ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 7| Issue: 12| December 2021|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2021: 8.047 || ISI Value: 1.188

ACCURATE AND EFFECTIVE OUTLIER DETECTION IN HIGH DIMENSIONAL DATA SETS BASED ON THE INTEGRATED LOCAL OUTLIER FACTOR (LOF) & PARZEN WINDOW (PW) METHOD

Isha[1], **Prof. Rajni Kori**[2]

[1]*Dept. of Computer Science and Engineering, LNCT, Bhopal, Bhopal, INDIA*
[2]*Dept. of Computer Science and Engineering, LNCT, Bhopal, Bhopal, INDIA*

ABSTRACT

With the rapid development into the new research application of outlier detection has been studied broadly in Machine Learning. Many traditional outlier detection techniques do not work well in such an environment. Therefore, developing up-to-date outlier detection methods becomes urgent tasks. Various methods for detecting different kinds of outliers in high-dimensional data sets from two different perspectives, i.e. detecting the outlying aspects of a data object and detecting outlying data objects of a data set.

In this proposed work, here we present an integrated methodology Local outlier factor (LOF) and Parzen window technique for the identification of outliers which is suitable for datasets with higher number of variables than observations for high dimensional dataset. Our integrated methodology is that the outliers can be detected without training datasets or prior knowledge about the underlying process that produces the dataset. Therefore, the local density based outliers can be computed very efficiently.

INDEX TERMS— ***Outlier detection, k nearest neighbours (k-NN), local outlier factor (LOF), intrinsic dimension.***

I. INTRODUCTION

Data quality is a prerequisite for any reliable, quantitative type of analysis and the large data sets which are becoming more common present specific challenges to the task of monitoring and ensuring data quality. One critical aspect of data quality monitoring is outlier detection i.e. the identification of values which either are obviously mistaken or seem to be unjustified from a business side perspective. Classical statistical methods, which underpin analytical tools supporting analysis of large data sets, are sensitive to the presence of outliers and may be led, consequently, to present a distorted picture of the reality due to the presence of outlier values leading to erroneous conclusions.

On another hand, a challenge for outlier detection in large datasets is that the development of purely automatic and efficient processes is required [1]. While large datasets present a number of challenges regarding data processing and analysis e.g. related to the difficulty of examining the data, it offers also the possibility to utilize the richness of the dataset for outlier detection. While the uncertainty of a large dataset may be larger compared to a smaller one (except e.g. in the case of surveys, where larger data allow more precise answers to be obtained from the data), the size of the dataset compels the development of methods which should make the optimal usage of the existing data.

A strand of research has focused on outlier identification on datum's that comprise vectors of numerical (or also categorical) attributes i.e. outliers when all dimensions of each datum are considered (hereafter called 'multi-dimensional outliers'). Indicative references from this vast literature include [2-4] However; the identification of multi-dimensional outliers presupposes a "clean" dataset with respect to the individual values in the sense that large deviations of exacting examinations of precise variables from their anticipated values characterize the true behaviour of the individual variables rather than outliers. In the reverse case, the detection of multi-dimensional outliers will also be have an effect on by distortions reasoned by outliers in entity values. In addition, in the high dimensional dataset several of these outliers may not be become aware of by multi-dimensional process as they possibly will be "hidden" within the granular dataset and not concern considerably the multi-dimensional distance metrics utilized by the individual outlier detection process. Consequently, it is significant to deal with the concern of outliers in single variables. Detecting outliers is to categorize the objects that significantly diverge from the common allocation of the data. Finding outlier point in main data streams can be valuable in many research areas such as analysis and scrutinizing of network traffic data e.g., connection-oriented records, web log, wireless sensor networks and financial transactions, etc.

Figure: Outlier Detection points

To encourage this research is that outliers accessible in real data streams are set in a few lower-dimensional subspaces. At this time, a subspace refers to as the real data space of outlier points. The survival of projected outliers is inspired that as the data dimensionality reduce on outlier data have a tendency to develop into regularly far-away from each other. Thus, the high esteem of data point's outlierness will be converted into progressively more fragile and thus undistinguishable.

