Outlier Detection in High Dimensional Data Streams to Detect Lower Subspace Outliers Effectively

Bhagyashri Karkhanis
Dept. of Computer Science and Engineering
Oriental Institute Of Science and Technology
Bhopal, India

Sanjay Sharma (Asst. Prof)
Dept. of Computer Science and Engineering
Oriental College of Technology
Bhopal, India

***Abstract* — Continuous, real-time processing of vast amounts of rapidly changing data lies at the heart of many modern and emerging applications. Detecting outliers is to recognize the objects that expansively turn out-of-the-way beginning the widespread allocation of the authentic data. Such that items may be distinguish as apprehensive data items as a result of the dissimilar method of production. Various algorithms have already occupational well in such a background for discovering outlier point. Various machine learning techniques are extending modern outlier detection techniques develop into persistent undertakings.**

Index Terms— Outlier detection, k nearest neighbours (k-NN), local outlier factor (LOF), intrinsic dimension.

I. INTRODUCTION

Various quantities of connected devices grows the velocity and volume of the data they produce and consume also grow. This massive data is scattered across many physically distributed sites, incurring high communication costs. The identification of outliers (i.e., data objects that do not fit well to the general data distribution) is very important in many practical applications. Application examples are the finding of fraud in financial transactions data, the identification of measurement errors in scientific data or the investigation of activities at statistics data. Various researches on such type of unsupervised problem of outlier detection precede the neighborhood by concerning collection methods [3]. Ensemble techniques, i.e., merging the findings or results of human being learners to a combined normally more consistent and enhanced accuracy are well started in the supervised context of classification or regression [1]. In unsupervised learning, the theoretical underpinnings are less clear but can be drawn in analogy to the supervised context as it has been done for clustering ensembles [2]. Here they presents two unsupervised algorithms to detect outlier observations whose aberrant behavior is hidden in lower dimensional subspaces or cannot be identified with the use of a single detector.

Most such applications arise in very high-dimensional domains. For instance, the credit card data set contains transaction records described by over 100 attributes [4]. To detect anomalous motion trajectories in surveillance videos, we have to deal with very high representational dimensionality of pixel features of sequential video frames [5]. Because of the notorious “curse of dimensionality", most proposed approaches so far which are explicitly or implicitly based on the assessment of differences in Euclidean distance metric between objects in full-dimensional space do not work efficiently. Conventional methods to detect distance-based outliers [6] or density-based outliers [7] suffer from the high computational difficulty for high-dimensional adjacent neighbour search.

Application area of main data streams consist of network traffic, telecommunications data, financial market data, and data from sensors that scrutinize the weather and environment, surveillance video and soon. Outlier detection from stream data can find items i.e. objects or points that are uncharacteristic or unbalanced about the popular of items in the entire or a horizon/window of the stream.

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.

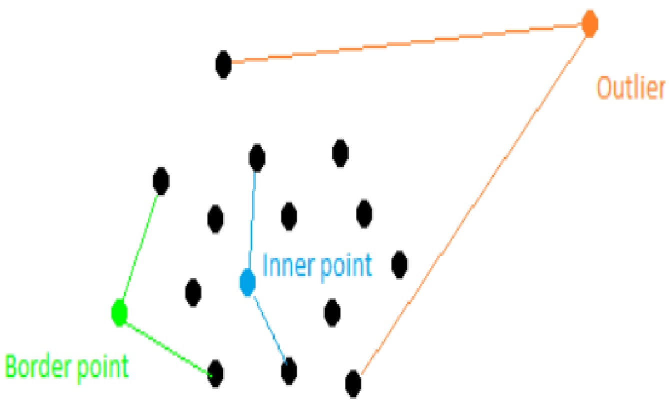


Figure: Outlier Detection points

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 outlier-ness will be converted into progressively more fragile and thus undistinguishable. As such, it is attractive to propose a various new techniques that well explain the disadvantages of these existing methods.

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. In many cases, the distributions are unknown. Moreover, the statistical tests are generalized to multivariate tests for datasets with more than one attribute. The multivariate statistical methods, however, are only effective for a extremely minute number of attributes. Therefore, distance-based outlier detection is practically preferred for multi-dimensional data by comparing the distance between the points in a dataset.

