RESEARCH ARTICLE OPEN ACCESS

Hybrid Approach for Outlier Detection in High Dimensional Data

Neha Randive, Sneha, Neha Singh, Ruby Singh, Deepa Abin
Computer Dept. Pimpri Cinchwad College of Engineering Pune, India

Abstract

It has been observed recently that the prominence of multidimensional data is increasing. Existing outlier detection techniques generally fail to work on multi-dimensional data. The need for analyzing high dimensional data has thus increased in today's data trends. It has enormous application in medical domain, network intrusion and satellite imagery. Even though there are existing methodologies supporting high dimensionality and outlier detection problems, most of them are computationally intensive and time consuming. Thus there arises a need for an efficient and hybrid algorithm which will help to analyze and address these problems.

Index Terms— Outlier Detection, Feature extraction, DBSCAN algorithm, K-Mean algorithm.

I. INTRODUCTION

Nowadays many applications are using high dimensional data for example medical systems, internet traffic and in molecular biology. High dimension data are huge and enormous and it is not easy to handle high dimensional data. High dimensional data has large number of attributes .The number of attributes ranges from 10 to 10000. Out of these attributes, only a small no of attributes are relevant .Thus it becomes difficult to analyze this huge amount of data. Outlier detection is an important data mining task and has been widely studied in recent years. Unlike clustering where similar data elements are grouped to form clusters, in outlier detection the elements which does not comply with other elements in the dataset are identified. There are various applications in which detection of outliers can be more helpful than clubbing the similar datasets such fault diagnosis in machines, fraud detection and network intrusion.

It becomes very difficult to detect outliers in high dimensional data. As the number of attributes increases the data becomes sparser and every data object in dataset acts as a potential outlier. There are various algorithms which are used for outlier detection do not require labeled data.

In this paper we propose a hybrid approach for outlier detection in high dimensional data by combining both density-based and distance-based approach. By combining these approaches we can take benefit of both density and distance-based clustering methods. Most of the clustering algorithms used for outlier detection in lower dimension datasets cannot be applied to high dimensional datasets due to the curse of dimensionality. DBSCAN (Density Based Spatial Clustering Application with Noise) a density based technique and K-means are combined in this hybrid approach.

II. OBJECTIVE OF STUDY

Detecting outliers for high-dimensional data with the help of various clustering techniques does not give efficient results. In high-dimensional space the data is sparse and the notion of proximity fails to retain its meaningfulness. The scarcity of the high dimensional data implies that that every point is an equally good outlier from the perspective of proximity based definition. Thus, the notion of finding meaningful outliers in this case becomes substantially more complex and non-obvious. Hence applying a single technique is not sufficient. And which compelled us to use a hybrid approach by combining two or more algorithms.

III. RELATED WORK

Outlier detection also termed as deviation detection, exception mining, and novelty detection. is an important problem that has attracted wide interest and different solutions. These solutions can be broadly classified into several major ideas:

Model-Based [2]: An explicit model of the domain is built (i.e., a model of the heart, or of an oil refinery), and objects that do not fit the model are flagged.
Disadvantage:This method requires a model to be built, which requires the input of a domain expert.It is often an expensive and difficult enterprise.

Connectedness [11]: In domains where objects are linked (social networks, network intrusion, geological and biological networks), objects with few links are considered potential anomalies.
Disadvantage: Connectedness approaches often deals with datasets which consists of linkage information

Density-Based [3]: Objects in low-density regions of space are flagged.

Disadvantage: Density based models require the careful settings of several parameters. It requires quadratic time complexity. It may rule out outliers close to some non-outliers patterns that has low density.

Distance-Based [1]: Given any distance measure, objects that have distances to their nearest neighbors that exceed a specific threshold are considered potential anomalies. Distance-based methods are much more flexible and robust. They are defined for any data type for which we have a distance measure and do not require a detailed understanding of the application domain

Cluster based approach [4]: The clustering based techniques involve a clustering step which partitions the data into groups which contain similar objects. The normal behavior of outliers is that they belong to any cluster, maybe small or large. Clustering based outlier detection techniques have been enveloped which make use of the fact that outliers do not belong to any cluster since they are very few and different from the normal instances.

K-Nearest Neighbor Based Approach [12]: K-nearest neighbor based schemes analyses each object with respect to its local neighborhood. The basic idea behind such schemes is that an outlier will have a neighborhood where it will stand out, while a normal object will have a neighborhood where all its neighbors will be exactly like it. The obvious strength of these techniques is that they can work in an unsupervised mode.

