Outlier Detection: The State of the Art and Research Directions in Machine Learning for Healthcare Anomaly Detection and Clinical Decision Support Systems

Prof. Giulia Conti[1]
[1] University of Toronto, Department of Data Science and Biomedical Informatics, Toronto, Canada

ABSTRACT

Outlier detection is a significant research area in data mining. An Outlier is a point or a set of points in a data set that exhibit a different behaviour compared to other objects. In static data sets, these anomalous patterns or variability can be detected using the traditional outlier detection techniques. The detection of outliers in data streams is a challenging task as changes or updations may take place during data processing and from multiple sources at a high rate. Here conventional detection techniques cannot be applied directly to achieve the time and space requirements in data streams. Combined approaches of several techniques are essential to extract outliers from such knowledge set. This paper focused on various algorithms used in density, distance, clustering based outlier detection among data streams and portrayed the metrices used in outlier detection. Also proposes a generalised framework to detect outlierness of an object in datastreams

KEYWORDS: Outlier, Data Stream ,Distance based , Density based,Clustering based.

1. INTRODUCTION

Now a day's enormous amount of unformatted raw data is circulating through internet and other data processing sources. Extract relevant information from the web for real world applications, we need a technique to sieve the valuable data from the raw data available in the internet. To extract relevant information from the raw data,the whole data should undergo some pre-processing steps. The main pre-processing methods [1] involved are data cleaning, selection, integration and transformation. Data mining techniques can then be applied to gather relevant knowledge from the web.

Data mining refers to the findings of patterns of interest from knowledge set through data analysis and data categorization. The knowledge sets can be categorized into static data sets and data streams. Static data sets are those data sets in which the data recorded will not be changed or updated. For example, while filling online application form we can only choose cities from the available list of cities for writing the exam. In dynamic data sets, changes or updations may take place during data processing. Data streams are static data sets received from single source or multiple sources arriving at high rate. Figure 1 gives an overview of the concept of Data mining.

Information containted on these sets of data may have flaws and faults at the time of recording and this may cause errors during process. Detecting the erroneous pattern is a very crucial task in data mining and such patterns are termed as anomalies or outliers. An outlier can be defined as a point or an object in a data set that is completely different from all the other objects. So mining the unusual pattern from the raw data is beneficial in the fields of intrusion detection, human gate analysis, credit card fraud detection etc. In real world, outliers may occur in any circumstances like errors that occur during the reading of a physical instrument or in a program fragment.

A variety of techniques are available to detect outliers in static data sets and data streams. The density based method is a generally accepted method for extracting outliers from static data sets. It finds out the anomalous point on the basis of local outliers [2]. But the detection accuracy of these results is poor in stream data. The reason is that at a time only a portion of data is available for detection and this leads to updation in results [3]. Incremental version of LOF method is adopted to detect outliers in data streams[4]. But this method couldn't satisfy the accuracy, space and time efficiency in results [5] [6]. This paper identifies the metrices that can be used in outlier detection and proposes a generalized frame work that depicts the combination of distance-density-clustering based approach to detect outliers in the information set.

Figure:

