Genetic Algorithm

towards Big Sensor Data in Internet of Things. IEEE Transactions on Industrial Electronics, 1–11.

[4] Inoue, J., Yamagata, Y., Chen, Y., Poskitt, C. M., & Sun, J. (2017). Anomaly Detection for a Water Treatment System Using Unsupervised Machine Learning. 2017 IEEE International Conference on Data Mining Workshops (ICDMW). P1-8.

[5] SADGALI, I., SAEL, N., & BENABBOU, F. (2019). Performance of machine learning techniques in the detection of financial frauds. Procedia Computer Science, 148, 45–54.

[6] Bauder, R. A., &Khoshgoftaar, T. M. (2017). Medicare Fraud Detection Using Machine Learning Methods. 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). P1-8.

[7] Stetco, A., Dinmohammadi, F., Zhao, X., Robu, V., Flynn, D., Barnes, M., … Nenadic, G. (2018). Machine learning methods for wind turbine condition monitoring: A review. Renewable Energy. P1-23.

[8] Althaf Hussain Basha, Y. Sri Lalitha "Student Performance Prediction – A Data Science Approach", Modern

\ Approaches in Machine Learning and Cognitive Science: A Walkthrough Studies in Computational Intelligence,

P.115-125, 2021.

[9] Bondu Venkateswarlu, GSV Prasada Raju, "Organ Donor Identification Through Improved K-Medoids Clustering", InternatIo Journal of Computer Science and technology (IJCST),Volume 5,Issue 3,pp.175-177, ISSN:0976-8491,2014.

[10] Sk Althaf Hussain Basha, Ch. Prakash, D. Mounika, G. Maheetha, "An Approach for Multi Instance Clustering of Student Academic Performance in Education Domain", IIJDWM Journal, Volume 3,Issue 1,pp.1-9,Feb.2013, ISSN: 2249-7161

[11] Liu, Y., Li, Z., Zhou, C., Jiang, Y., Sun, J., Wang, M., & He, X. (2019). Generative Adversarial Active Learning for Unsupervised Outlier Detection. IEEE Transactions on Knowledge and Data Engineering, 1–12.

[12] Moustafa, N., Choo, K.-K. R., Radwan, I., &Camtepe, S. (2019). Outlier Dirichlet Mixture Mechanism: Adversarial Statistical Learning for Anomaly Detection in the Fog. IEEE Transactions on Information Forensics and Security, 1–13.

[13] Rayana, S., Zhong, W., &Akoglu, L. (2016). Sequential

Ensemble Learning for Outlier Detection: A Bias-Variance Perspective. 2016 IEEE 16th International Conference on Data Mining (ICDM). P1-6.

[14] Chakraborty, D., Narayanan, V., & Ghosh, A. (2019). Integration of Deep Feature Extraction and Ensemble Learning for Outlier Detection. Pattern Recognition. P1-13.

[15] Christy, A., Gandhi, G. M., &Vaithyasubramanian, S. (2015). Cluster Based Outlier Detection Algorithm for Healthcare Data. Procedia Computer Science, 50, 209–215.

[16] Angelin, B., &Geetha, A. (2020). Outlier Detection using Clustering Techniques – K-means and K-median. 2020 4th InternationalConference on Intelligent Computing and Control Systems (ICICCS).p1-6.

[17] Sri Lalitha Y et al (2014) "Semantic Framework for Text Clustering with Neighbors" in ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India-Vol. II Advances in Intelligent Systems and Computing, P.261-271.

[18] Y. Sri Lalitha, et al (2020) "Efficient Tumor Detection in MRI Brain Images",in International Journal of Online and Biomedical Engineering, P. 122-131.

[19] Malini, N., & Pushpa, M. (2017). Analysis on credit card fraud identification techniques based on KNN and outlier detection. 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB). P1-4.

[20] Liu, H., Li, X., Li, J., & Zhang, S. (2017). Efficient Outlier Detection for High-Dimensional Data. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 1–11.

[21] Domingues, R., Filippone, M., Michiardi, P., &Zouaoui, J. (2018). A comparative evaluation of outlier detection algorithms: Experiments and analyses. Pattern Recognition, 74, 406–421.

[22] Nesa, N., Ghosh, T., & Banerjee, I. (2018). Non-parametric sequence-based learning approach for outlier detection in IoT. Future Generation Computer Systems, 82, 412–421.

[23] Nguyen, H., Cai, C., & Chen, F. (2017). Automatic classification of traffic incident's severity using machine learning approaches . IET Intelligent Transport Systems, 11(10), 615–623.

[25] Zhao, H., Liu, H., Ding, Z., & Fu, Y. (2018). Consensus Regularized Multi-View Outlier Detection. IEEE Transactions on Image Processing, 27(1), 236–248.

[24] Bondu Venkateswarlu and Prof G.S.V.Prasad Raju. 2013. "Mine Blood Donors Information through Improved K-Means Clustering." International Journal of Computational Science and Information Technology (IJCSITY) Vol.1,No.3, arXivpreprint arXiv:1309.2597

[26] Souza, A. M. C., & Amazonas, J. R. A. (2015). An Outlier Detect Algorithm using Big Data Processing and Internet of Things Architecture. Procedia Computer Science, 52, 1010–1015.

[27] Munoz-Organero, M., Ruiz-Blaquez, R., & Sánchez-Fernández, L. (2018). Automatic detection of traffic lights, street crossings and urban roundabouts combining outlier detection and deep learning classification techniques based on GPS traces while driving. Computers, Environment and Urban Systems, 68, 1–8.