Breunig [8] introduces the idea of local density-based outliers. An observation that deviates greatly from its neighbors regarding its local density is considered to be an outlier. The density is measured by the length of the k-nearest neighbor distances of its neighbors. Even though a local outlier may not deviate from all the other observations, it, indeed, may signal interesting information. For example, a network activity may be considered exceptionally high depending on the nature of the network. Specifically, a server dedicated for content streaming likely requires more bandwidth than email servers.

However, in the calculation of local density-based outliers, the computation of the k-nearest neighbor is costly. The local density-based method utilizes multidimensional index trees to rate up the k-nearest neighbor computation. A multidimensional index tree is a hierarchy of the boundaries of the set of points in the sub-trees. When the numeral of dimensions enhances the number of possible boundaries to contain the subsets explodes. Therefore, the presentation of the index trees deteriorates in high dimensions. One of our contributions in this thesis is to introduce a method that can compute the local density-based outliers efficiently. We observe that when the dataset is partitioned into multiple subspaces, the local outliers are still prominent in any of these subspaces. In many cases, an outlier deviates greatly from only some subsets of the attributes. In multi-dimensional data, the delta deviations can be accumulated. When the number of attributes is large, the accumulation can be significant. This accumulation can even be greater than the deviation of an outlier in the set of attributes when it shows up as a strong outlier. Therefore, such outliers will be overlooked in a high dimensional dataset.

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 [9] 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 [9] 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 [10] 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 [8] 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

The common problem of identifying outliers has been addressed by various approaches that can be approximately confidential as global versus local outlier models. An overall model shows by outlier model show the ways to a binary decision of whether or not a given data object is to finding outlier point. A local outlier idea was to a certain extent allocates a degree of outlierness to each object to an “outlier factor” is a value distinguishing each object in “how much” this object is an outlier. Many applications rank the outliers in the database and to get back the top-n outliers a local outlier approach is apparently desirable. A different arrangement of outlier approaches discriminates between supervised and unsupervised approaches.

In this paper [11], 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.

Here author has introduce a new outlier scoring method for finding local outliers to distinguish between inliers and outliers in the surrounding area of Continuous ID allows for inliers members of a subspace cluster with other members of the same cluster and distinguishes ability from non-members of local outliers. Local outliers have a tendency to increase in the estimated value of their continuous intrinsic dimension at uncertainty point is modeled with continuous random variable. The proposed [12] author has local outlier achieve IDOS well-known LOF outlier score can be summarized by explanation of the model of continuous intrinsic dimensionality initiated. Here author has to compare with IDOS; LOF is make public to have the potential for assessment of local density within local clusters i.e. groups of inliers than IDOS has in its measurement of local intrinsic dimensionality which would make it harder to discriminate outliers in the neighborhood of such clusters values of data objects. An experimental analysis shows that the correctness of IDOS considerably increase to finding outlier point based on scoring methods mainly when the real data sets are huge and high-dimensional datasets show their superiority in terms of both effectiveness and efficiency.

In this paper, they propose [13] novel parameter-free approach angle-based outlier detection (ABOD) and several alternatives evaluating data points. This approach the consequences of the “curse of dimensionality” are alleviated on mining high-dimensional data where distance-based approaches often fail to offer high quality results. The basic concept of this method ABOD, they proposed two alternatives: Fast ABOD as acceleration suitable for low-dimensional but big data sets and LB-ABOD, a filter-refinement approach as acceleration suitable also for high-dimensional data. The main advantage of this proposed method on any constraint collection manipulating the feature of the accomplished ranking and here author has try to find rank the best candidates for being an finding outlier point with high precision and recall value. Here experimental assessment has to compare angle-based outlier detection to the well-started distance-based technique LOF for a variety of artificial data set and a real life data set and give you an idea about angle-based outlier detection to achieve mainly well on high-dimensional data.