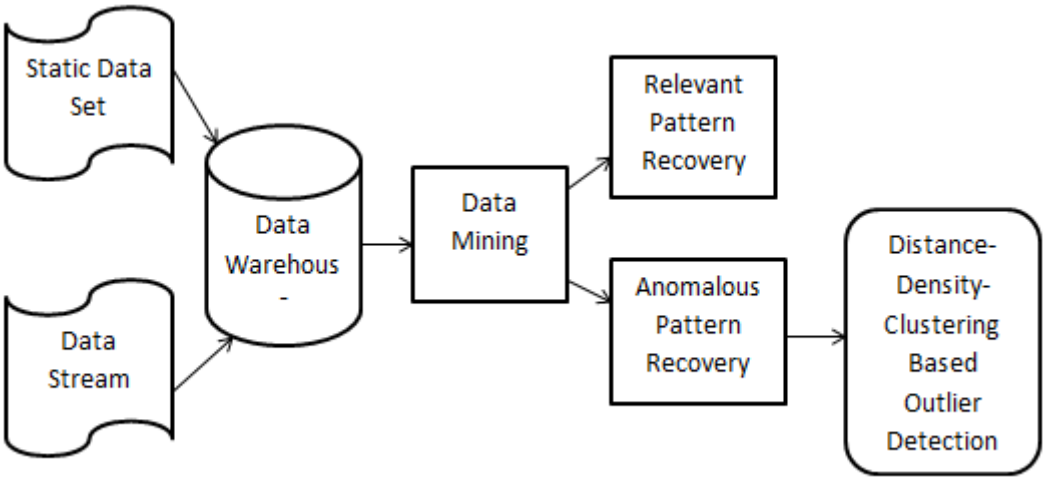

Figure 1 Data Mining: An Overview

This paper is organized as follows: Section II introduces the related works done so far. Section III identifies the design metrics and proposes an architectural design of the method. Section IV provides certain research directions and Section V concludes the paper.

2. RELATED WORKS

Most of the works on outlier detection is focused on static data sets in which all data are available in memory. Generally the outlier detection mechanisms can be classified into five categories. They are depth based, distribution based, clustering based, distance based and density based[7]. This is shown in figure 2.

Figure:

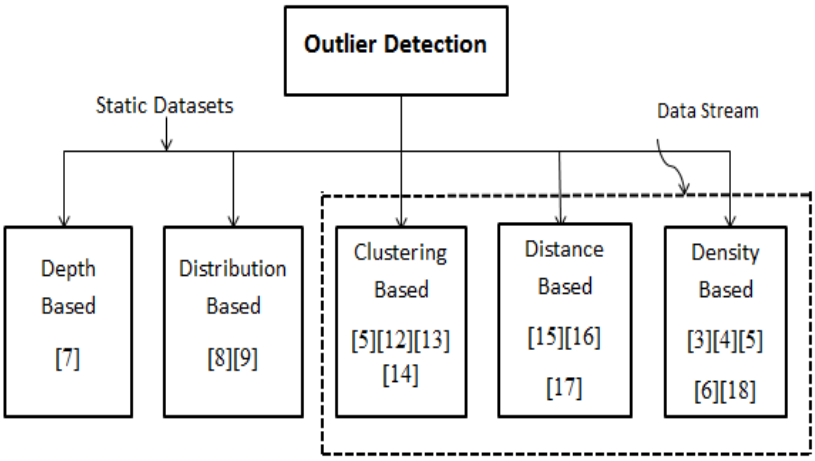

Figure 2 Classification of outlier detection

Depth based approach is an example of a model based approach independent of data distribution in which it arranged the data points in a convex hull [7]. Here the points which are lying on the border are considered as outliers and inliers are those that are concentrated on the center of the convex hull.

In distribution based outlier detection the points are arranged on the basis of a probabilistic model[8][9]. The main disadvantage of this technique is the lack of knowledge on the underlying data distributions.

The remaining categories are distance based, clustering based and density based, which can also be applied to detect anomalies from data streams [5]. Clustering techniques are used to build a cluster model on the basis of

the underlying data distribution[10][11]. Clusters having smaller density are interpreted as outliers and are omitted from the clustering model. It is focused on make clustering rather than detecting outliers [12] [13][14].
The distance based technique calculates the distance between the points in the data sets [15]. If an object is farthest from all the other points, it is considered as an outlier and all the other points are inliers. Distance based method have local and global variants. Local distance method can extract the outliers more effectively from stream data [16][17].

Density based method adopted a more effective way to search and clean the outliers [18]. Like distance based technique it has both local and global variants. It is based on the number of points around an object or the density of a cluster. The basic method adopted here is LOF(Local outlier factor),which finds out the nearest neighbor and the local reachability density of each point [6].

3. DESIGN METRICES AND ARCHITECTURAL DESIGN

This section depicts the metrices used in performance evaluation and an architectural design for performing the outleirness of an object.

The performance evaluation is interpreted in terms of three factors - ROC curve, run time and number of data points residing in the memory.ROC(Receiver Operating Characteristics) is plotted using true positive rate versus the false positive rate. The model can distinguish the points and assign the points to the appropriate classes.