IV. PRELIMINARIES

- High-Dimensional Data – In high-dimensional space the data is sparse and the notion of proximity fails to retain its meaningfulness. Each dimension in a dataset refers to an attribute .The scarcity of the high dimensional data implies that that every point is an almost equally good outlier from the perspective of proximity-based definition.

Thus, the notion of finding meaningful outliers in this case becomes substantially more complex and non-obvious.

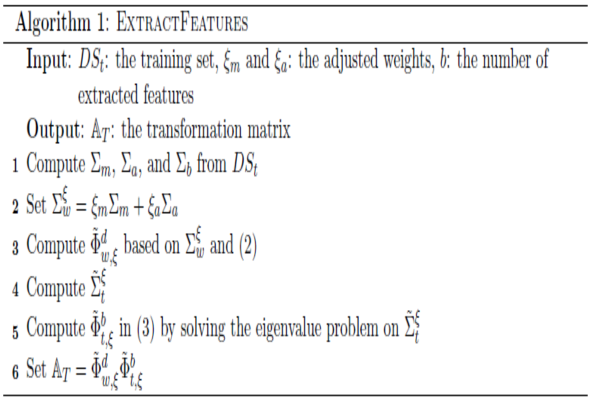
- Data Chunk – For high dimensional data it is not possible to store complete data at once, for processing we divide it into data chunks of same size. Chunk size is specified by the user and depends upon the nature of data. In our scheme an object will be examined over multiple consecutive data chunks before declaring it as outliers because a cluster may be split over two data chunks or new cluster may be emerging and an outlying object of current data chunk may become inlier when similar data objects occur in the next chunks.
- Weight - Weight of an attribute gives its degree relevance in data mining. Larger weights are assigned to relevant attribute and smaller to noisy attributes. Sum of all weights is always one. So all distance calculations in updation, clustering and outlier detection part of proposed scheme use corresponding attribute weights.
- Variance Matrix – It is a matrix of 1×*m* dimensions where m is number of attributes. Each entry of this matrix stores sum of distance of all object from their corresponding cluster in corresponding attribute.
- Candidate Outlier - An object is candidate outlier if it deviates more than the given threshold from normal clusters based upon deviation criteria. We have used same criteria as defined in for scattered outliers because it is useful even when clusters are of different density and size.
- Real Outlier - A candidate outlier becomes real outlier when it fulfils deviation criteria up to L data chunks.

V. METHODOLOGY

The approach here is hybrid. We first divide the dataset into a number of data chunks so as to simplify the execution. The data chunks need to be made noise free. Thus the process of data-preprocessing is necessary. This helps in data redundancy, reducing the noise in the data and also normalizing it.

1) Feature Extraction

Being able to reduce the number of data dimensions helps to overcome the lack of data and avoid the over-fitting issue. Recognizing this need, the *Dimensionality Reduction/Feature Extraction for OUTlier Detection* (DROUT) [11] an efficient method for feature extraction. A relevant set of features is then rxtracted, and finally transforms the testing set where detection algorithms are applied using the features obtained as in Fig.1 below..



```
Algorithm 1: ExtractFeatures
  Input: $DS_t$: the training set, $\xi_m$ and $\xi_a$: the adjusted weights, $b$: the number of
         extracted features
  Output: $\mathbb{A}_T$: the transformation matrix
Compute $\Sigma_m$, $\Sigma_a$, and $\Sigma_b$ from $DS_t$
Set $\Sigma_w^{\xi} = \xi_m \Sigma_m + \xi_a \Sigma_a$
Compute $\tilde{\Phi}_{w,\xi}^{d}$ based on $\Sigma_w^{\xi}$ and (2)
Compute $\tilde{\Sigma}_t^{\xi}$
Compute $\tilde{\Phi}_{t,\xi}^{b}$ in (3) by solving the eigenvalue problem on $\tilde{\Sigma}_t^{\xi}$
Set $\mathbb{A}_T = \tilde{\Phi}_{w,\xi}^{d} \tilde{\Phi}_{t,\xi}^{b}$
```

Fig 1: Feature extraction [11]

By performing eigenspace regularization, we are able to mitigate the loss of discriminated information during the feature extraction process [11]. Furthermore, different from other techniques on feature extraction, in DROUT, eigenvalue regularization and feature extraction are performed on *weight-adjusted* scatter matrices instead of normal ones. The target of these matrices is at outlier detection where class cardinalities (normal class v/s. outlier class) are highly unbalanced. This helps DROUT to work better than existing techniques in mining anomalies. The process of feature extraction reduces the number of attributes and thus DBSCAN algorithm can be applied efficiently.