As increasing dimensions of data objects it is difficult to find out data points which are not fitting in group i.e. cluster called outlier. This method is using to finding outlier point has significant in real life applications area of fraud detection, intrusion detection and various areas in which increasing data dimensions. Here author has to propose another method to divides original high dimensional data set in subspace clusters using subspace clustering method and here they try to improved k-means algorithms outlier cluster is establish which is additional amalgamated with other clusters depending upon compromise task. Various outlier clusters which are not going to combine find final outlier cluster. Here author [14] investigates various researches over many concepts of high dimensional data mining, information retrieval to finding outlier point in multi dimensional data ensemble subspace clustering, spam detection, improved k-means algorithm based on association rules. As these types of data is require to information systems so all these concepts can be used for improvement in data mining as well as machine learning methods. All these approaches are helpful for designing many strong applications for information retrieval. One application can be Spam Outlier Detection using Ensemble subspace clustering. In which spam outliers in analysis dataset of e-commerce can be detected. In this subspace clustering can be done trailed by outlier detection and again ensemble with other subspaces for enormous accuracy. In progress if we append spam detection logic then there will not be any concern for fraud reviews by someone. Whatever clusters are recognized as an outlier cluster from high dimensional data sets these can be highlighted or in some cases make some authorized essential accomplishments against all these entities. Second is they can put into practice elimination logic in datasets so that while performing data analysis when outliers are detected primarily if coming data is belonging to same dimension set will be rejected form adding it to the database.

In this paper, here they propose [15] 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.

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: ***Identify outliers in the complete space.***

A. To get distance-based outliers but make use of the ranks of distance as an alternative of the unconditional expanse in outlier detection.
B. For each entity *o*, discover its k-nearest neighbors: $nn_1(o), \ldots, nn_k(o)$
C. The weight of object o:

$$w(\boldsymbol{o}) = \sum_{i=1}^{k} dist(\boldsymbol{o}, nn_i(\boldsymbol{o}))$$

D. All objects are ranked in weight descending order.
E. Top-*l* objects in weight are amount produced as outliers (*l*: user-specified parm)
F. Make use of space-filling curves for rough calculation: scalable in both time and space w.r.t. data size and dimensionality.

Step 2: ***Dimensionality reduction***

A. Occupations only when in lower-dimensionality standard illustrations can at rest be well-known from outliers.
B. Conventionally PCA: Heuristically, the principal components with low variance are choosing for the reason that on such dimensions normal objects are expected close to each other and outliers frequently turn from the majority of classes in form of A-Classes, B- Classes and C-Classes.

Step 3: ***Widening predictable outlier detection: Hard for outlier interpretation***.

Step 4: ***To get acquire outliers in much lower dimensional subspaces: simple to take why and to what amount the thing is an outlier***

- E.g., find outlier data in certain subspace: calculate *average subspace A classes>> average subspace B classes << average subspace C classes*.
- Calculate (D, ω, ε) to find all interesting subspaces;
- Rank the reported interesting subspaces in a list L;

Step 5*:* ***A local outlier factor and parzen window method for outlier detection.***

- Project records onto different subspaces to get an region whose concentration is much lower than average.
- Discretize the information into a framework with φ equi-depth (why?) sections.
- Investigate for expanses that are considerably sparse
- Think about a k-d cube: k ranges on k dimensions with n objects.
- If objects are autonomously deal out the anticipated number of objects falling into a k-dimensional area is $(1/\varphi)^k n = f^k n$, and compute the standard deviation is

$$\sqrt{f^k(1-f^k)n}$$

- The sparsity coefficient of cube C:

$$S(C) = \frac{n(C) - f^k n}{\sqrt{f^k(1-f^k)n}}$$

- If S(C) < 0, C contains less objects than anticipated
- The new negative the sparser C is and the further expected the objects in C are outliers in the subspace.

