e-ISSN:2582-7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 6, Issue 3, March 2023

Impact Factor: 7.54

6381 907 438 6381 907 438 ijmrset@gmail.com www.ijmrset.com

| DOI:10.15680/IJMRSET.2023.0603008 |

Automatic Detection Outliers from High Dimensional Data Using Unsupervised Learning Framework

R Krishna Chaitanya[1], Dr. Jaideep Gera[2]

PG Scholar, Dept. of CSE, ST Marys Group of Institutions Guntur, AP, Guntur, India[1]

Associate Professor, Dept. of CSE, ST Marys Group of Institutions Guntur, AP, Guntur, India[2]

ABSTRACT: Anomaly discovery is a fashion for chancing an unusual point or pattern in a given set. The term anomaly is also appertained to as outlier. Before, the data mining experimenters were concentrated on other ways like bracket and clustering. Outlier are set up as a part of data sanctification process. Still, view passed a change in 2000 when experimenters set up discovery of abnormal effects can help working the real-world problems seen in damage discovery, fraud discovery, discovery of abnormal health condition and intrusion discovery. There are three kinds of anomalies which are appertained to viz., point anomaly, contextual anomaly, and collaborative anomalies. However, it's called a point anomaly, If a single case in a given dataset is different from others with respect to its attributes. However, it's called contextual anomaly, If the data is anomalous in some environment. We proposed a frame named Unsupervised Learning grounded 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 ameliorate conciseness of clusters made besides determining outliers. The algorithm exploits an enhanced interpretation of K-Means clustering fashion. A prototype is erected to validate the mileage of the frame and the underpinning algorithm. Different standard datasets and criteria are used in the empirical study. The experimental results revealed that the NODM- HDD shows better performance over the state of the art.

I. INTRODUCTION

Outliers play pivotal part in operations like complaint opinion, fraud discovery ways and cyber security to mention many. Unsupervised literacy ways like clustering are extensively used, in the area of machine literacy, towards outlier discovery. Still, utmost of the being styles didn't consider binary tasking benefits of using clustering that not only renders quality clusters but also identifies outliers effectively. ML grounded outlier discovery styles similar asare set up in the literature. Still, they used different operation disciplines similar as business, networks, Wireless Sensor Network (WSN), Internet of effects(IoT) etc. Evolutionary approaches for outlier discovery are delved in. Generative Adversarial Network(GAN) is used for outlier discovery as studied. Ensemble approaches for perfecting delicacy are set up. Clustering grounded approaches are bandied in. From the literature, it's understood that there are colorful approaches for outlier discovery. Still, we believe that unsupervised approaches with optimization give better performance for outlier discovery. We also set up that Holo entropy metric grounded approach helps in detecting outliers while performing clustering to have binary benefits. In this paper we proposed an approach that exploits clustering in a new way. A frame known as unsupervised literacy grounded 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 ameliorate conciseness of clusters made besides determining outliers. The algorithm exploits an enhanced interpretation of K- Means clustering fashion. A prototype is erected to validate the mileage of the proposed frame and the underpinning algorithm. Different standard datasets and criteria are used in the empirical study. The experimental results revealed that the NODM- HDD shows better performance over the state of the art. Our benefactions are as follows.

1. A frame known as unsupervised literacy grounded Outlier Detection Framework(UL- ODF) is proposed and enforced.

2. An algorithm named Novel Outlier Detection Method in High Dimensional Data(NODM- HDD) is defined.

3. A prototype is erected to validate the mileage of the proposed frame and the underpinning algorithm.

Other sections in the paper are as follows. Section 2 make a review different styles of outlier discovery that provides needed gaps on the exploration. Section 3 presents the proposed frame while section 4 provides evaluation methodology. Section 5 provides results of empirical study while Section 6 concludes the work. Anomaly discovery algorithms of low dimensional data aren't suitable for high dimensional data. However, high dimensional dataset has d> 10 attributes, If d is dimension or trait of a dataset. High dimensional data deteriorates as a result of “dimensionality curses”. In general, the outlier or anomaly can be set up using distance grounded or viscosity grounded algorithms. These algorithms measure the distance between data cases and on the supposition that abnormal data point will be down from other data points where anomalies are set up. Still, in the case of high dimensional situation, the data becomes meager and all the data points look normal. On the base of the algorithms can be classified as supervised, semi supervised and unsupervised. When data markers for both normal and anomalous are known, they're distributed as supervised. However, it's appertained to as semi supervised algorithms, If only the data marker of normal is known. However, it's unsupervised algorithm, if data markers of both normal and anomalous are unknown. Numerous real world operations don't contain data markers. Manually labeling of data is a precious task.