Figure:

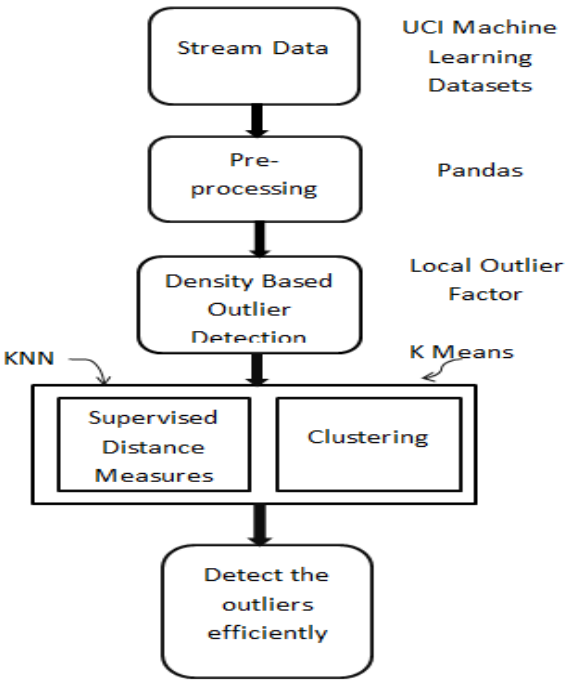

Figure 3 Architectural design

The stream data actually obtained from the data ware house contains missing and redundant values. These duplicate and missed values can be eliminated with the help of a pre-processing tool known as Pandas. Pandas is an open source data analysis tool available in the Python library. The actual method is performed using the combined algorithms of density based method (LOF),distance(KNN) and clustering-based method(K-means). The proposed architectural design is depicted in figure 3.
Terms used in Area under curve(AUC) and ROC:

Formulae:

TPR/Sensitivity = TP / (TP+FN) ,TP -- True Positive, FN -- False Negative (i)

Specificity = TN/(TN+FP) ,TN -- True Negative , FP -- False Positive (ii)

FPR = (1-specificity) (iii)

FPR = FP/(TN+FP) (iv)

4. RESEARCH DIRECTIONS

The outliers can be detected through the combinations of algorithms specified in distance, density and clustering based methods. As a research direction, the detection of anomalies in the data stream can be predicted with the help of a learning model. The actual input of the learning model is data stream and the output is the outliers detected by the learned model. Figure 4 gives an overview of the proposed model.

The data sets used are Lympho, Vowels, Letter Recognition and Pendigits from UCI machine learning datasets. Table 1 lists the number of points and the number of attributes used in these UCI data sets. The Lymphography data sets consist of 4 classes out of which 2 are quite small. The 2 small classes are merged and considered as outlier classes compared to other big classes. The Japanese vowel data set consists of nine male speakers, uttering two Japanese vowels /ae/ successively.

The letter recognition data sets identify the black and white rectangular pixels which denote the 26 capital letters in the English alphabet. Pendigits(Pen-Based recognition of Hand written Digits data sets) create a database consisting of 250 samples from 44 writers.

Figure:

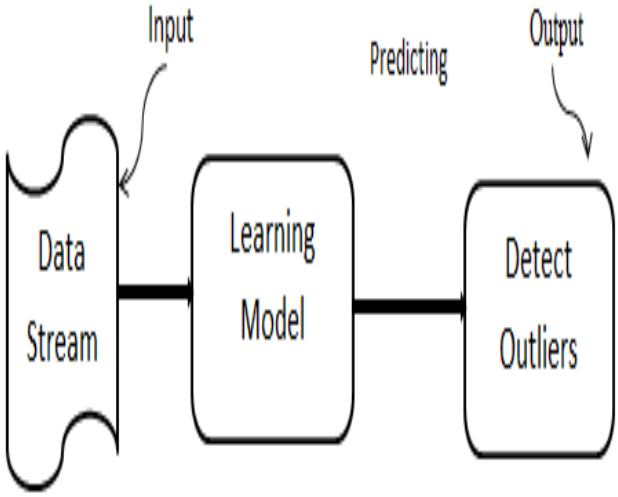

Figure 4 Research Directions

Table:

Table 1. UCI datasets

Data set	Number of points	Number of Attributes
Lympho	148	18
Vowels	1456	12
Letter Recoginition	1600	32
Pendigits	6870	16

5. CONCLUSION AND FUTURE WORK

The proposed method is more convenient and efficient approach for detecting the outliers from data stream as it is a combinations of distance, density and clustering based approaches. Supervised distance measures in the distance based technique gives more accuracy and lesser time complexity in results. Here it also studies the metrices used for detecting the degree of anomalous point in the data streams.