II. Problem Statement

Outlier detection from stream data can find items i.e. objects or points that are abnormal or irregular regarding the common of items in the entire or a horizon/window of the data stream. They make available a collection of solutions to embark upon these disadvantages. But they focus on several of these open research issues, such as the relationship between numerous characteristic reduction methods and the resulting classification accuracy. The main objective is to recognize a set of features that best estimated the novel data without classification result. Other problems are based on computational cost of feature reduction algorithms for upcoming data requires developing computationally efficient feature reduction techniques which can be achieved concurrently. To finding outlier point various algorithms are originated upon statistical modeling techniques it can be any of them whether predictive or direct. Predictive techniques use tagged data using training sets to produce a finding outlier point data model i.e. contained by which outliers reduce for a domain which is subsequently utilized to categorize original data objects. Bit direct techniques are consist of deviation, proximity, statistical clustering and density based techniques pass on to those in which tagged training sets are occupied and for that explanation the organization of objects as finding outlier point is implemented through the measurement of statistical heuristics. Although characteristically more composite than predictive techniques, direct methods are not as much of constrained as detection is not dependent upon pre-defined models.

III. Outlier Detection in Large and High Dimensional Data

Today's the outliers can be detected by conducting hypothesis tests against an assumed distribution of the underlying process that produces the dataset. Only in sensible data or small dimensional data subspaces can considerable outlierness of data be scrutinized. For the reason that most of modern outlier detection methods achieve outlier detection in the full data space thus the projected outliers cannot be found by any of these techniques. This will show the way to a loss of remarkable and potentially valuable unusual patterns hidden in high-dimensional data streams. However their dimensions exploited for calculating point's outlierness are not gradually update and lots of techniques involve various scans of outlier data making them incompetent of handling data streams. For example, [5][6] use the Sparsity Coefficient to calculate data sparsity and this is based on technique that has to be cut down frequently from the data stream. This will be expensive and such updates will require multiple scans of data. [7-8] use data sparsity metrics that are involving the concept of concept of k-nearest neighbours (k-NN) to calculate distance.

IV. Learning Methods

Semi-supervised Methods: When training data is either available for normal observations or outliers but not both the semi-supervised methods are used to produce the boundaries for the known classes. For example, when the normal observations are known, an observation that falls outside the boundary of normal observations is an outlier. The one-class support vector machine [10] is a commonly used for semi-supervised outlier detection. Support vector machine (SVM) is a classification method using hyperplanes to partition a training dataset. Scholkopf et al [10] apply the kernel method to transform the data such that the points close to the origin are treated as another class. Therefore, the origin and the training data are used to construct the decision boundaries. The problem can also be generalized to the problem of classification with multiple classes by constructing multiple hyperplanes for each class. If a new observation does not belong to any class, it is an outlier.

Unsupervised Methods: In unsupervised methods, clustering methods are used to detect outliers. A clustering method groups similar observations using some objective function. Compared with the classification methods, the clustering methods can group data without training datasets. The clustering methods can be utilized to detect outliers by comparing the observations with the identified clusters. If an observation is far from the cluster centroids, it is declared an outlier. Although kmean is fast, it cannot detect clusters with different densities or shapes. In such cases, we can use density-based clustering methods such as the kernel method, SNN [11] in order to identify cluster centroids. There are several limitations of using clustering methods for outlier detection. The goal of clustering techniques is to detect clusters, not outliers. It is not optimal in outlier detection. For instance, an outlier close to a centroid can still be an outlier. An outlier possibly incorrectly unsigned as a normal observation, whereas a normal observation may be flagged as an outlier.

Distance-based and Density-based: In order to overcome the limitations of statistical and clustering methods in outlier detection, Knorr et al introduce a pruning strategy to facilitate speed up the algorithm. The advantage of the method is that it can detect outliers without any assumption about the fundamental allocation of a dataset. However, in some applications, the outliers of interest may not be the farthest observations. An outlier can be characterized by its most similar observations instead. Consequently, Breunig et al [9] introduce a density-based method that detects outliers with respect to their local densities. An observation that is far regarding its local region is think about an outlier. The method appears to be useful in practice.

V. LITERATURE SURVEY

In this paper [12], author has tried to develop a better learning method to identify outliers out from normal observations. The concept of this learning method is to make use of local neighbourhood information of an observation to determine whether it is an outlier or not. To confine the neighborhood information precisely an idea local neighbourhood information concept called LPS is initiated to compute the anomalous degree of an apprehensive observation. Formally, the LPS are dependable with the perception of nuclear norm and can be acquired by the procedure of low-rank matrix approximation. Furthermore, distinct offered distance-based and density-based detection methods the recommend technique is robust to the parameter k of k-NN embedded within LPOD. Using this method they are effectiveness algorithms on applying various outlier data sets. Experimental outcomes give you an idea about that the LPS are good at ranking the most excellent candidates for individual outliers and the show of LPOD is capable at many characteristics. While LPOD make use of k-NN to get neighbourhood information its competence relies on k-NN and its concert will be influenced by the distance formulation of k-NN to some area.

