Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 7– No. 24, December 2018 – www.caeaccess.org

Efficient Model to Reduce False Positives using Outliers Detection in Big Data

Esraa Samir Ahmed
Master Student
Faculty of Computers and Information
Helwan University

Laila A. Abd-Elmegid
Assistant Professor
Faculty of Computers and Information
Helwan University

Hala Abdel-Galil
Associate Professor
Faculty of Computers and Information
Helwan University

ABSTRACT
Emerging fields like Internet of Thing (IoT), sensor data, mobile computing, and social media are driving new forms and sources of data with distinct features. Big data is the term used to describe such type of data. Analytics of big data aims to use advanced analytic techniques on very large data sets that are collected from different sources in different formats. Mining anomalies from big data is a powerful mining task that is used mainly in critical systems. Applications work with big data require efficient outliers detection system. Outlier detection system in big data need to be efficiently designed to cope with its distinct features of volume, speed, complexity, and variety. The huge volume of outliers detected in big data is a barrier for outlier's diagnosis and analysis. Due to the cost of analysis each anomaly, outlier detector needs to be accurate as possible. Minimization of false positive alerts is a key feature that increases the accuracy of the detector system. This paper present a new propose model for reducing the false positive alerts using outliers detection. The proposed model uses cluster analysis by DBscan algorithm to highlight the outliers and then validates these outliers and reduces false positive alerts using Support Vector Machine (SVM) algorithm. The experimental study proves the efficiency of the proposed model with reported accuracy equals to 99%.

Keywords
Big Data; DBscan; False Positive; SPARK; SVM

1. INTRODUCTION
Big Data is a term used to describe large and/or complex data sets where traditional data processing techniques are inefficient. Big data requires special techniques that can be used to store and process the exponentially increasing data sets which contain structured, semi structured and unstructured data. Big data does not only mean that the volume of data is big. There are other main characteristics of big data including velocity in which the data is changed and created rapidly and variety that refers to data stored in multiple formats. Volume, variety, and velocity present big data 3vs characteristics. The 3vs is extending over time by adding more distinct characteristics of big data. Many fields of science and engineering are utilizing big data technologies, including biological, biomedical, and physical sciences [1].

One of the most important analysis processes that performed on big data is the anomaly detection process. Anomaly is a data object with behaviour that are very different from expectation. Anomaly is also called outlier. Clustering analysis is a popular technique used for the purpose of outlier detection. Clustering aims to group data points with similar characteristics together in clusters [2]. Consequently, the remaining data points that have not be assigned to any cluster are marked as anomalies or outliers.

False positive alert is identifying normal data point as outlier and generating alert by the detector system. Security system overwhelmed with large number of false positives can create major problems and subsequently the detector system will lose its credibility.

this paper present a new propose model that combines clustering and classification techniques to increase the accuracy of anomaly detection process and decreases false positive alerts.

The paper is structured as follows: Section Two presents the cluster analysis as one of the methods for outlier detection. A literature review is given at Section Three. Whereas, Section Four illustrates in details the proposed model. The experimental study is shown in Section Five. Finally, a conclusion of the research work is summarized at Section Six

2. CLUSTER ANALYSIS-BASED OUTLIER DETECTION
One of the popular methods of outliers detection is using the clustering technique. In such approach data points are clustered into set of clusters according to the specified similarity measurement. Consequently, data points with different characteristics are not assigned to any cluster and remarked as outliers [3].

There are two famous clustering-based outlier detection methods; distance-based method and density-based method.

1. The distance-based outlier detection method is depending on the distance between the data points. In this method a parameter called radius is specified. For any data point, if this point is surrounding by a set of neighbours within the radius, it is a normal data object. Otherwise, it is marked an outlier.

K-means algorithm is one of the distance-based clustering algorithms. K-means clustering is a type of unsupervised learning that is used with unlabelled data. The objective of the algorithm is to classify data into groups having the number of groups given a prior (variable K). The algorithm implements an iterative procedure to assign each data point to one of K clusters. Feature similarity is the criteria used in the clustering process.