II. LITERATURE REVIEW

Anomaly detection in high-dimensional data is becoming a fundamental research area that has various applications in the real world. As such, a large body of research has been devoted towards addressing this problem. Nevertheless, most existing surveys focus on the individual aspects of anomaly detection or high dimensionality. For example, Agrawal and Agrawal [2] provide a review of various anomaly detection techniques, with the aim of presenting a basic insight into various approaches for anomaly detection. Akoglu et al. [3] present several real-world applications of graph-based anomaly detection and concluded with open challenges in the field. Chandola et al. [4] present a survey of several anomaly detection techniques for various applications. Hodge and Austin [1] present a survey of outlier detection techniques by comparing techniques’ advantages and disadvantages. Patcha and Park [5] have conducted a comprehensive survey of anomaly detection systems and hybrid intrusion detection systems by identifying open problems and challenges. Jiang et al. [6] present a survey of advanced techniques in detecting suspicious behavior; they also present detection scenarios for various real-world situations. Sorzano et al. [7] categorize dimensionality reduction techniques, along with the underpinning mathematical insights. Various other surveys can also be observed, such as those by Gama et al. [8], Gupta et al. [9], Heydari et al. [10], and Jindal and Liu [11], Pathasarathy [12], Phua et al. [13], Tamboli et al. [14], and Spirin et al. [15], which further highlight the problems either in anomaly detection or in high-dimensional data.

A limited amount of work has been done that emphasizes anomaly detection and high dimensionality problems together, either directly or indirectly. Zimek et al. [16] present a detailed survey of specialized algorithms for anomaly detection in high dimensional numerical data; they also highlight important aspects of the curse of dimensionality. Parsons et al. [17] present a survey of the various subspace clustering algorithms for high-dimensional data and discuss some potential applications in which the algorithms can be used.

III. EXISTING WORK

In existing framework, proposed an exception discovery strategy that is intended to perform on high layered information. High layered information acts novel difficulties such like low thickness of information and expanded computational intricacy. In this strategy endeavors to resolve the two issues. In the first place, they apply PCA to decrease the element of the information. The utilization of PCA considers an uninformed misfortune dimensionality decrease. Then they use KDE to show the thickness circulation of the information and select the top K focuses with the most minimal likelihood as exceptions. KDE is an adaptable assessment procedure that produces regular guess of the basic circulation on the information. The subsequent calculation accomplishes vigorous forecast results and quick execution times. The proposed technique is tried on both manufactured and genuine information. They use F1 as the scoring metric to represent imbalanced class dissemination. Then, at that point, tests on two-layered manufactured information and seven-layered genuine information demonstrate that PKDE is equipped for creating serious outcomes in low layered setting. Further tests on high layered genuine information confirm that PKDE proceeds too or better than other benchmark strategies.

Disadvantages of Existing Framework

- High-layered information presents extraordinary difficulties in exception discovery process. The vast majority of the current calculations neglect to appropriately resolve the issues coming from countless elements.

- Specifically, exception discovery calculations perform inadequately on informational collection of little size with an enormous number of elements.

- The current technique is neglected to address the difficulties of managing high-layered information by extending the first information onto a more modest space and utilizing the intrinsic construction of the information to compute inconsistency scores for every information point.

- Assuming the exceptions are non-haphazardly conveyed, they can diminish ordinariness. It builds the blunder fluctuation and diminishes the force of measurable tests. They can cause inclination or potentially impact gauges. They can likewise affect the fundamental supposition of relapse as well as other measurable models.