Zengyou He et. al. [13], called Discover CBLOF, provides outlier score known as the Cluster-based Local Outlier Factor (CBLOF) for each data point. The CBLOF measurement captures the size of the cluster in which the data object belongs, and the distance of the object to its cluster centroid. Initially they divided the data set to be set into clusters with a squeezer algorithm, and these clusters were divided into two clusters, namely large clusters and small clusters, based on the number of points in each collection. For each data point they count CBLOF and announce clusters.

Bay and Schwabacher et. al. [14] has shown that with enough detailed information, a simple pruning step can lead to the usual confusion of nearby neighborhood searches that will descend almost to the line. After calculating the nearest neighbors by a data point, the algorithm sets the outer limit of any data point to the weak outlier points obtained so far. Using this pruning process, the process discards nearby items, which is why it is unpopular. The performance of this algorithm is largely based on three assumptions, the violation of which could lead to malicious operation.

In [15], a feature bagging approach for finding outliers is suggested. It combines results from many outlier detection algorithms that are implemented using various features. The outlier detection algorithm uses a small set of randomly selected elements from the original feature set. As a result, each outlier detector identifies different outliers, and provides all data objects with outlier scores associated with the potential of becoming outliers. The outlier scores calculated by the individual outlier detection algorithms are then combined with the purpose of finding the better-quality outliers.

In this paper, here they propose [16] a hybrid semi-supervised anomaly detection model for high-dimensional data. Here author has using proposed detection model that consists of two parts: a deep auto encoder (DAE) and a together *k*-nearest neighbor graph- (*K*-NNG) based anomaly detector. The deep auto encoder (DAE) is promoting from the ability of nonlinear mapping method and to begin with only trained the essential features of data objects in unsupervised mode and to transform into high-dimensional data. In this method they are sharing of the training dataset is more dense in the compact feature dimensional data space to various nonparametric KNN-based detect anomaly detectors method with a part of a real life dataset rather than using the whole specific training set and this process greatly condenses the computational charge. Experimental results and statistical significance analysis shows that proposed method is evaluated on several real-life datasets and their performance confirms that the proposed hybrid model improves the anomaly detection accuracy and also they reduces the computational complexity than standalone algorithms.

We showed in previous work that this assumption is not true [17] and might even result in a incorrect density estimation. Therefore, we additionally evaluate a modified version of

CBLOF which simply neglects the weighting, referred to as unweighted-CBLOF (uCBLOF). The results of uCBLOF using a simple two-dimensional dataset where the color corresponds to the clustering result of the preceding k-means clustering algorithm. Similar to the nearest-neighbor based algorithms, the number of initial clusters k is also a critical parameter. Here, we follow the same strategy as for the nearest-neighbor based algorithms and evaluate many different k values. Furthermore, k-means clustering is a non-deterministic algorithm and thus the resulting anomaly scores can be different on multiple runs. To this end we follow a common strategy, which is to apply k-means many times on the data and pick the most stable result. However, clustering-based anomaly detection algorithms are very sensitive to the parameter k, since adding just a single additional centroid might lead to a very different outcome.

VI. PROPOSED METHODOLOGY

Input: A data set D, integer k, n, threshold φ, ε, and f^k.
Output: Top N outliers, and minimal number of interesting subspaces.

Step 1: Locate the outliers across the space.
Step 2: Dimensional reduction
Step 3: Extending predictable external identities: Externally difficult to understand.
Step 4: Getting Outliers in Very Small Dimensional Subspaces: Why and How Much Content is Easy to Take Out
Step 5: Parzen window method for local external factor and outdoor detection.