2. The density-based outlier detection method.is depending on measuring the density of the data points. It defines clusters as dense regions of data

points in the space surrounded by low density regions of data points.

DBSCAN is a density-based clustering algorithm which is used to discover clusters of arbitrary shape. DBSCAN groups data points based on a two parameters; a distance measurement (Radius) and a minimum number of points.it categorizes the data objects according to these parameters to three categories; Core point: that is a point has more than the specified minimum number of points within the radius, Border point that has less than the specified minimum number of points but it exists in the neighbourhood of a core point; Outlier point which is neither a core point nor a border point.

In general, k-means algorithm suffers from high false positive alerts. However, in dynamic environment, the problem getw worse as it can not efficiently discover outliers according to its processing mechanism. The algorithm starts with clustering the existing data points into clusters according to the specified similarity measurement. When a wave of new data points comes, the k-means algorithm analyze only these new data points and compare their features according to the features of the previously developed clusters. It does not add any new clusters. Consequently, if it identified a data point as an anomaly in the first phase and there are other data points similar to its features in the new data set, the k-means algorithm will consider the new data points as anomalies also instead of creating a new cluster for these data points.

DBscan algorithm overcomes the main problems of K-means algorithm as it results in smaller false positive alerts. Also, it works efficiently in dynamic environment through repeating the clustering process for the complete data set for each wave of new data. As a result, if there are data points with similar features exist, new cluster will be created [2,3].

Figure 1 shows how the two algorithms work in dynamic environment as described in the previous sections.

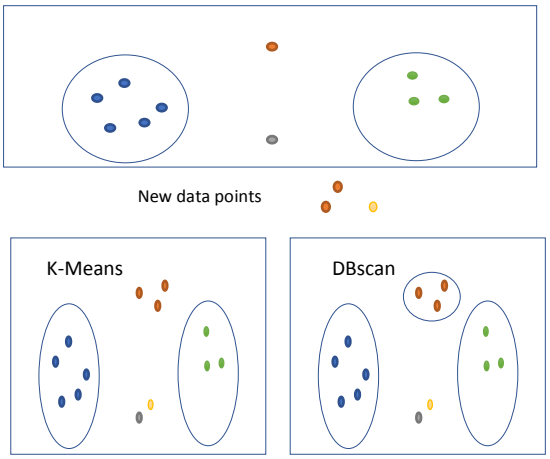

Figure 1: K-Means and DBscan Algoirthms

3. LITERATURE REVIEW

The work by YADIGAR [5] proposed a model for outlier detection using k-means and distance-based algorithms. The model has been applied on Spark environment. The experimental study reported faster processing time of the proposed model when comparing it with The Gaussian Mixture Model (GMM). However, regarding to the accuracy, both models result in the same number of clusters and similar outliers record.

Laura et al [6] introduced a new system for outliers detection in big data. the system utilizes two techniques; relative Entropy and Pearson correlation. The study implemented the system on Spark and ran it on real data sets. The experiments showed that relative entropy is best used for detecting gradual changes. However, Pearson correlation is best used for detecting sudden changes.

A research for context-aware anomaly detection in embedded systems introduced in [7]. In this work the authors stated that all detection methods designed for embedded systems cannot be applied on dynamic environments because all of them don't recognize the context changes and consequently they generate false alarms. The research proposed a system that can recognize the context changes react to the changes in the environment. Firstly, it identifies the context of a test data stream by using KNN clustering algorithm and feeds this context test data stream into the detector. Next, it loads the corresponding probability matrix into the detector. The context is assumed to be constant until an alert is generated. Once an alert is generated it identifies the context of the new incoming test subsequence. Finally, according to the context of the new incoming test subsequence, the generated alarm is marked as outlier or ignored.