IV. PROPOSED SYSTEM

An exception recognition system named Novel Anomaly Discovery Technique in High Layered Information (NODM-HDD) is proposed. Its oddity lies in its fundamental components in managing double errands of productive bunching and accordingly disconnecting anomalies really. The structure results into a bunch of groups and a few focuses as exceptions that are of much worth in determining business knowledge (BI). The structure is shown in engineering outline. It accepts high layered information as info and results in exceptionally smaller bunches and accurately distinguished anomalies. The given information is taken as info and its component space is separated. Subsequently, the element space is separated into many parcels. Here an essential technique, for example, K-Means is utilized for parceling. Since essential parts are additionally handled, basic K-Means is seen as adequate. There are benefits in separating information into starting parts with comparing lattice. To begin with, the network can show data relating to group having a place which is vital for exception location setting. Second, the parallel space as framework is a lot simpler to identify exceptions gave downright elements. As investigated in [12] we involved Holo entropy metric for exception discovery that is reasonable for the work in this paper. Holo entropy is the summation of entropies acquired from all ascribes. It depends on data hypothesis and handles well when there is downright information.

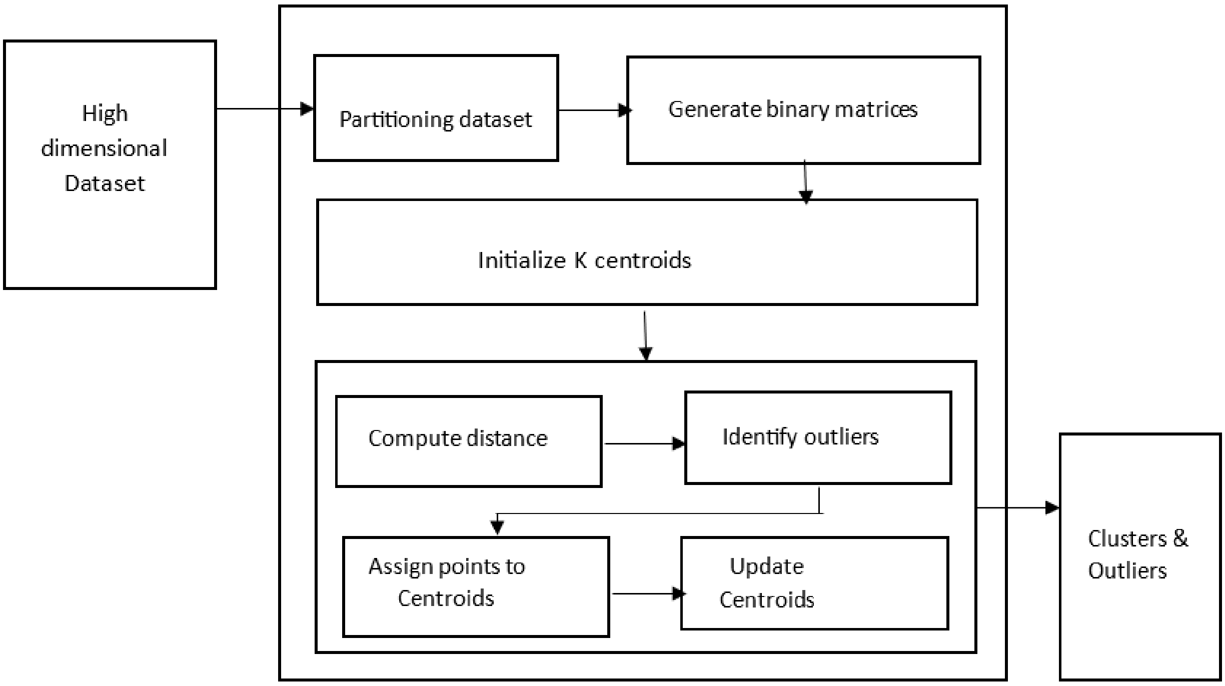

Fig 1: Architecture of Proposed System

Two double matrices are deduced from the partitioned data. One is deduced originally while the other is the optimized double matrix. These matrices are used in order to initialize K centroids. After this step, an iterative process is involved in order to cipher distance, identify outliers, assign points to centroids and update centroids. This process continues until confluence. Eventually, the process results in compact clusters and rightly linked outliers. While discovering points to be included in clusters, the frame contemporaneously discovers outliers that are insulated from clusters. Therefore the frame reflects a clustering medium which is non-exhaustive where some data points aren't assigned any cluster markers. Similar points are outliers that are used to made well-conditioned informed opinions in different real world operations