VII. PROPOSED ARCHITECTURE

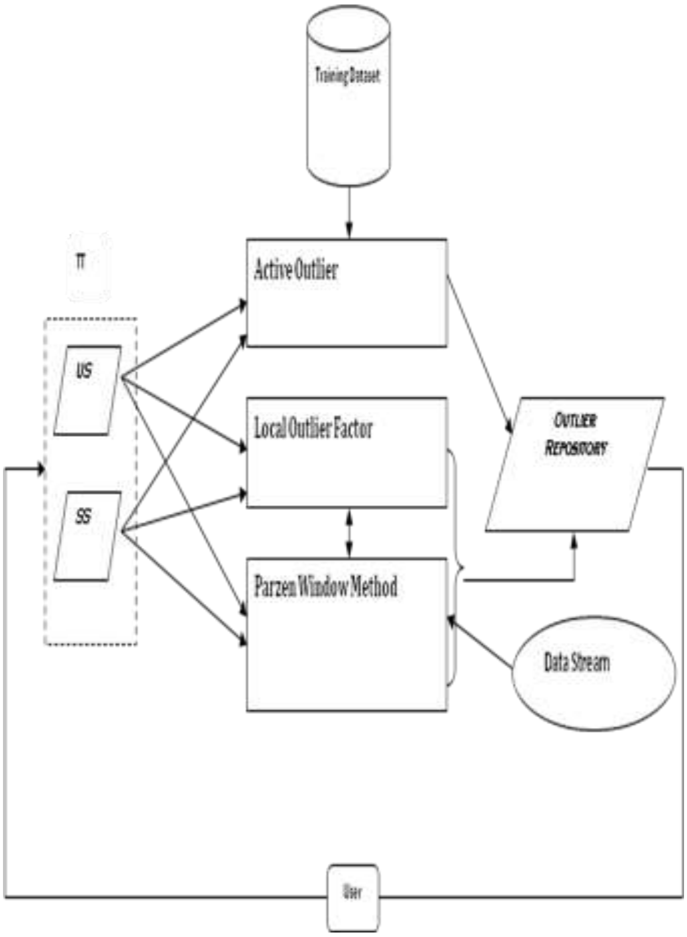
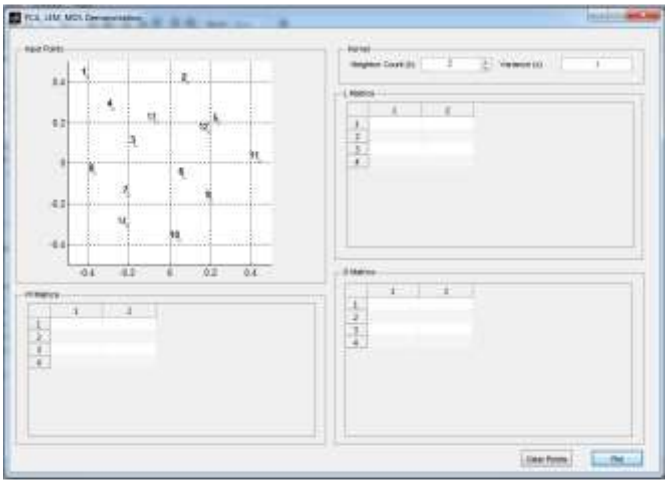
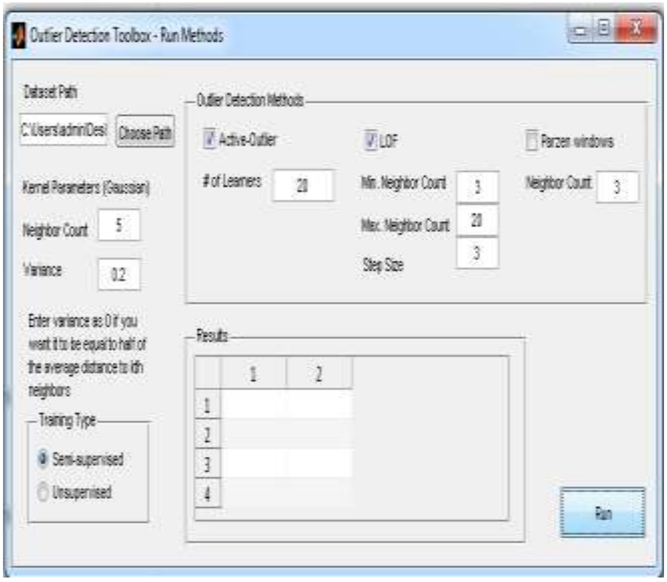


Figure: Overview of Learning Stage Outlier Detection

VIII. EXPERIMENTAL OUTCOMES





Outlier Detection Toolbox - Run Methods

Dataset Path: C:\Users\admin\Des| Choose Path

Kernel Parameters (Gaussian)

