Anomaly detection in text data that represented as a graph using dbscan algorithm

ANOMALY DETECTION IN TEXT DATA THAT REPRESENTED AS A GRAPH USING DBSCAN ALGORITHM

ASMA KHAZAAL ABDULSAHIB

University of Baghdad, College of Education for Human Science-Ibn Rushed, Baghdad, Iraq.
E-mail: h_asamaa@yahoo.com

ABSTRACT

Anomaly detection is still a difficult task. To address this problem, we propose to strengthen DBSCAN algorithm for the data by converting all data to the graph concept frame (CFG). As is well known that the work DBSCAN method used to compile the data set belong to the same species in a while it will be considered in the external behavior of the cluster as a noise or anomalies. It can detect anomalies by DBSCAN algorithm can detect abnormal points that are far from certain set threshold (extremism). However, the abnormalities are not those cases, abnormal and unusual or far from a specific group, There is a type of data that is do not happen repeatedly, but are considered abnormal for the group of known. The analysis showed DBSCAN using the improved algorithm can detect this type of anomaly. Thus, our approach is effective in finding abnormalities.

Keywords: *Anomaly Detection, Enhanced DBSCAN algorithm, Unsupervised anomaly detection and Concept Frame Graph (CFG).*

1. INTRODUCTION

Anomaly detection is an imperative issue being addressed by many researchers inside differing research regions and application areas. Numerous anomaly detection methods have been particularly created for Particular application areas, Others are non-specific. Abnormal alludes to issue of the discovering models in the data that don't adjust to the normal conduct. These non-accommodating models are regularly appointed as anomalies, dissonant perceptions, exemptions, distortions, outliers, idiosyncrasies or contaminants in various application areas. These outliers and anomalies are two terms most utilized ordinarily with regards to detects anomaly. Find Anomaly detection in the widespread use in a variety of uses, for example, fraud discovery for insurance, health care or for credit cards and military surveillance of enemy activities, Intrusion detection for Cybersecurity, and fault detection in integrity of critical systems, The significance of the anomalies is due to the fact that the anomalies in the data translate into a major operation information (and decisive in many cases) in a wide range of application areas.

Detect anomalies or extreme values in the data that studied in the community statistics at the beginning of the century 19th [Edge worth 1887] After a period of time, it has developed several methods miscellaneous for detection of anomalies in many research societies techniques. Several of these technologies were developed precisely in some application areas, while others are more general.

In this study, we address a type of case that has not been addressed in previous studies. However, the abnormalities are not those cases, abnormal and unusual or far from a specific group, There is a type of data that is do not happen repeatedly, but are considered abnormal for the group of known. These cases, which do not occur frequently, will be addressed in this research using the clustering algorithm called enhanced DBSCAN. Where enhanced DBSCAN algorithm can detect abnormal points that are far from certain set threshold (extremism). The analysis showed DBSCAN using the improved algorithm can detect this type of anomaly. Thus, our approach is effective in finding abnormalities.

1.1 What Are Anomalies?

Anomalies are patterns in text that don't fit in with a very much characterized idea of ordinary conduct. From figure1 can notice the information has two ordinary districts, D1 and D2 as most perceptions located in these two areas. Points that are far from the areas, e.g., points A1, A2, and points at district A3, are anomalies. Anomalies may be instigated in the text for an assortment of reasons, for example, pernicious action, e.g., credit card extortion, digital interruption, fear based oppressor action or breakdown of a framework, yet the majority of the reasons has a common feature that is it fascinating to the analyst. The "intriguing quality" or genuine importance of anomalies is a key component of inconsistency recognition. Figure 1 shows anomalies in a data set.

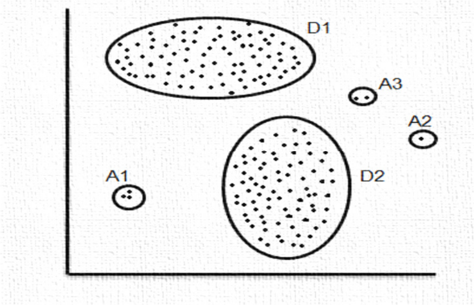

Figure 1: illustrate anomaly

Detect anomalies systems compared to activities with "ordinary standard". Anomaly detection systems have two major advantages. The advantage preferred is their ability to recognize obscure assaults since they can display the typical operation of a system and distinguish deviations from this model. The second advantage is the customization capacity of the normal dynamic profiles for each framework, application, and system. However, the anomaly detection approach has its drawbacks like the complexity of the system, false alarms high and the difficulty of detecting such an event, which are being alarmed [1].