3.2.1 Benefits OF PROPOSED Framework

- The strategy works by finding lower layered projections which are locally scanty, and can't be found effectively by animal power procedures on account of the quantity of blends of conceivable outcomes.
- This method for exception identification enjoys upper hands over straightforward distance based anomalies which can't defeat the impacts of the dimensionality revile.
- We likewise showed how to carry out the method actually for high layered applications by utilizing a transformative inquiry procedure.
- We propose an original exception identification calculation in view of head part examination and bit thickness assessment.
- Mathematical investigations on manufactured and genuine information show that our strategy performs well on high-layered information. Specifically, the proposed technique outflanks the benchmark strategies as estimated by the F1-score. Our technique likewise delivers better-than-normal execution times contrasted with the benchmark strategies.

V. EXPERIMENTAL RESULTS

The proposed algorithm is evaluated using the datasets aforementioned and the results are compared with different outlier techniques such as K-Means, LOF [32], COF [33], LDOF [34], FABOD [35], iForest [36], OPCA [37], TONMF [38] and K-Means--[39]. 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

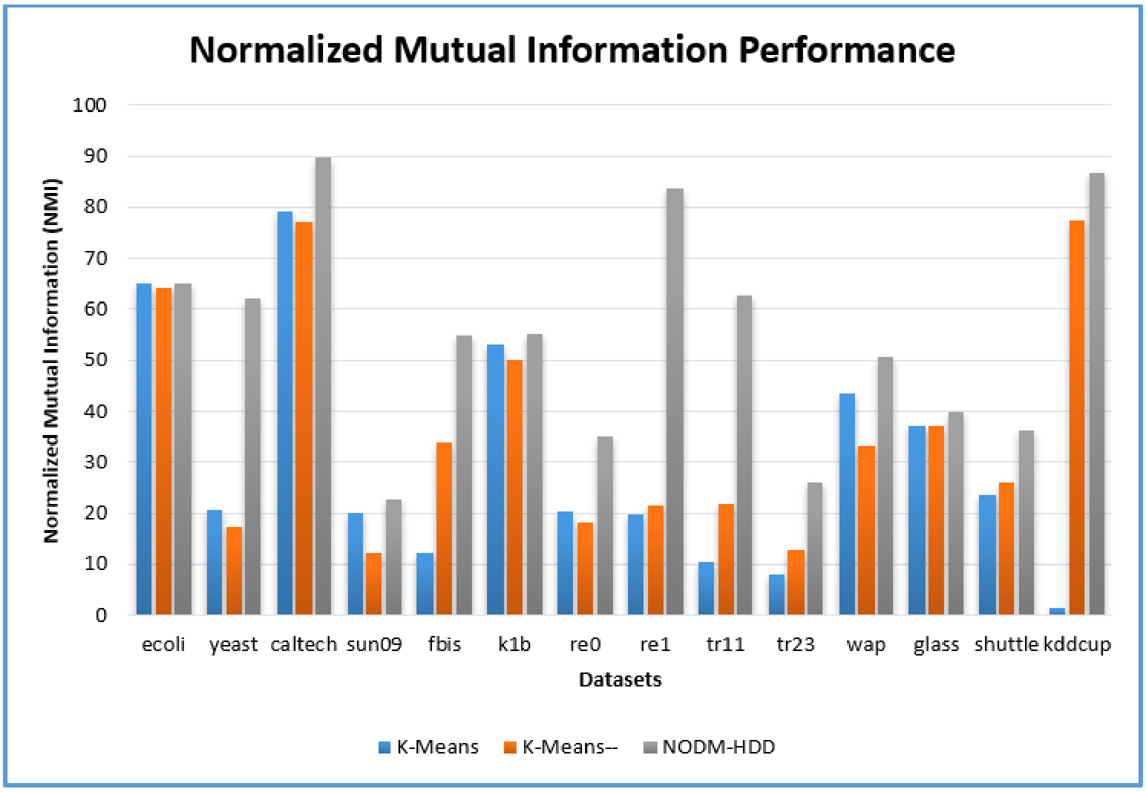

Figure 2: Performance evaluation in terms of NMI against different datasets

As introduced in Figure 2, the exhibition of the proposed calculation is contrasted and existing techniques against various datasets. The significant perception is made with NMI measure. Higher in NMI shows better execution. The level pivot shows the benchmark datasets utilized in exact review while NMI measure is displayed in vertical hub. A significant perception is that there is different execution in light of the dataset and its qualities. One more perception is that the NMI execution of the proposed calculation NOMD-HDD is superior to that of existing strategies for all the datasets reliably