Neighbor Count: 5

Variance: 0.2

Enter variance as 0 if you want it to be equal to half of the average distance to kth neighbors

Training Type

● Semi-supervised

○ Unsupervised

Outlier Detection Methods

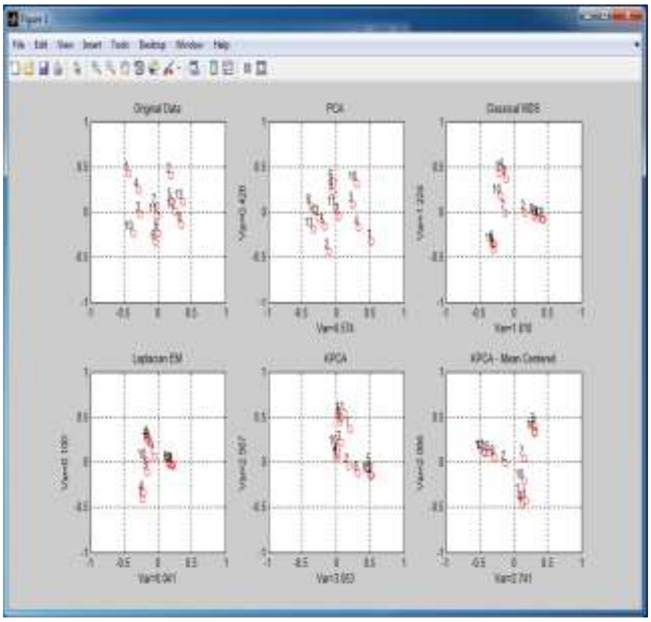
☑ Active-Outlier — # of Learners: 20

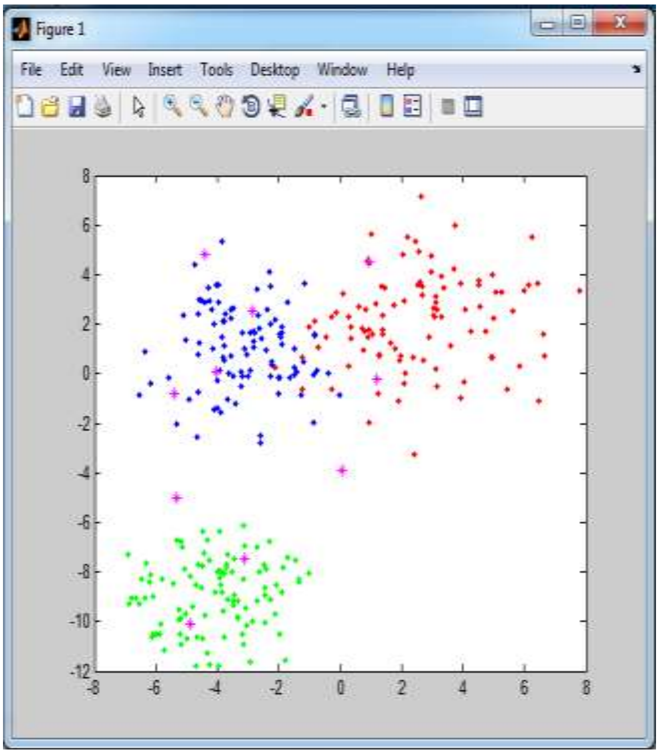
☑ LOF — Min. Neighbor Count: 3; Max. Neighbor Count: 20; Step Size: 3

