International Journal of Current Engineering and Technology
ISSN 2277 - 4106

Available at http://inpressco.com/category/ijcet

Research Article

Survey on Clustering Algorithm & Diagnosing Unsupervised Anomalies for Network Security

Priti K.Doad[Å*] and Mahip M.Bartere[Å]

[Å]Department Department of Computer Science & Engineering,, G.H. Raisoni College of Engineering & Management, Amravati,Maharashtra,India.

Accepted 29 October 2013, Available online 25 December 2013, **Vol.3, No.5 (December 2013)**

Abstract

Data clustering is a process of putting similar data into groups. A clustering algorithm partitions a data set into several groups such that the similarity within a group is larger than among groups. This paper reviews four types of clustering techniques- k-Means Clustering, K-Median Clustering, Density Based Clustering, Filtered clustered. Performance of the 4 techniques are presented and compared. In this paper, we also discussed completely unsupervised approach to detect the attack, without relying on signature, labeled traffic & training. Also discussed limitations of supervised network attacks in an increasingly complex & ever evolving internet. To show the feasibility of such knowledge-independent (unsupervised) approach, we develop UNADA, Unsupervised Network Anomaly Detection Algorithm. UNADA uses novel & robust multi-clustering based detection technique and evaluate its ability to detect & characterize network attack without any previous knowledge. The evidence of traffic structure provided by these multiple clustering is then combined to produce abnormality ranking of traffic flows using correlation-distance based approach. Additionally, we compare its performance against previous unsupervised detection methods using traffic from two different networks.

Keywords: *Data clustering, Density based Clustering, Filtered cluster, K-Means clustering, K-Median clustering, Unsupervised Anomaly Detection.*

1. Introduction

Clustering is a division of data into groups of similar objects. Each group, called a cluster, consists of objects that are similar between them- selves and dissimilar compared to objects of other groups. Cluster analysis is a very important technology in Data Mining. It divides the datasets into several meaningful clusters to reflect the data sets' natural structure. Cluster is aggregation of data objects with common characteristics based on the measurement of some kind of information. There are several commonly used clustering algorithms, such as K-means, Density based and Hierarchical and so on. Data clustering is a process of putting similar data into groups. A clustering algorithm partitions a data set into several groups such that the similarity within a group is larger than among groups. Clustering is an unsupervised classification mechanism where a set of patterns (data), usually multidimensional is classified into groups (clusters) such that members of one group are similar according to a predefined criterion. Clustering of a set forms a partition of its elements chosen to minimize some measure of dissimilarity between members of the same cluster .Clustering algorithms are often useful in various fields like data mining, pattern recognition , learning theory etc.

2. Related work

Comparisons between Data Clustering Algorithms Osama Abu Abba, Computer Science Department, Yarmouk University, Jordan . This paper is intended to study and compare different data clustering algorithms. The algorithms under investigation are: k-means algorithm, hierarchical clustering algorithm, self-organizing maps algorithm and expectation maximization clustering algorithm. All these algorithms are compared according to the following factors: size of dataset, number of clusters, type of dataset and type of software used. Some conclusions that are extracted belong to the performance, quality, and accuracy of the clustering algorithms.

2.1 k-means Clustering Algorithm

k-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different

*Corresponding author: **Priti K.Doad**; **Mahip M.Bartere** is M.Tech Scholar

location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more. Finally, this algorithm aims at minimizing an objective function know as squared error function given by:

$$J(y) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (\|x_i - y_j\|)^2$$

Where,
'||xi - yj||' is the Euclidean distance between xi & yj
'ci' is the number of data points in i^{th} cluster.
'c' is the number of cluster centers.

Algorithmic steps for k-means clustering

Let $X = \{x_1, x_2, x_3, \ldots\ldots, x_n\}$ be the set of data points and $y = \{y_1, y_2, \ldots\ldots, y_c\}$ be the set of centers.
1) Randomly select *'c'* cluster centers.
2) Calculate the distance between each data point and cluster centers.
3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..
4) Recalculate the new cluster center using:

$$y_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

Where, *'c_i'* represents the number of data points in i^{th} cluster.
5) Recalculate the distance between each data point and new obtained cluster centers.
6) If no data point was reassigned then stop, otherwise repeat from step 3).

Advantages

1) Fast, robust and easier to understand.
2) Relatively efficient: O (tknd), where n is # objects, k is # clusters, d is # dimension of each object, and t is # iterations. Normally, k, t, d << n.
3) Gives best result when data set are distinct or well separated from each other.