2) *DBSCAN Algorithm*

In this section DBSCAN algorithm is used for clustering current data chunk. It does not require number of clusters and can find the arbitrary shape clusters. DBSCAN output is a set of clusters and outlying objects. Outlying objects are considered as those which have possibility of being real outlier known as ‘Candidate Outliers’. Small size clusters of DBSCAN may be group of outliers or these may be portion of a cluster that yet to be come in next data chunk and has been split over to chunks. So these objects are also treated as candidate outliers and feed to outlier detection module. For clustering the current data chunk weights of previous phase are taken and that are then updated in weighted k-mean clustering module. DBSCAN parameters MinPts & epsilon are updated using equations in Fig 2.

$$\left.\begin{array}{l} \text{Epsilon} = \dfrac{\sum\limits_{i=1}^{k} \text{Avg Intra } (Ci)}{k} \\ \text{where} \quad \text{Avg Intra } (Ci) = \dfrac{\sum\limits_{i=1}^{n} \sum\limits_{j=1, j\neq i}^{n} \text{Dist } (Oi, Oj)}{2 \times n} \end{array}\right\} \quad (1)$$

$$\left.\begin{array}{l} \text{Min Pts} = \text{Avg No. of Objects with a distance of Epsilon} \\ \qquad \text{from a object in cluster of smallest density} \\ \\ \text{where Density } (Ci) = \dfrac{\text{No. of Object } (Ci)}{\text{Radius } (Ci)} \end{array}\right\} \quad (2)$$

Fig 2: MinPts calculation[2]

In above equations k is the number of clusters, n is the total number of objects comprises all clusters. C*i* represent i*th* cluster and o*i* represent i*th* object. Dist ( ) is the distance between two objects. Based on these equation parameters of DBSCAN are updated in MinPts & epsilon updation section.

3) *Updation Module*

Objects of DBSCAN clusters are assigned to their nearest cluster centres of previous phase clustering. For initial phase DBSCAN clusters are considered as it is and their mean are taken for centres. But in other phases after assigning objects to centres these newly assigned objects are used to calculate current centres and variance matrix. Updated centres are calculated using additive property, by summing up current and previous phase centres.

$$\left.\begin{array}{l} \text{w}_j = \begin{cases} 0 & if\ D_j=0 \\ \dfrac{1}{\sum_{t=1}^{h} \left[\frac{D_j}{D_t}\right]^{\frac{1}{\beta-1}}} & if\ D_j \neq 0 \end{cases} \\ where \text{Dj} = \sum_{l=1}^{k} \sum_{i=1}^{n} u_{i,l} d(x_{i,j}, z_{l,j}) \end{array}\right\} \quad (3)$$

$$\sum_{j=1}^{m} w_j = 1 \text{ and } 0 \leq w_j \leq 1 \qquad (4)$$

In above equations w*j* represents weight of j*th* attribute and h is the number of variable where D*j* = 0.*β* is the parameter supplied by user, k gives the number of clusters, n is the number of objects, u*i,l* is one if i*th* object is member of l*th* cluster otherwise it is zero, d is the distance measure like Euclidian or could be any other measure. Z*l* is center of l*th* Cluster.

Here, to calculate updated weights D*j* is equal to sum of previous phase and current phase variance matrix entries corresponding to j*th* attribute. During the current chunk processing we are considering both old and current statistics (centres and weights).

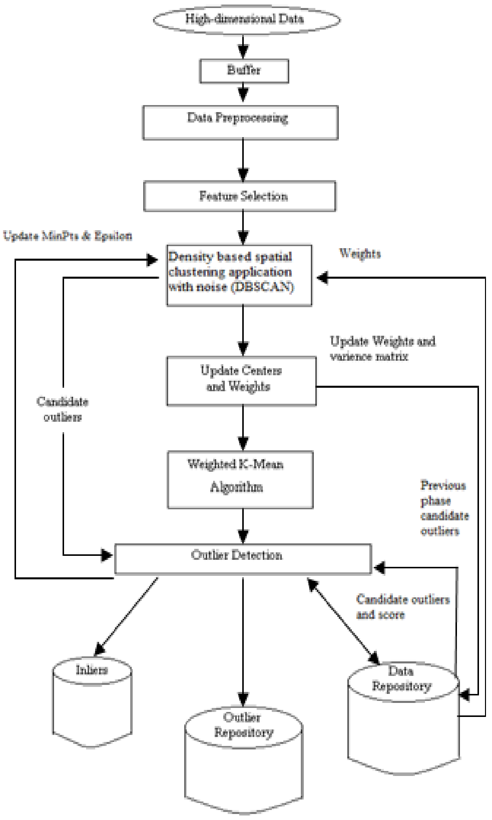

High-dimensional Data
Buffer
Data Preprocessing
Feature Selection
Update MinPts & Epsilon
Density based spatial clustering application with noise (DBSCAN)
Weights
Update Weights and varience matrix
Update Centers and Weights
Candidate outliers
Weighted K-Mean Algorithm
Previous phase candidate outliers
Outlier Detection
Candidate outliers and score
Inliers
Outlier Repository
Data Repository