In this study using a new method to enhance the performance of the algorithm where apply of the many technical challenges that have to be handled before the adoption of anomaly detection systems. The new strategy, including change over content to a graph before applying unsupervised methods (DBSCAN) for dealing with the issue of anomaly detection. DBSCAN is a critical algorithm since what is required is just one scan over the database. What's more, it doesn't require any Predefined cluster number to be worked. Likewise, DBSCAN is built clusters which utilize separate transitivity in view of a density measure characterized by the client. Dens are archives that have numerous co-went to documents around them. What DBSCN do is take advantage of a fixed threshold value to determine areas "dense" in the space of the document. This method can not identify a dense and free points as a result of continuous appreciation of the limit on every point in space. Therefore, it is often collected the all document space in one cluster.

The density-based method very successful in the implementation. It doesn't require the user to identify the number of clusters. Another has a major favorite feature of this application. In light of the possibility that in a space object, must be collected dense objects together in one cluster, the establishment of methods depends on the density. And accordingly, a cluster characterized as a region has the highest density of points from its neighboring points. For any point in space, where the related points. Page on the Internet, and the intensity gets to be significantly higher when there are many pages that share to happen with it.

1.2 Detection Anomaly In Data

The procedures of anomaly detect in this area principally identify innovative points or a news articles, events or news in a set of documents. The reason of anomalies Because of the new wonderful topic or case abnormal. The data in this area are usually very sparse and high dimensional. Additionally, the data contain the temporal side since gathered the documents on the passage of time. There is a challenge to the techniques of detection of anomalies in this area to deal with the significant differences in the documentation affiliated to a single category or one topic [1].

This paper presents one of the unsupervised techniques (DBSCAN) where the procedure of this exploration incorporates two phases. Principal phases is pre-processing documents, or changing the text into a proper valuable data and afterward convert all documents to graph called Concept Frame Graph (CFG). The second phase DBSCAN is used to detect abnormalities and analyze the results using the measurement evaluation. Then compare the result with the same algorithm without strengthening the algorithm to find out whether the text represented as a graph affect the work DBSCAN algorithm to detect anomalies.

In DBSCAN, the clustering methodology is not the same as run of the typical clustering approaches. Can identify the outlier points (abnormal) that do not lend themselves or not similar to any clusters by the DBSCAN. The results of the cluster approach represent a group that takes the earnings per share as a measure of cluster distance threshold. DBSCAN requiring an addition to distance metric, a minimum number of points minpts within the same group to denote them a cluster. For instance, in figure 2 if the minpts value is specified as 3, 1, 2, 4, and 6 will be set apart as clusters utilizing DBSCAN. On the other hand, 3 and 5 will be characterized as outliers because the cluster must contain the appropriate number of points for its formation. Correspondingly, If have been determined the values of minpts as 5, then group 3, group 6, and group 5 will be appointed as anomalies utilizing DBSCAN.

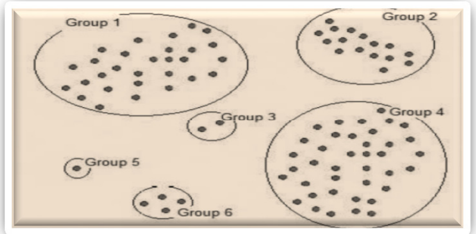

Figure 2: A Sample Data Set For Distance-Based Anomaly Detection Approach

This paper presents one of unsupervised methods (DBSCAN) where the process of this research includes two stages. The first stage is pre-processing the documents, i.e. transforming the documents into an appropriate useful data representation and then enhanced DBSCAN algorithm by parsing data and convert all documents to graph called Concept Frame Graph (CFG). The second stage uses DBSCAN to detect anomalous and analyzing the results using evaluation measurement. Then compare the result with the same algorithm without enhanced algorithm to see which algorithm better to reduce the problem of anomaly problem.

2. RELATED WORK

Detect anomalies is a test field that utilizations different strategies for a variety of applications. Since the writing on detecting anomalies is extensive exceptionally, in this section we provide a survey of the previous reviews that need to be done with DBSCAN algorithm specifically and unsupervised algorithms in general and its effect on the detection of anomalies. It provides [3] strategies for the detection of the anomaly based on the graph. What is more, provide another method to detect the consistency of the graph, with applications to detect abnormalities. Where in this study, the researchers assumed these strategies will demonstrate it is helpful both to detect abnormalities and to decide on the possibility of fruitful anomalies within the graph. They give the test comes around to take advantage of all the real-world data network intrusion and data artificially.