Another work for context-aware anomaly detection presented by Michael [8]. It proposed a new system that for contextual anomaly detection in big data. Working with big data poses challenges due to the huge volume and high velocity of such data. Accordingly, it is too hard and expensive to apply collective anomaly detection algorithms on big data. For this reason, the proposed system tends to use a point anomaly detection algorithm which is fast but less accurate to detect anomalies and then applies on these anomalies only the collective anomaly detection algorithm which is more accurate but more computationally expensive. The collective anomaly detection algorithm will determine whether the anomaly was contextually anomalous or not. This approach allows the algorithm to be suitable with the big data requirements because all data will be detected by a fast, non-expensive algorithm "point anomaly detection algorithm" and only small set of data will need to be detected by the more expensive algorithm " collective anomaly detection algorithm".Examples of application-oriented research for outliers detection in various fields include [9], [10], and [11]

4. PROPOSED MODEL

This section explains in details the framework of the proposed model to efficiently mine outliers from big data. The main objective of the proposed model is to increase the accuracy of the results by minimizing the false positive errors. Accordingly, the model utilizes the clustering and the classification techniques together. The main idea of the proposed model is to use a clustering algorithm to highlight the clusters of the normal data points and the outliers and then validate the outliers by a classification algorithm. The model consists of three main phases as shown in Figure 2.

Phase 1: Data pre-processing

As agreed in the data mining community: No quality data, No quality mining results. The pre-processing of data is an essential step to guarantee the reliability of the mining results. The main activities of this step include: data cleaning, data integration, data transformation, data reduction, and data discretization. this proposed model focus primary on two activities; data cleaning to fill missing values, remove noisy

data, and resolve inconsistences, and data reduction to obtain reduced data sets with the same characteristics of the original data set. Sampling is the implemented technique for data reduction used by the proposed model.

Phase 2: Cluster analysis-based outlier detection

Clustering algorithms can detect outliers as a by-product of the clustering process. The Clustering algorithms works on the principle that object that is not assigned to any cluster should be considered as outlier. As previously mentioned in Section 2, DBscan algorithm offers distinct features over K-means algorithm including; a smaller number of false positive anomaly alerts and no required predefined attributes (number of clusters). The proposed model uses DBscan algorithm for clustering analysis. The outputs of this phase are the set of clusters of normal data and the set of outliers.

Phase 3: Support Vector Machine

Support vector machine (SVM) is a classification algorithm that uses a line called hyper-plane to separate classes. This hyper-plane can be identified as follow; First the data points are plotted on the space then a set of lines called "hyper plane" are drawn to separate the classes. The objective is to select the best hyper plane that separate the classes. To do so, it's important to sure that this line actually separates the classes. If there are several lines that separate the classes, the line has the maximum distance from the nearest class or data point is selected. The distance between hyper-plane and its n earest point is called margin. The proposed model uses SVM as a classifier algorithm to validate the outliers resulted from phase 2. The process is performed by using the clusters of normal data points as a training set and the outliers as a testing set. By using SVM, the number of false positive anomaly alerts is reduced and consequently leading to more accurate results.

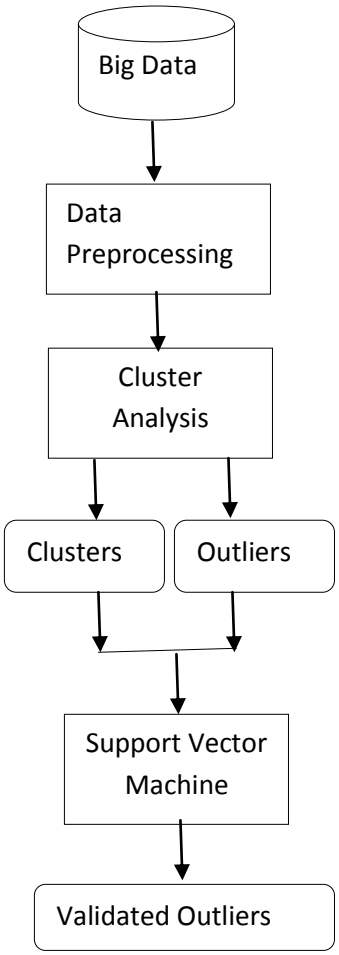

Figure 2: The proposed model for accurate outliers' detection from big data

5. EXPERIMENTAL STUDY

The aim of the experimental study is to prove that the proposed model accurately mine outliers from big data better than the well-known DBscan algorithm. The experiments conducted on a machine with the following configurations:

GPU:1xTesla K80, compute 3.7, having 2496 CUDA cores, 12 GB GDDR5 VRAM; CPU: 1xsingle core hyper threaded Xeon Processors@2.3 Ghz; RAM: 12.6 GB; Disk: 33 GB on SPARK environment

The used dataset, individual household electric power consumption, [4] contains 2075259 records gathered between December 2006 tell November 2010 (47 months) for power consumption of a house in Paris. At the beginning, the first phase of the proposed model for data pre-processing is applied. All the issues related to missing data, invalid attributes, and inconsistencies have been resolved. The cleaning process results in 2049280 valid records. The representation of the data set is showed in Figure 3. The data reduction process is done through sampling the data. Four different samples have been selected to conduct four experiments. The information about each sample exists at Table 1.

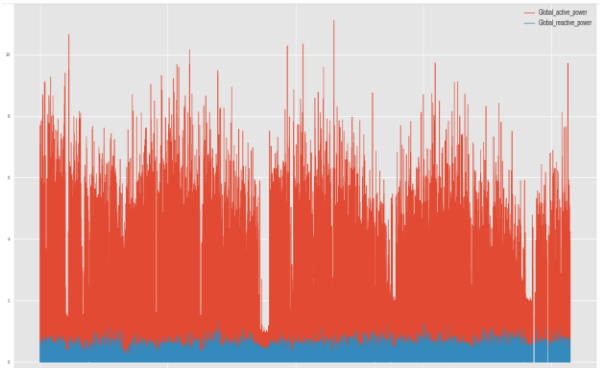

Figure 3: Representation OD the data set

Table 1: Information of samples of the data set

Sample no.	No. of records	Criteria of sampling
1	68324	A sample every 30 minutes
2	81990	A sample every 25 minutes
3	102485	A sample every 20 minutes
4	136639	A sample every 15 minutes
5	146399	A sample every 14 minutes

The first set of experiments apply the DBscan algorithm on the four samples of data to discover the outliers at each sample. The results of the four experiments are represented in Figure 4. Figure 5, Figure 6, and Figure 7