[28] Maniruzzaman, M., Rahman, M. J., Al-MehediHasan, M., Suri, H. S., Abedin, M. M., El-Baz, A., & Suri, J. S. (2018). Accurate Diabetes Risk Stratification Using Machine Learning: Role of Missing Value and Outliers. Journal of Medical Systems, 42(5). P1-17.

[29] Ren, J., Guo, J., Qian, W., Yuan, H., Hao, X., &Jingjing, H. (2019). Building an Effective Intrusion Detection System by Using Hybrid Data Optimization Based on Machine Learning Algorithms. Security and Communication Networks, 2019, 1–11.

[30] Erkuş, E. C., &Purutçuoğlu, V. (2020). Outlier Detection and Quasi-periodicity Optimization Algorithm: Frequency Domain Based Outlier Detection (FOD). European Journal of Operational Research. P1-19.

[31] Sk Althaf Hussain Basha, Naga Raju Devarakonda, Shaik Subhani, " Outliers Detection in Regression Analysis using Partial Least Square Approach", ICT and Critical Infrastructure: proceedings of the 48th Annual Convention of Computer Society of India- Springer, Vol II Advances in Intelligent Systems and Computing, Volume 249, pp. 125-135, Visakhapatnam, December 2013,ISBN: 978-3-319-03095-1.

[32] Varmedja, D., Karanovic, M., Sladojevic, S., Arsenovic, M., &Anderla, A. (2019). Credit Card Fraud Detection - Machine Learning methods. 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH).

[33] Davaslioglu, K., &Sagduyu, Y. E. (2019). Trojan Attacks on Wireless Signal Classification with Adversarial Machine Learning. 2019 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN). P1-6.

[34] Sandosh, S., Govindasamy, V., &Akila, G. (2020). Enhanced intrusion detection system via agent clustering and classification based on outlier detection. Peer-to-Peer Networking and Applications. P1-8.

[35] Bhatti, M. A., Riaz, R., Rizvi, S. S., Shokat, S., Riaz, F., & Kwon, S. J. (2020). Outlier detection in indoor localization and Internet of Things (IoT) using machine learning. Journal of Communications and Networks, 22(3), 236–243.

[36] Ma, B., Yuan, L., Xu, S., Zheng, K., Huang, F., Li, R., & Yuan, P. (2020). Positive Active Power Outlier Detection based on One-Class SVM. 2020 12th IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC). P1-4.

[37] Ayesha M, S.K. Althaf Hussain Basha, S. V. Raju, Y. S. Lalitha: A Brief Research on Deep Learning Models," International Journal of Computer Engineering and Applications, Volume XIII, Issue VI, November. 20, ISSN 2321-3469, pp. 1-7.

[38] S. Wu and S. Wang, "Information-theoretic outlier detection for large scale categorical data," IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 3, pp. 589–602, 2013.

[39] H. Liu, M. Shao, S. Li, and Y. Fu, "Infinite ensemble clustering," Data Mining and Knowledge Discovery, no. 1-32, 2017.

[32] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in ACM sigmod record, vol. 29, no. 2, 2000, pp. 93–104.

[40] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, Jörg Sander, "LOF: Identifying Density-Based Local Outliers", Proc. ACM SIGMOD 2000 Int. Conf. On Management of Data, Dalles, TX, pp.1-12,2000

[41] J. Tang, Z. Chen, A. W.-C. Fu, and D. W. Cheung, "Enhancing effectiveness of outlier detections for low density patterns," in Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2002.

[42] K. Zhang, M. Hutter, and H. Jin, "A new local distance-based outlier detection approach for scattered real-world data," in Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2009.

[43] Hans-Peter Kriegel, Matthias Schubert, and Arthur Zimek, "Angle-based outlier detection in high-dimensional data", In Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 444–452, 2008..

[44] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in Proceedings of IEEE International Conference on Data Mining, 2008.
[45] Y. Lee, Y. Yeh, and Y. Wang, "Anomaly detection via online over sampling principal component analysis," IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 7, pp. 1460–1470, 2013.
[46] R. Kannan, H. Woo, C. C. Aggarwal, and H. Park, "Outlier detection for text data," in Proceedings of SIAM International Conference on Data Mining, 2017.
[47] Sk. Althaf, H. Basha, A. Govardan, S. V. Raju and N. Sultana, "MICR: Multiple Instance Cluster Regression for Student Academic Performance in Higher Education", International Journal of Computer Applications (0975–8887), vol. 14, no. 4, January 2011.
[48] Basha Althaf. H., Govardhan, A., Raju, S. V., & Sultana, N, " A Comparative Analysis of Prediction Techniques for Predicting Graduate Rate of University", European Journal of Scientific Research, 46(2), 186-193,2010.
[49] S. Chawla and A. Gionis, "k-means-: A unified approach to clustering and outlier detection," in Proceedings of SIAM International Conference on Data Mining, 2013.
[50] Althaf H. B., Ramesh S. K., Kumar, Y.R., Govardhan A. & Mohd. Z. A., "Predicting Student Academic Performance Using Temporal Association Mining", International Journal of Information Science and Education, Vol. 2, No. 1, pp. 21-41, 2012.
[51] Samson Anosh Babu P, Chandra Sekhara Rao, Annavarapu, Suresh Dara, "Clustering-based hybrid feature selection approach for high dimensional microarray data" Chemometrics and Intelligent Laboratory Systems,Vol. 213,104305 , 2021,