VII. PROPOSED ARCHITECTURE

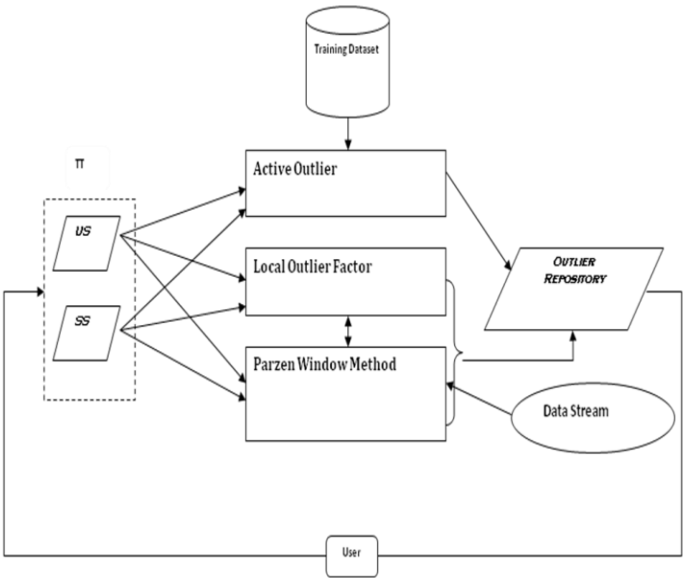
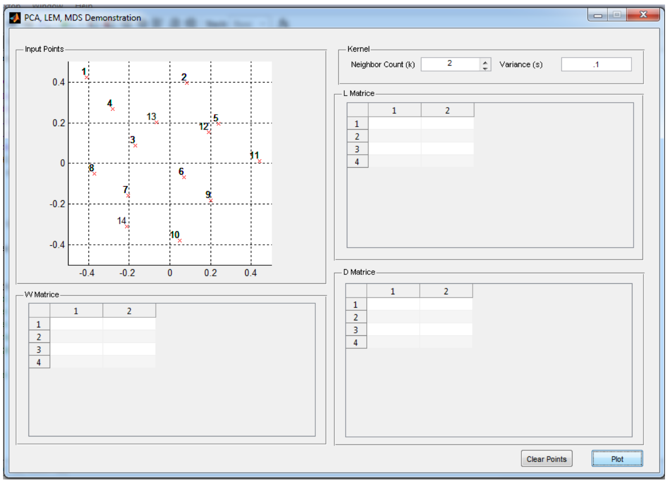


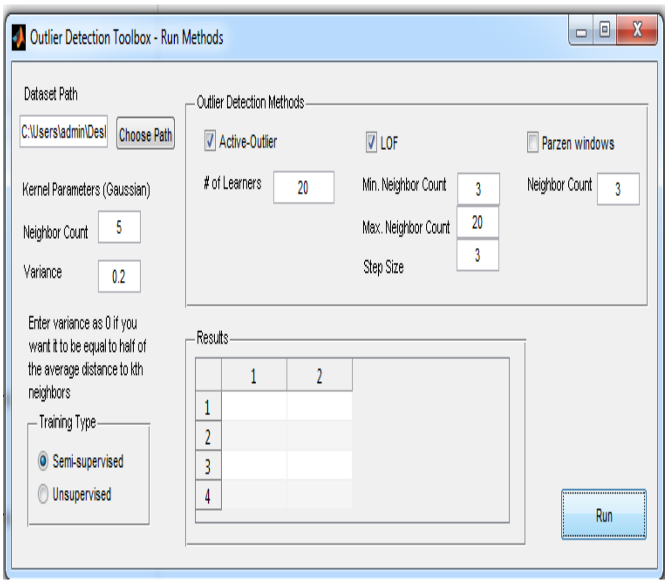
Figure: Overview of Learning Stage Outlier Detection