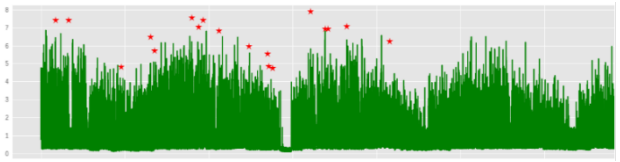

Figure 4: Applying Dbscan A lgoirthm on Sample #1

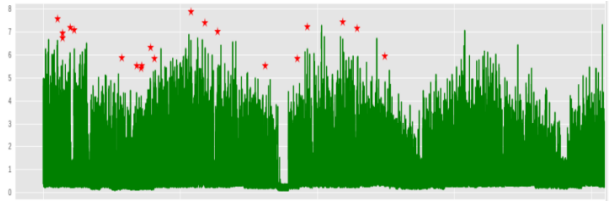

Figure 5: Applying Dbscan A lgoirthm on Sample #2

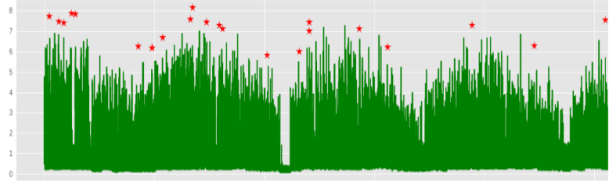

Figure 6: Applying Dbscan A lgoirthm on Sample #3

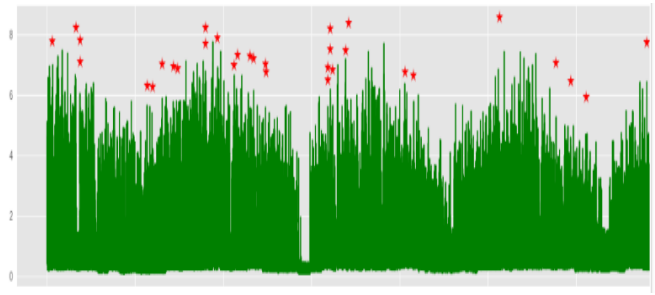

Figure 7: Applying Dbscan A lgoirthm on Sample #4

The next set of experiments process the output of applying DBscan algorithm on each sample and validate it using SVM as structured on the proposed model. The results show that using SVM has a marked effect on the reduction on number of false positive alerts with accuracy reported from SVM equals to 0.99 for all experiments. The comparison between the number of outliers of each sample using DBscan algorithm and the proposed model is represented in Table 2.

Table 2: Performance of the Proposed Model

Sample No.	No. of outliers (DBscan)	No. of outliers (Proposed Model)
1	18	3
2	20	5
3	23	12
4	32	12
5	36	13

6. CONCLUSION

In this study, the techniques of clustering and classification have been utilized together for efficient outlier detection on big data with minimal false positive alerts. DBscan algorithm has been chosen for clustering the data points and highlighting the proposed outliers while Support Vector Machine algorithm is used for validating the resulted outliers and remove false positives. All the experiments proved the efficiency of the proposed model with accuracy equals to 99%.

7. REFERENCES

[1] Kaufman, L., & Rousseeuw, P. J. 2009. Finding Groups in Data: An Introduction to Cluster Analysis (Vol. 344). John Wiley & Sons, United States.

[2] Min Chen, Simone A. Ludwig, and Keqin Li, Clustering in Big Data , “K29224_C016” — 2017/1/12

[3] Han, Jiawei, and Micheline Kamber. Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann Publishers, 2001.

[4] Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[5] YADIGAR ERDEM, CANER OZCAN "Fast Data Clustering And Outlier Detection Using K-Means Clustering On Apache Spark" International Journal of Advanced Computational Engineering and Networking, ISSN: 2320-2106, Volume-5, Issue-7, Jul.-2017

[6] L Rettig, M Khayati, P Cudr-Mauroux et al., "Online anomaly detection over Big Data streams", 2015 IEEE International Conference on, pp. 1113-1122, 2015.

[7] Ehsani-Besheli, F., Zarandi, H.R.: Context-aware anomaly detection in embedded systems. In: Zamojski, W.,Mazurkiewicz, J., Sugier, J., Walkowiak, T., Kacprzyk, J. (eds.) DepCoS-RELCOMEX 2017. AISC, vol. 582, pp. 151–165. Springer, Cham (2018).

[8] M. A. Hayes, M. A. Capretz, "Contextual anomaly detection framework for big sensor data", Journal of Big Data, vol. 2, no. 1, pp. 1-22, 2015.

[9] Janković, S., Zdravković, S., Mladenović, S., Mladenović, D., & Uzelac, A. (2016). The Use of Big Data Technology in the Analysis of Speed on Roads in the Republic of Serbia. Proceedings of the Third International Conference on Traffic and Transport Engineering (ICTTE Belgrade 2016), Belgrade, Serbia, 219-226.

[10] Lee, W-J. (2017) 'Contextual air leakage detection in train braking pipes', in Benferhat, S., Tabia, K. and Ali, M. (Eds.): International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2017: Advances in Artificial Intelligence: From Theory to Practice, pp.191–200, Springer, Cham, Arras, France.

[11] Berhane Araya, Daniel, "Collective Contextual Anomaly Detection for Building Energy Consumption" (2016). Electronic Thesis and Dissertation Repository. 4027