□ Parzen windows — Neighbor Count: 3

Results

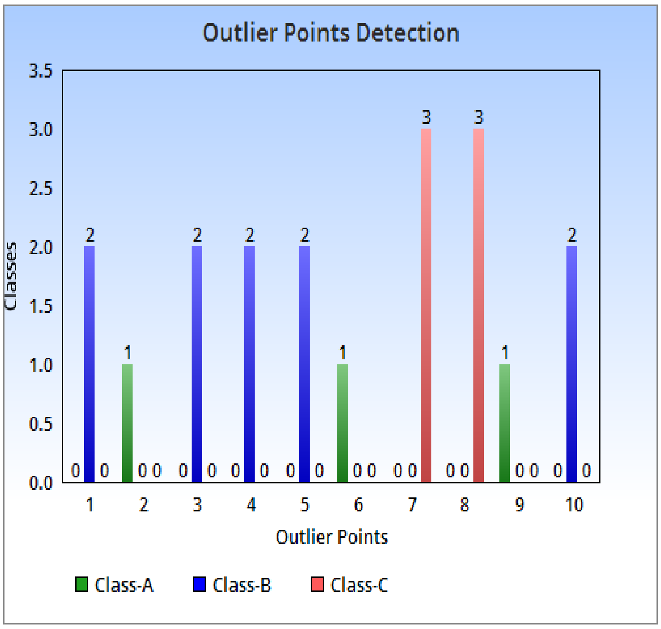
	1	2
1		
2		
3		
4		

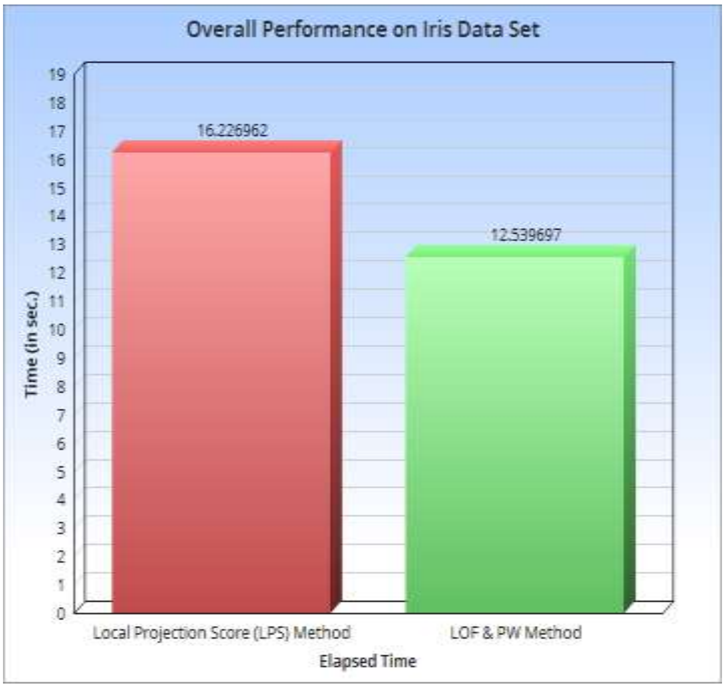


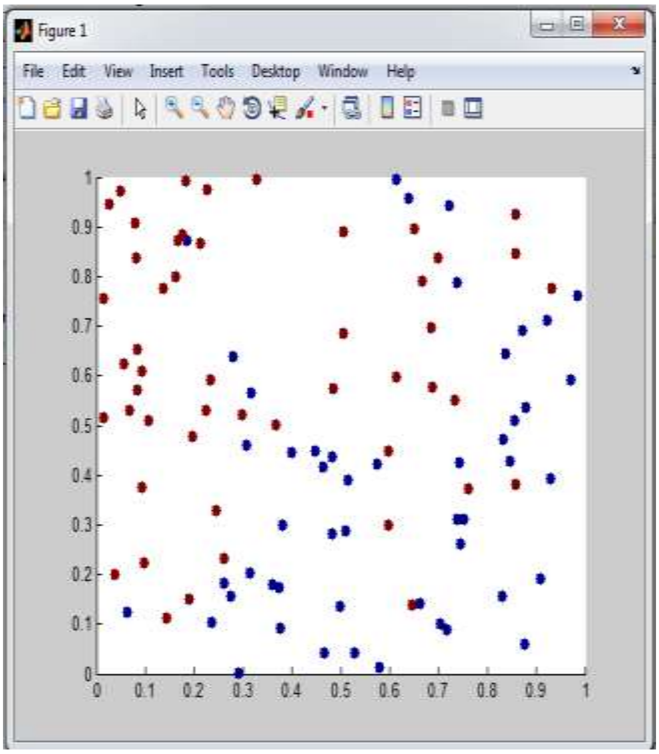


Select outlier points

it belong to the third class
it belong to the first class
it belong to the second class
it belong to the first class
it belong to the first class
it belong to the second class
it belong to the second class
it belong to the third class
it belong to the first class
it belong to the third class