Detecting anomalous points from the data streams leads to better performance in many applications like video surveillance,wireless sensor networks, credit card fraud detection etc.

This fragment should obviously state the foremost conclusions of the exploration and give a coherent explanation of their significance and consequence.

As a future work, we can generate a learning model which will predict the outlierness of an object without the involvement of algorithms.

REFERENCES

[1] Sergio Ramírez-Gallegoa, Bartosz Krawczyk , Salvador Garcíaa, Micha Wozniak ,Francisco Herreraa,, "A survey on data preprocessing for data stream mining: Current status and future directions.," in Proc.ELSEVIER Neurocomputing, 2017.

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

[3] Shiblee Sadik,Le Gruenwald, "Research Issues in Outlier Detection for Data Streams," SIGKDD Explorations ,2015.

[4] Dragoljub Pokrajac,Aleksandar Lazarevic,Longin Jan Latecki, "Incremental Local Outlier Detection for Data Streams," in Procthe 2007 IEEE Symposium onComputational Intelligence and Data MiningYear: 2007

[5] Mahsa Salehi, Christopher Leckie, James C. Bezdek,Tharshan Vaithianathan ,Xuyun Zhang,"Fast Memory Efficient Local Outlier Detectionin Data Streams",," in ProcIEEE Transactions on Knowledge and Data Engineering (Volume: 29, Issue: 3, March 1 2016

[6] M. Salehi, C. Leckie, J. C. Bezdek, T. Vaithianathan,"Local Outlier Detection for Data Streams in SensorNetworks: Revisiting the Utility Problem",in proc.IEEE International Conference on IntelligentSensors, Sensor Networks and Information Processing,2015 April

[7] Sreevidya S S,"A Survey on Outlier Detection Methods",in proc.(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (6) , 2014, 8153-81565

[8] K. Yamanishi, J.-I. Takeuchi, G. Williams, and P. Milne," Onlineunsupervised outlier detection using finite mixtures withdiscounting learning algorithms",in SIGKDD,2000

[9] K. Yamanishi , J.i. Takeuchi,"A unifying framework for detectingoutliers and change points from non-stationary time seriesdata",in SIGKDD,2002

[10] Dr. T. Christopher, T. Divya ,"A Study of Clustering Based Algorithm forOutlier Detection in Data streams",in proc the UGC Sponsored National Conference on Advanced Networking and Applications, 2015.

[11] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu,, "A framework forclustering evolving data streams," in ProcVLDB,2003}

[12] S. Guha, A. Meyerson, N. Mishra, R. Motwani, and L. O'Callaghan,"Density-based clustering over an evolving data stream with noise",in SIAM Conference on Data Mining,IEEE Transactions on Knowledge and Data Engineering,2003

[13] Dr. S. Vijayarani1 ,Ms. P. Jothi,"Detecting Outliers in Data streams using Clustering Algorithms",in proc International Journal of Innovative Research in Computer and Communication Engineering ,2013

[14] G. Pyun and U. Yun," Clustering data streams: Theory and practice",in proc. Applied Intelligence,2014

[15] Edwin M. Knox , Raymond T. Ng,," Algorithms For Mining Distance-Based Outliers in Large Data sets", in Proc of the 24th VLDB Conference New York, USA, 1998

[16] Biao Huang ,Peng Yang," KNN Based Outlier Detection Algorithm in Large Dataset", International Workshop on Education Technology and Training and International Workshop on Geoscience and Remote Sensing,2008.

[17] Edwin M. Knorr1, Raymond T. Ng,Vladimir Tucakov, and V. S. Tseng,"A Study of Clustering Based Algorithm for Outlier Detection in Data streams", The VLDB Journal,2000

[18] Sridhar Ramaswamy,Rajeev Rastogi,Kyuseok Shim," Efficient Algorithms for Mining Outliers from Large Data Sets", Advanced Information Technology Research Center at KAIST, 2000