Where [11] give a broad study of anomaly detection procedures created in machine learning and measurable domains. A wide audit of anomaly discovery methods for numeric and typical data is introduced in this review [10]. Additionally, the specialists have examined anomaly based instruction detection, advantages and disadvantages of abnormality discovery, unsupervised and supervised anomaly detection. What's more, the outcomes demonstrate the unsupervised give preferred outcomes over supervised methods [4].

In [1] the researchers talked about various routes in which the issue of anomaly. Anomaly Detection Identification, sorts of anomaly detection, different methods of abnormality discovery. Their points of interest and drawbacks. And furthermore examined Clustering Based Algorithm. Where using DBSCAN to recognize anomalies in time series data and compared it with the measurable anomaly detection method. The outcomes indicate utilizing DBSCAN algorithm give the great outcome in this aspect.

Also in [5] Concentrate on the revelation of anomalies in a monthly temperature information utilizing DBSCAN method. DBSCAN method is a density-based clustering algorithm that has the ability to find abnormal information. In the empirical assessment, looked at the effects of the DBSCAN algorithm with the consequences of a measurable technique. The examination demonstrated that DBSCAN has a many preferences over the statistical approach on finding abnormalities.

View [2] a novel engineering for an intrusion detection framework in the artificial immune system. At the point, when the innate immune suggested utilizing unsupervised machine learning techniques. As indicated by analyses and think about between more one algorithm presume that among different strategies, DBSCAN clustering is powerful and has the best potential for this reason.

3. MATERIAL AND METHODS

The system proposed for clustering data representation following these steps: First, must have been collecting data where the data set that is utilized as a part of this research is a collection of KDD Cup group 99 data, which is extracted from the network traffic DARPA 98. The quantity of test of data set was 22545, which was enough to evaluate and compare the performance between the enhancing DBSCAN using the graph and the same algorithm without the graph [2]. What's more, dispose of the undesirable words through a few procedures it's called the pre-processing. The main objective of the preprocessing stage is to dispose of the phrasing and the pointless words from the documents that we have, and in this manner get the words or essential components in documents, the documents are prepared as follows: Tokenization method: Segmentation the sentences or phrases into several parts (Tokens), commonly words. More advanced techniques, derived from the field of NLP, linguistic structure of the content analysis of the choice of terms or pieces (the succession of words, for example, noun phrases [8]. Then applying Stemming Algorithm where all documents stemming using porters algorithms Porter, 1980 [12]. Stemming is a familiar technique which is used in text mining research. It is done by omitting the suffixes of a word and then bringing it back to the original root because stemming decreases the word complexity without causing great loss of information for typical applications (especially for bag-of-words).

3.1 Anomaly Detection Techniques

The method using to the enhanced DBSCAN algorithm, including many phases when must in the beginning parsing the all documents where the parsing algorithms means produces a tree of sentences correlated to each other where the relationship between words per sentence and the relationship with the rest of the sentence can be found. The aim of the parser is to analyze the input sentence and the output corresponding parse tree (most preferred). In this research, Standard English Grammar rule [9] is used to obtain word dependencies. Is dependent on a type of grammar which is declarative formed in order to be calculated in many possible ways. After complete parsing the output become as matrix means create an (n-1) × (n-1) matrix as a list of lists in Python, and then [7] converts the documents are parsed to graph. There are several types of graphs in Text Representation Schemes. In this research represent text using graphical schemes such as Concept Frame Graph (CFG), where representation of every document as a graph, the nodes compatible to words which could be seen as meta-descriptions of the document, while the edge representation the relations between pairs of words. Where the previous study [6] indicates that the texts represented as a graph improves the overall quality of performance. In this step introduce the technique for detection anomaly based a graph. The goal of anomalous substructure detection is to study on the entire graph, and to submit a report on the foundations of an unusual sub graph inside all graph. We consider that the best sub graph to be the one that minimizes the following value:

$$F1 (S, G) = DL (G | S) + DL (S)$$

Where G is the entire Concept Frame Graph (CFG), S is the sub graph; DL (G|S) is the description length of G after compressing it using S, and DL (S) is the description length of the sub graph. Figure 3 represents an example of a graph.