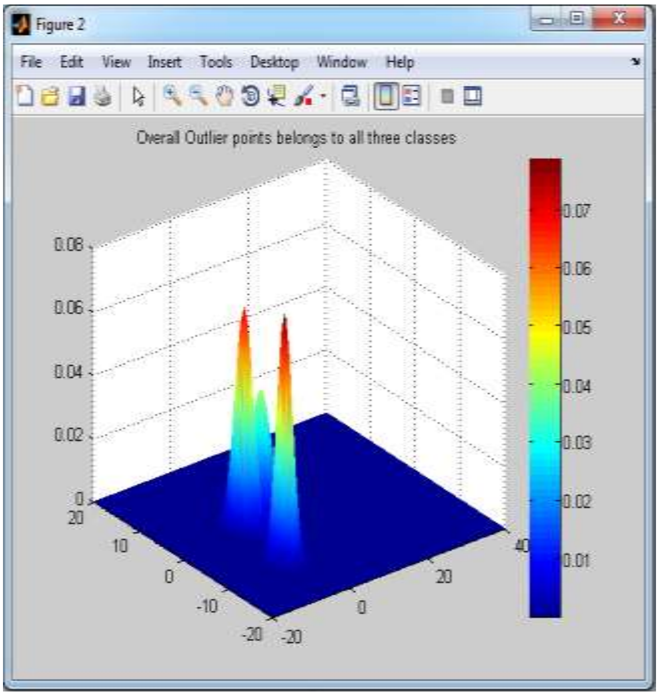
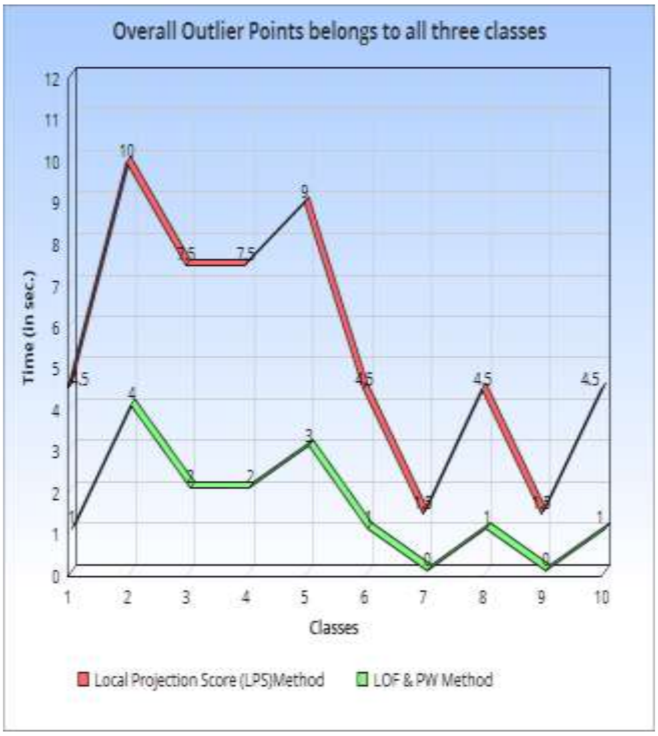


Figure 2
File Edit View Insert Tools Desktop Window Help
Overall Outlier points belongs to all three classes
0.08
0.06
0.04
0.02
0
20
10
0
-10
-20
-20
0
20
40
0.07
0.06
0.05
0.04
0.03
0.02
0.01

'Sepal Length' [0.9423]	'Sepal Width' [0.8453]	'Petal Length' [1]	'Petal Width' [1]
'Sepal Length' [0.9145]	'Sepal Width' [0.8785]	'Petal Length' [1]	'Petal Width' [1]
'Sepal Length' [0.9234]	'Sepal Width' [0.1341]	'Petal Length' [1]	'Petal Width' [1]
'Sepal Length' [0.9345]	'Sepal Width' [0.7134]	'Petal Length' [1]	'Petal Width' [1]



IX. CONCLUSION

This is an important measure of the machine learning task with many important applications such as medical diagnostics, detection of scams and intrusion detection to find outliers. Because of the large number of data objects in real life applications, outlier trail detection faces various challenges to detect these outlier points, so when we effectively reduce the dimensionality, they either increase the value of the data object or combine traditional algorithms to produce robust approximations to both high dimensional data and low dimensional data. High-dimensional data can be seen as part of a different challenge of outlier detection. The amount of data increases, but the dimensionality also increases: a larger set of lower dimensional sensors can be seen as a high dimensional multivariate time series.