Disadvantages

1) The learning algorithm requires a priori specification of the number of cluster centers.
2) The use of Exclusive Assignment - If there are two highly overlapping data then k-means will not be able to resolve that there are two clusters.
3) The learning algorithm provides the local optima of the squared error function.
4) Applicable only when mean is defined i.e. fails for categorical data.
5) Unable to handle noisy data and outliers.
6) Algorithm fails for non-linear data set.

2.2 K-medians

The K-medians clustering algorithm is also an important clustering tool because of its well-known resistance to outliers. K-medians, however, is not trivially adapted to produce normalized cluster centers. We introduce a new algorithm (called MN), inspired by spherical K-means, that integrates with K medians clustering to produce locally optimal normalized cluster centers. We now review the K-medians algorithm , which is used when one wishes to minimize the total 1-norm distance from each point to its nearest cluster center. K-medians is quite similar to K-means**.** Since we now work with 1-norm distance instead of squared Euclidean distance, our objective is stated as:
We start with a partitioning of the data as in K means.
Initialize $t = 1$.
1. For each point, find the closest cluster center as
Measured via 1-norm distance.
2. Compute the new set of cluster centers by computing the median of the cluster. In other words,
for each dimension compute the median value for that dimension over all points in the cluster. We use the median because the median is the point that minimizes the total 1-norm distance from all points to it.
3. Terminate if the stopping condition is met. Increment *t* and go to step 1 otherwise.
In a similar fashion to K-means, steps 1 and 2 of K medians are guaranteed not to increase the objective *Q*.

2.3 Density Based Clustering Algorithm

Density based clustering algorithm has played a vital role in finding non linear shapes structure based on the density. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is most widely used density based algorithm. It uses the concept of density reachability and density connectivity. Density Reachability - A point p is said to be density reachable from a point q if point p is within ε distance from point q and q has sufficient number of points in its neighbors which are within distance ε.Density Connectivity - A point p and q are said to be density connected if there exist a point r which has sufficient number of points in its neighbors and both the points p and q are within the ε distance. This is chaining process. So, if q is neighbor of r, r is neighbor of s, s is neighbor of t which in turn is neighbor of p implies that q is neighbor of p.
Algorithmic steps for DBSCAN clustering
Let X = {x1, x2, x3... xn} be the set of data points. DBSCAN requires two parameters: ε (eps) and the minimum number of points required to form a cluster (minPts).

1) Start with an arbitrary starting point that has not been visited.
2) Extract the neighborhood of this point using ε (All points which are within the ε distance are neighborhood).
3) If there are sufficient neighborhoods around this point then clustering process starts and point is marked as visited else this point is labeled as noise (Later this point can become the part of the cluster).
4) If a point is found to be a part of the cluster then its ε neighborhood is also the part of the cluster and the above procedure from step 2 is repeated for all ε neighborhood points. This is repeated until all points in the cluster is determined.
5) A new unvisited point is retrieved and processed, leading to the discovery of a further cluster or noise.
6) This process continues until all points are marked as visited.

Advantages

1) Does not require a-priori specification of number of clusters.
2) Able to identify noise data while clustering.
3) DBSCAN algorithm is able to find arbitrarily size and arbitrarily shaped clusters.

Disadvantages

1) DBSCAN algorithm fails in case of varying density clusters.
2) Fails in case of neck type of dataset.
3) Does not work well in case of high dimensional data.

2.4 UNADA (Unsupervised Network Anomaly Detection Algorithm)

UNADA, an Unsupervised Network Anomaly Detection Algorithm for knowledge-independent detection of anomalous traffic. UNADA uses a novel clustering technique based on Sub-Space-Density clustering to identify clusters and outliers in multiple low-dimensional spaces. The evidence of traffic structure provided by these multiple clustering is then combined to produce an abnormality ranking of traffic flows, using a correlation-distance-based approach. Unsupervised Network Anomaly Detection Algorithm has several advantages-

- It works in a completely unsupervised fashion, which means that it can be directly plugged-in to any monitoring system and start to work from scratch, without any kind of calibration or previous knowledge.
- It combines robust clustering techniques to avoid general clustering problems such as sensitivity to initialization, specification of number of clusters, or structure-masking by irrelevant features.
- It automatically builds compact and easy-to-interpret signatures to characterize attacks, which can be directly integrated into any traditional security device.
- It is designed to work on-line, using the parallel structure of the proposed clustering approach.