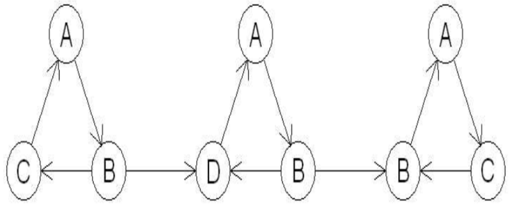

Figure3: Concept Frame Graph (CFG)

When each substructure appears twice, as in the example above this leading to generate and evaluate all substructures, this substructure will be ranked as the best. If another iteration will be run, then will first compress the graph by replacing the instances of this substructure with a new vertex. As with many algorithms involving graphs. After creating documents using the new method to represent the text is now applying the DBSCAN algorithm to detect the anomalies.

3.2 Dbscan Algorithm

DBSCAN is a spatial aggregation on the basis of density which can identify anomalies in the data series algorithm. It requires two parameters of knowledge, they are as follows: a minimum of points minpts and a neighborhood distance epsilon (EPS). For a given point, it is called a point in the distance EPS Neighbors of that point. In the case that a number the neighboring points greater than minpts, this gathering of points is known as a cluster. The fundamental data in DBSCAN called core points, border points for the other points and anomalous points for outliers. Core points are those that have in any event minpts number of points in the distance EPS. The points are not fundamental points, but neighbors of core points called the border points. While the points that are not the core or the border points called abnormal points. DBSCAN algorithm is "density-based clustering algorithm". Its preference is that it can detect clusters with discretionary shapes besides; it can deal with noise also. The algorithm normally considered the clusters, as the dense areas of items in the information space which are isolated by areas of low density objects. Two of the input parameters at the algorithm, MinPts and ε. For comprehension the procedure of DBSCAN, must be presented the how this algorithm works in data represented as a graph.

The algorithm starts its work as following: It checks the ε-neighborhood of every point (sub graph) of the dataset, and if around this region a bigger number of objects than MinPts exist then it is known as a core object. Every cluster is developed from a core object by gathering those points that are straightforwardly density-reachable from the core point. The algorithm is ended if there is no more points exist that can belong to a cluster. These objects are dealt with as noise that couldn't be allocated to any cluster by the algorithm. Figure 4 displays the algorithmic work.

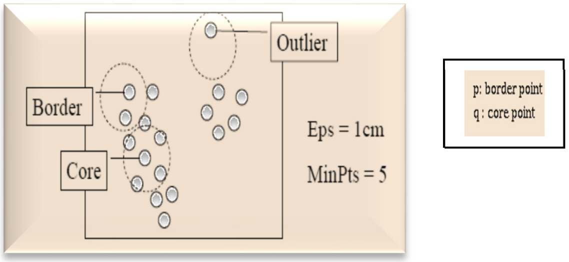

Figure4: DBSCAN algorithm

4. RESULT DISCUSSION

The outcome is investigated utilizing the Precision, Recall, F-measure and Accuracy as shown in (equ. 1, 2, 3, 4). Table 1 shows the Evaluation result of detection anomaly text represented as a graph using DBSCAN compared with the previous study [2] using the same algorithm to detect anomalous natural text without representing text as a graph.

" $\boldsymbol{Precision(P) = TP/TP + FP}$ (1)
$\boldsymbol{Recall(R) = TP/TP + FN}$ (2)
$\boldsymbol{F\ Measure\ (F) = \ 2 * P * R/P + R}$ (3)
$\boldsymbol{Accuracy\ (A) = TP + TN/TP\ + FP + FN + TN}$ (4)"

Where:

"TP is true positive: case was positive and predicted positive
TN is true negative: case was negative and predicted negative
FP is false positive: case was negative but predicted positive
FN is false negative: case was positive but predicted negative"

Table 1: Shows The Evaluation Result Of Detection Anomaly

Methods	Dataset	Precision	Recall	F-measure	Accuracy
Enhance DBSCAN	KDD-Cup 99 dataset	0.965	0.939	0.926	0.940
DBSCAN	KDD-Cup 99 dataset	0.987	0.589	0.738	0.771

From the table can notice that the value of recall when clustering documents utilizing enhanced DBSCAN algorithm are higher compare with DBSCAN directly without enhancing algorithm as 0.939,0.589 respectively. This means that the accuracy value is high. Also the rate of f-measure is more in case using enhanced DBSCAN where 0.926 while in another case is 0.738.