Fig 3: Architecture Diagram

4) Clustering Module

In this module data chunk is again clustered using weighted k-mean[1] clustering method. It is a partitioning clustering algorithm. Updated centres and weights from updation module are used as initial centre and initial weight in clustering. Weighted-k-mean is a k-mean based clustering algorithm which includes iterative weight updation over the clustering process according to equation (1) and (2). In such a manner that intra cluster dissimilarity minimize and inter cluster dissimilarity maximize. Clusters, weights and variance matrix are the output of this module. Clusters are input to outlier detection module. Cluster centres, weights and variance matrix are saved in data repository and old once are discarded.

5) Outlier Detection

Outlier detection process is a combination of three steps. In first step candidate outliers are found out using the clusters, deviation criteria [2] and threshold and their score is initialized to one. Candidate outliers and their score are saved to metadata repository. In second step previous phase candidate outliers are tested whether they still satisfy deviation criteria or not.

If they satisfy then their score is checked if it is less then L then it is incremented by one and these objects with their score are added to candidate outliers of current phase on metadata repository. If previous phase candidates satisfy deviation criteria and their score is equal to L then these objects are real outlier and stored in outlier repository. We check candidate outliers for consecutive L data chunks before declaring them as outliers or inliers. When new candidate outliers and scores are saved in metadata repository, old once are deleted.

VI. Conclusion

We are proposing an outlier detection method where we first group the data (having similar characteristics) in to number of clusters. Due to reduction in size of dataset, the computation time will reduce considerably. Then we would take threshold value from user and calculate outliers according to given threshold value for each cluster. Outliers will be obtained within a cluster. Hybrid approach would take less computation time. This will be applicable for numerical data, later which can be further modified for textual mining too. Complex and varying real world datasets can be considered for greater experimentation.

References

[1] J. Han and M. Kamber, Data Mining: Concepts and Techniques, J. Kacprzyk and L. C. Jain, Eds. Morgan Kaufmann, 2006, vol. 54, no. Second Edition.

[2] Yogita and D. Toshniwal, \A framework for outlier detection in evolving data streams by weighting attributes in clustering," in Proceedings of the 2nd International Conference on Communication Computing and Security, India, 2012.

[3] S. Ramaswamy, R. Rastogi, and K. Shim, \E_cient algorithms for mining outliers from large data sets," in Proceedings of the 2000 ACM SIGMOD international conference on Management of data, ser. SIGMOD '00. New York, NY, USA: ACM, 2000

[4] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, \Lof: identifying density- based local outliers," in Proceedings of the 2000 ACM SIGMOD international conference on Management of data, ser. SIGMOD '00. New York, NY, USA: ACM, 2000

[5] Z. He, X. Xu, and S. Deng, \Discovering cluster based local outliers," Pattern Recognition Letters, vol. 2003

[6] M. Elahi, K. Li, W. Nisar, X. Lv, and H. Wang, \E_cient clustering-based outlier detection algorithm for dynamic data stream," in Proceedings of the 2008 Fifth

International Conference on Fuzzy Systems and Knowledge Discovery - Volume 05, ser. FSKD '08. Washington, DC, USA: IEEE Computer Society, 2008

[7] F. Angiulli and F. Fassetti, \Detecting distance-based outliers in streams of data," in Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, ser. CIKM '07. New York, NY, USA: ACM, 2007

[8] S. Subramaniam, T. Palpanas, D. Papadopoulos, V. Kalogeraki, and D. Gunopulos, Online outlier detection in sensor data using non-parametric models," in Proceedings of the 32nd international conference on Very large data bases, ser. VLDB '06.

[9] M. S. Sadik and L. Gruenwald, DBOD-DS : Distance Based Outlier Detection for Data Streams. Springer, 2011

[10] F. Angiulli, S. Basta, and C. Pizzuti, \Distance-based detection and prediction of outliers," IEEE Trans. on Knowl. and Data Eng., vol. 18, no. 2, pp. 145160, Feb. 2006.

[11] Feature Extraction for Outlier Detection in High Dimensional Spaces by Hoang Vu Nguyen,2012.

[12] Jiawei Han, Micheline Kamber, ―Data Mining Concepts and Techniques.