VIII. EXPERIMENTAL RESULTS



Neighbor Count (k) 2

Variance (s) .1



Dataset Path: C:\Users\admin\Desl

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

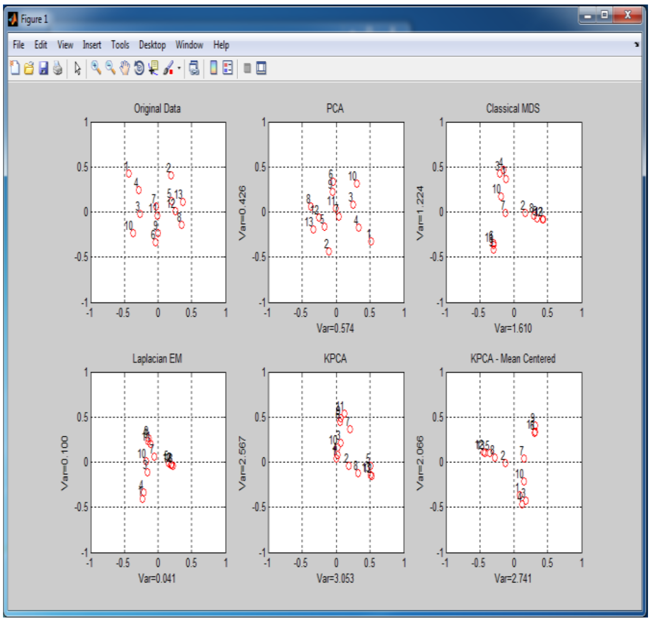
☑ Active-Outlier # of Learners 20

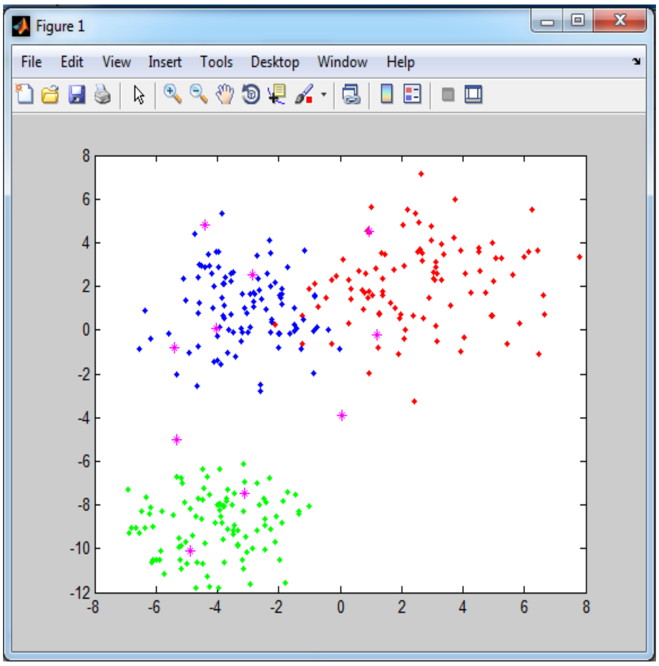
☑ LOF Min. Neighbor Count 3 Max. Neighbor Count 20 Step Size 3