VI. CONCLUSION AND FUTURE WORK

Outlier discovery is a necessary task that can be reused as part of real world operations. Still, utmost of the being styles dealt with unsupervised styles for clustering and outlier discovery independently. There's need for having an intertwined approach that leverages cluster performance and lead to outlier discovery. In this paper we proposed a frame known as unsupervised literacy grounded Outlier Detection Framework (UL- ODF). An algorithm named Novel Outlier Detection Method in High Dimensional Data (NODM- HDD). The algorithm has mechanisms to ameliorate conciseness of clusters made besides determining outliers. The algorithm exploits an enhanced interpretation of K-Means clustering fashion. A prototype is erected to validate the mileage of the proposed frame and the underpinning algorithm. Different standard datasets are used in the empirical study. Different criteria are used to estimate 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 discovery. Still, our work is grounded on only unsupervised literacy grounded approach. It lacks the advantages of a supervised system taking benefits of ground verity from unsupervised system. Thus, in our unborn work, we exploit both supervised and unsupervised literacy ways by defining a mongrel algorithm to descry outliers in high dimensional data.

REFERENCES

[1]. Hodge V, Austin J. A survey of outlier detection methodologies. ArtifIntell Rev. 2004; 22(2):85–126.
[2]. Agrawal S, Agrawal J. Survey on anomaly detection using data mining techniques. Procedia Comput Sci. 2015; 60:708–13.

[3]. Akoglu L, Tong H, Koutra D. Graph based anomaly detection and description: a survey. Data Mining KnowlDiscov. 2015; 29(3):626–88.

[4]. Chandola V, Banerjee A, Kumar V. Anomaly detection: a survey. ACM ComputSurv. 2009; 41(3):15.

[5]. Patcha A, Park J-M. An overview of anomaly detection techniques: existing solutions and latest technological trends. ComputNetw. 2007; 51(12):3448–70.

[6]. Jiang M, Cui P, Faloutsos C. Suspicious behavior detection: current trends and future directions. IEEE Intell Syst. 2016; 31(1):31–9.

[7]. Sorzano COS, Vargas J, Montano AP. A survey of dimensionality reduction techniques. arXiv preprint arXiv :1403.2877. 2014.

[8]. Gama J. Knowledge discovery from data streams. London: Chapman and Hall/CRC; 2010.

[9]. Gupta M, Gao J, Aggarwal CC, Han J. Outlier detection for temporal data: a survey. IEEE Trans Knowl Data Eng. 2014; 26(9):2250–67.

[10] Heydari A, aliTavakoli M, Salim N, Heydari Z. Detection of review spam: a survey. Expert Syst Appl. 2015;42(7):3634–42.

[11] Jindal N, Liu, B. Review spam detection. In: Proceedings of the 16th international conference on World Wide Web. ACM. 2007. pp. 1189–90.

[12] Parthasarathy S, Ghoting A, Otey ME. A survey of distributed mining of data streams. In: Data streams. Springer; 2007. pp. 289–307.

[13] Phua C, Lee V, Smith K, Gayler R. A comprehensive survey of data mining-based fraud detection research. arXiv preprint arXiv:1009.6119. 2010.

[14] Tamboli J, Shukla M. A survey of outlier detection algorithms for data streams. In: Computing for sustainable global development (INDIACom), 2016 3rd international conference on. IEEE. 2016. pp. 3535–40.

[15] Spirin N, Han J. Survey on web spam detection: principles and algorithms. ACM SIGKDD ExplorNewsl. 2012;13(2):50–64.

[16] Zimek A, Schubert E, Kriegel H-P. A survey on unsupervised outlier detection in high-dimensional numerical data. Stat Anal Data Mining ASA Data Sci J. 2012;5(5):363–87.

[17] Parsons L, Haque E, Liu H. Subspace clustering for high dimensional data: a review. ACM SIGKDD ExplorNewsl. 2004;6(1):90–105.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor
7.54

ISSN
INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com