In this paper, we provide the integrated methodology to identify the Local outlier factor (LOF) and parzen window technique suitable for datasets with a greater number of variables than the observations of higher dimensional datasets. Our method aims to use all the contextual information in the dataset to effectively identify in a dataset to detect outliers. This feature provides a flexible approach to application in large dimensional datasets and greatly improves high dimensional table data capability and accuracy for malicious outlier detection methods. Our integrated methodology is to find outliers without training dataset or prior knowledge of the underlying process of creating a dataset. Outliers are combined with individual subgroups. Therefore, local density based on outliers can be calculated very efficiently. Another challenge in outlier detection in high dimensional data is that outliers are often suppressed when too many dimensions do not display outliers to degrade in high dimensional data. Experimental analysis shows that our algorithm demonstrates the efficiency and effectiveness of integrated methodology in identifying outliers in high-dimensional data streams.

REFERENCES

1. *Maciá-Pérez F., Berna-Martinez J., Fernández Oliva A., Ortega, M. Abreu, 2015. Algorithm for the detection of outliers based on the theory of rough sets. Decision Support Systems, 75, pp. 63-75.*
2. *Otey, M., Ghoting, A., Parthasarathy, S., 2006. Fast distributed outlier detection in mixed-attribute data sets. Data Mining and Knowledge Discovery 12, 203-228.*
3. *Koufakou, A., Georgiopoulos, M., 2010. A fast outlier detection strategy for distributed high- dimensional data sets with mixed attributes. Data Mining and Knowledge Discovery, 20, 259-289.*
4. *Kutsuma, T., Yamamoto, A. 2017. Outlier detection using binary decision diagrams. Data Mining and Knowledge Discovery, 31, 548-572.*
5. *C. C. Aggarwal and P. S. Yu. Outlier Detection in High Dimensional Data. In Proc. of 2001 ACM SIGMOD International Conference on Management of Data (SIGMOD'01), Santa Barbara, California, USA, 2001.*
6. *C. Zhu, H. Kitagawa and C. Faloutsos. Example-Based Robust Outlier Detection in High Dimensional Datasets. In Proc. of 2005 IEEE International Conference on Data Management (ICDM'05), pp 829-832, 2005.*
7. *J. Zhang, M. Lou, T. W. Ling and H. Wang. HOS-Miner: A System for Detecting Outlying Subspaces of High-dimensional Data. In Proc. of 30th International Conference on Very Large Data Bases (VLDB'04), demo, pages 1265-1268,Toronto, Canada, 2004.*
8. *J. Zhang, Q. Gao and H. Wang. A Novel Method for Detecting Outlying Sub-spaces in High-dimensional Databases Using Genetic Algorithm. 2006 IEEE International Conference on Data Mining (ICDM'06), pages 731-740, Hong Kong, China, 2006.*
9. *Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J., "LOF:identifying density-based local outliers," SIGMOD Rec., vol. 29, no. 2, pp. 93–104, 2000.*
10. *Scholkopf, B., Platt, J. C., Shawe-Taylor, J. C., Smola, A. J., and Williamson, R. C., "Estimating the support of a high-dimensional distribution," Neural Comput., vol. 13, no. 7, pp. 1443–1471, 2001.*
11. *Ertoz, L., Steinbach, M., and Kumar, V., "Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data," in Proceedings of the third SIAM international conference on data mining, pp. 47–58, Society for Industrial and Applied, 2003.*
12. *Huawen Liu, Member, IEEE, Xuelong Li, Fellow, IEEE, Jiuyong Li, Member, IEEE, and Shichao Zhang, Senior Member, IEEE "Efficient Outlier Detection for High-Dimensional Data" IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS, 2017.*

13. *Zengyou He, Xiaofei Xu, and Shengchun Deng. Discovering cluster-based local outliers. Pattern Recognition Letters, 24(9-10):1641–1650, 2003.*
14. *Stephen D. Bay and Mark Schwabacher. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD 03), pages 29–38, New York, NY, USA, 2003. ACM.*
15. *Aleksandar Lazarevic and Vipin Kumar. Feature bagging for outlier detection. In Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, KDD '05, pages 157–166, New York, NY, USA, 2005. ACM.*
16. *Hongchao Song, Zhuqing Jiang, Aidong Men, and Bo Yang, "A Hybrid Semi-Supervised Anomaly Detection Model for High Dimensional Data" Comput Intell Neurosci. 2017.*
17. *Amer M, Goldstein M. Nearest-Neighbor and Clustering based Anomaly Detection Algorithms for RapidMiner. In: Simon Fischer IM, editor. Proceedings of the 3rd RapidMiner Community Meeting and Conferernce (RCOMM 2012). Shaker Verlag GmbH; 2012. p. 1–12.*