□ Parzen windows Neighbor Count 3

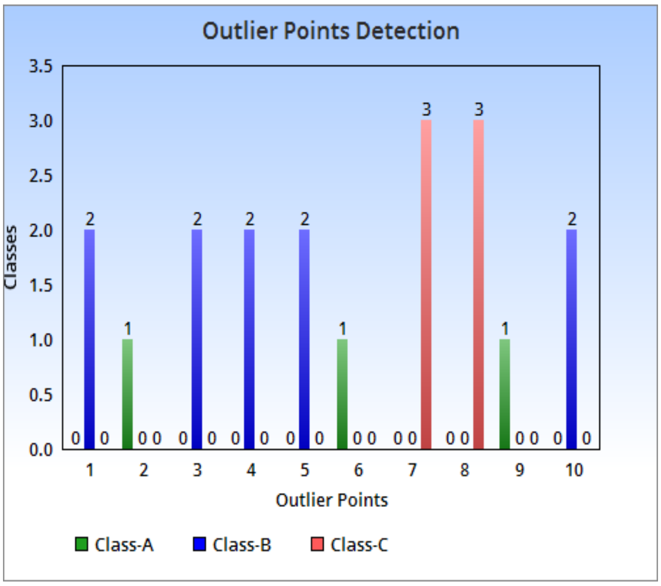
Results

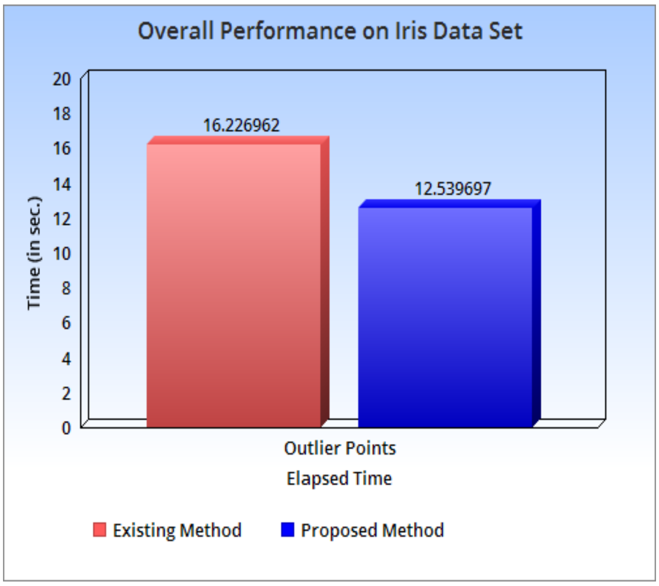
	1	2
1		
2		
3		
4		

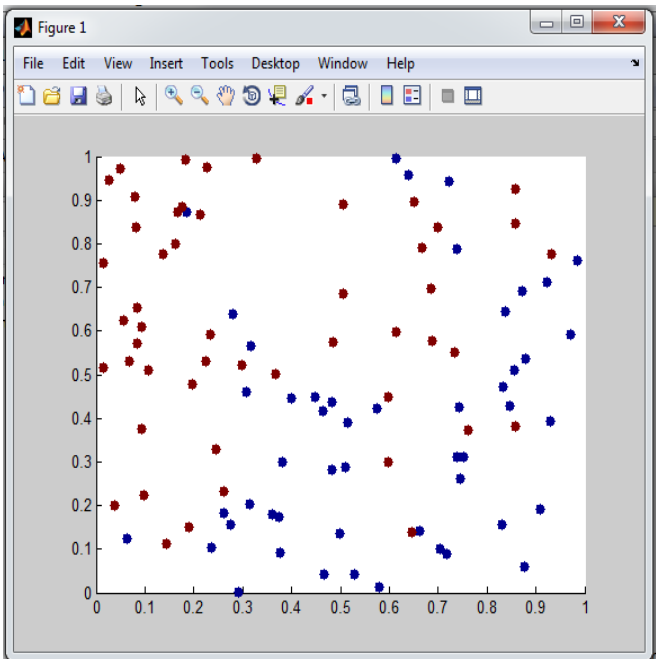




Select outlier points
it belong to the second class
it belong to the first class
it belong to the second class
it belong to the second class
it belong to the second class
it belong to the first class
it belong to the third class
it belong to the third class
it belong to the first class
it belong to the third class







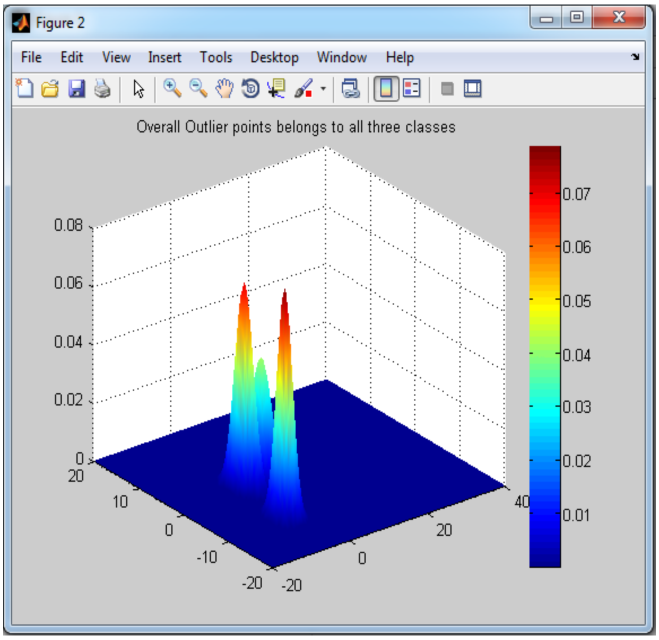
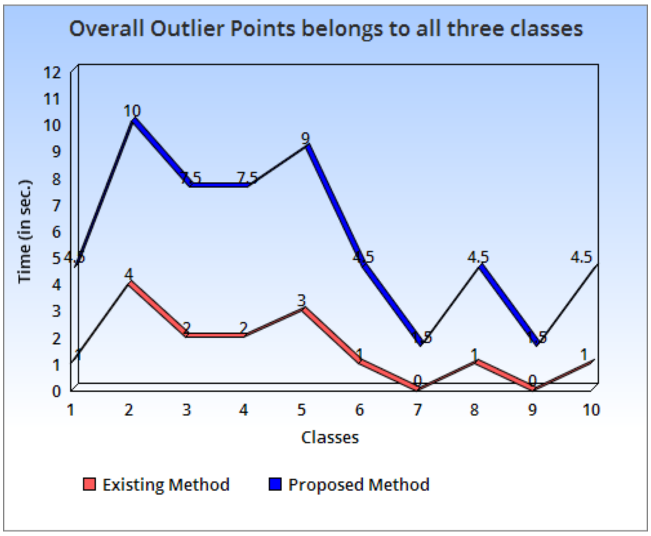