Finally, the value of accuracy is high if the enhanced algorithm as shown in the table is used where its value is 0.940. As indicated by the outcomes are appearing in table 1, the DBSCAN is ideal for identifying anomalous when the content represented as a graph before utilizing this algorithm. Figures 5 introduce the evaluation values for this study compare with the same algorithm without graph through a bar graph. The DBSCAN method can detect anomalous points which are far the cluster a specific edge in sub graph (extremes). In any case, anomalous points are extraordinary points, means the data that doesn't occur frequently also called abnormal points. So, the contribution of this study lie when using the Enhanced DBSCAN algorithm can discover these kinds of anomalies. That is why we can say that according to the results of our approach is effective in detecting these cases of anomalies.

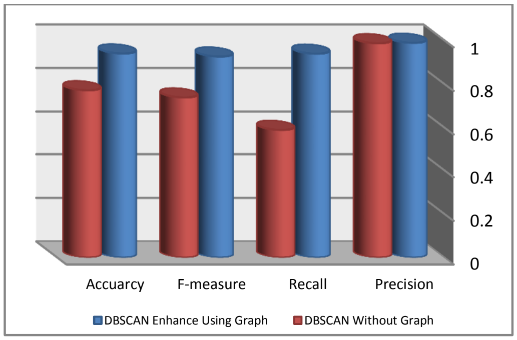

Figure 5: Evaluation Values Of DBSCAN Algorithms

6. CONCLUSION AND FUTURE WORK

This paper has investigated the use of the enhanced DBSCAN algorithm in the anomaly detection to handle situations that do not occur frequently. We have applied enhanced DBSCAN to detect anomalies in text data, where applied a preprocessing step to remove seasonality and then convert all documents to graph called Concept Frame Graph (CFG) then applying DBSCAN algorithm on the data set and compared it with the same algorithm and same data directly without processing the data means without convert documents to graph. Then analyzing the results of the performance measurements, the conclusion can be drawn that the intrusion detection based on an enhanced DBSCAN algorithm achieves the highest recognition accuracy than other method .

REFRENCES

[1] V. CHANDOLA, A. Banerjee, and V. Kumar, “Anomaly Detection: A Survey,” ACM computing surveys, 2009.

[2] Farhoud Hosseinpour, 2Payam Vahdani Amoli, 3Fahimeh Farahnakian, 4Juha Plosila and Timo Hämäläinen" Artificial Immune System Based Intrusion Detection:Innate Immunity using an Unsupervised Learning Approach"International Journal of Digital Content Technology and its Applications(JDCTA) Volume 8, Number 5, October 2014

[3] Caleb C. Noble, Diane J. Cook"Graph-Based Anomaly Detection"SIGKDD ’03, August 24-27, 2003, Washington, DC, USA. Copyright 2003 ACM 1-58113-737-0/03/0008.

[4] Satinder Singh1, Guljeet Kaur2 "Unsupervised Anomaly Detection In Network Intrusion Detection Using Clusters" Proceedings of National Conference on Challenges & Opportunities in Information Technology (COIT-2007) RIMT-IET, Mandi Gobindgarh. March 23, 2007.

[5] Mete ÇELİK, Filiz DADAŞER-ÇELİK, Ahmet Şakir DOKUZ "Anomaly Detection in Temperature Data Using DBSCAN Algorithm" 2011.

[6] Abdulsahib, A. K. (2015). Graph based text representation for document clustering. Universiti Utara Malaysia.

[7] Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: O'Reilly Media Inc.

[8] Dolamic, L., & Savoy, J. (2008). Stemming approaches for East European languages Advances in Multilingual and Multimodal Information Retrieval (pp. 37-44): Springer.

[9] Dhillon, I. S., & Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering. Machine learning, 42 (1-2), 143-175.

[10] AGYEMANG,MALIK & BARKER,KEN&ALHAJJ, (2006) "A COMPREHENSIVE SURVEY OF NUMERIC AND SYMBOLIC OUTLIER MINING TECHNIQUES" , UNIVERSITY OF CALGARY, 2500 UNIVERSITY DRIVE N.W. CALGARY, AB, CANADA T2N.

[11] Hodge, V. J., & Austin, J. (2004). A survey of outlier detection methodologies. Artificial intelligence review, 22(2), 85-126.

[12] Porter, (1980) "An algorithm for suffix stripping", Program, Vol. 14 Iss: 3, pp.130 - 137 International Symposium on INnovations in Intelligent Systems and Applications