Figure 2
File Edit View Insert Tools Desktop Window Help
Overall Outlier points belongs to all three classes
0.08
0.06
0.04
0.02
0
20
10
0
-10
-20
-20
0
20
40
0.07
0.06
0.05
0.04
0.03
0.02
0.01

'Sepal Length' 'Sepal Width' 'Petal Length' 'Petal Width'
[0.9586] [0.8796] [1] [1]
'Sepal Length' 'Sepal Width' 'Petal Length' 'Petal Width'
[0.9586] [0.8796] [1] [1]
'Sepal Length' 'Sepal Width' 'Petal Length' 'Petal Width'
[0.9586] [0.1204] [1] [1]
'Sepal Length' 'Sepal Width' 'Petal Length' 'Petal Width'
[0.9172] [0.7592] [1] [1]



IX. CONCLUSION

To finding outlier point it is an important measurement of machine learning task with many critical applications, such as medical diagnosis, fraud detection, and intrusion detection. Due to huge quantity of data objects in real-life applications outlier detection faces various challenges to find these outlier points, so as we reduce the dimensionality efficiently they enlarge data object value or they combine traditional algorithms to build strong approximation for both high dimensional data and low dimensional data. High-dimensional data can be seen as part of the variety challenge of big data. Volume of data increases, but also the dimensionality: a large set of low-dimensional sensors can be seen as a high-dimensional multivariate time series.

Here we proposed the concept of anomalous patterns where the outliers follow certain patterns. The anomalous patterns imply that the deviation may not just be random. In our experiments we have revealed that top outliers are not always interesting. We have presented the adaptive dual-neighbor method to detect anomalous patterns based on the size and deviation. Our research analysis give you an idea about that our algorithm can discover interesting patterns which are undetected by clustering or outlier detection methods.

REFERENCES

[1] L. Rokach. Ensemble-based classifiers. Artif. Intell. Rev., 33:1-39, 2010.

[2] Z.-H. Zhou. Ensemble Methods. Foundations and Algorithms. CRC Press, 2012.

[3] A. Zimek, R. J. G. B. Campello, and J. Sander. Ensembles for unsupervised outlier detection: Challenges and research questions. SIGKDD Explor. 15(1):11-22, 2013.

[4] R. Wheeler and J. S. Aitken. Multiple algorithms for fraud detection. Knowledge Based Systems, 13(2-3):93-99, 2000.

[5] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In Proceedings of CVPR'10, pages 1975-1981, 2010.

[6] S. Ramaswamy, R. Rastogi, and K. Shim. E_cient algorithms for mining outliers from large data sets. In Proceedings of SIGMOD'00, pages 427-438, 2000.

[7] S. Papadimitriou, H. Kitagawa, P. B. Gibbons, and C. Faloutsos. Loci: Fast outlier detection using the local correlation integral. In Proceedings of ICDE'03, pages 315{326, 2003.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] Jonathan von Brunken, Michael E. Houle, and Arthur Zimek, “Intrinsic Dimensional Outlier Detection in High-Dimensional Data” NII-2015 -003E, Mar. 2015.

[13] Hans-Peter Kriegel, Matthias Schubert, Arthur Zimek” Angle-Based Outlier Detection in High-dimensional Data” ACM 978-1-60558-193, 2008.

[14] Suresh S. Kapare, Bharat A. Tidke, “Spam Outlier Detection in High Dimensional Data: Ensemble Subspace Clustering Approach” IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (3), 2015, 2326-2329.

[15] Hongchao Song, Zhuqing Jiang, Aidong Men, and Bo Yang, “A Hybrid Semi-Supervised Anomaly Detection Model for High Dimensional Data” Comput Intell Neurosci. 2017.