2.4.1 Unsupervised Anomaly Detection through Clustering

The unsupervised anomaly detection step takes as input all the IP flows in the flagged time slot. At this step UNADA ranks the degree of abnormality of each of these flows, using clustering and outliers analysis techniques. For doing so, IP flows are analyzed at two different resolutions, using either IPsrc or IPdst aggregation key. Traffic anomalies can be roughly grouped in two different classes, depending on their spatial structure and number of impacted IP flows.1-to-N anomalies and N-to-1 anomalies. 1-to-N anomalies involve many IP flows from the same source towards different destinations; examples include network scans and spreading worms/virus. On the other hand, N-to-1 anomalies involve IP flows from different sources towards a single destination; examples include DDoS attacks and flash-crowds. Using IPsrc key permits to highlight 1-to-N anomalies, while N-to-1 anomalies are more easily detected with IPdst key. Without loss of generality, let Y = {y1, ..,yn} be the set of n aggregated- flows (at IPsrc or IPdst) in the flagged slot. Each flow yi∈ Y is described by a set of m traffic attributes or features, like number of sources, destination ports, or packet rate. Let $x_i \in R^m$ be the corresponding vector of traffic features describing flow $y_i$and
X = {x1 , .., xn } the complete matrix of features, referred to as the feature space UNAD is based on clustering techniques applied to X. The choice of both keys for clustering analysis ensures that even highly distributed anomalies, which may possibly involve a large number of IP flows, can be represented as outliers.

3. Proposed work

Firstly create log file. For create log file, we using software for capturing web pages files & other network traffic. This can allow you to see the data coming in from- and going out to- your computer, such as instant messages, e-mails, and Web pages. In this case find out maximum data flow in current log file means details about packet size, source IP address, and Destination IP address etc. Then apply sliding time windowing scheme and Aggregation Process for traffic flow; means total sum of byte transact in that window, total byte send or transact per unit time. Creation of feature space matrix & apply various clustering algorithm required.

Using K-means Clustering algorithm and declare smallest group of cluster as outlier. And time for that we have to trace back into feature space matrix, aggregation and log file. Use trace data to Create signature for anomalous flow. Signature will be logged and updated the signature table. Signature table can be used for online detection anomalous flow.

4. Conclusion

To conclude this paper; we have present a survey of different type of Classification Technique. In this study, the basic concept of clustering & clustering technique are given. The processes of grouping a set of physical

or abstract object into classes of similar objects are named as clustering. Clustering is a significant task in data analysis and data mining applications. There are different types of Clustering algorithms partition-based algorithms such as K-Means, K-median, density-based algorithms. In this paper, the completely unsupervised algorithm for detection of network attacks. It uses exclusively unlabeled data to detect and characterize network attacks, without assuming any kind of signature, particular model, or canonical data distribution. This allows detecting new previously unseen network attacks, even without using statistical learning.

References

Amineh Amini, Teh Ying Wah,, Mahmoud Reza Saybani, Saeed Reza Aghabozorgi Sahaf Yazdi (20100), A Study of Density-Grid based Clustering Algorithms on Data Streams ,*IEEE 8th International Conference on Fuzzy Systems and Knowledge Discovery*, vol.3, pp.1652-1656.

Anoop Kumar Jain, Prof. Satyam Maheswari (2012), Survey of Recent Clustering Techniques in Data Mining, *International Journal of Computer Science and Management Research*, pp.72-78.

Pavel Berkhin (2002), A Survey of Clustering Data Mining Techniques, pp.25-71.

Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar, Nidhi Gupta (2012), A Comparative Study of Various Clustering Algorithms in Data Mining, *International Journal of Engineering Reserch and Applications* (IJERA), Vol. 2, Issue 3, pp.1379-1384.

A. K. Jain (2010), Data Clustering: 50 Years Beyond K-Means, *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–66.

Anna Sperotto, Gregor Schaffrath, Ramin Sadre, Cristian Morariu, Aiko Pras and Burkhard Stiller (2010), *An Overview of IP Flow-Based Intrusion Detection*, IEEE communications surveys & tutorials, vol. 12, No. 3, Third Quarter.

Jiong Zhang and Mohammad Zulkernine (2006), Anomaly Based Network Intrusion Detection with Unsupervised Outlier Detection *IEEE International Conference on Communications.*

Rui Xu (May 2005), Student Member, IEEE and Donald Wunsch II, Fellow, IEEE, *Survey of Clustering Algorithms*, IEEE Transactions on Neural Networks, Vol. 16, NO. 3.S. Hansman, R. Hunt *A Taxonomy of Network and Computer Attacks*, in Computers and Security, vol. 24 (1), pp. 31-43, 2005

Fred and A. K. Jain (2005) Combining Multiple Clustering Using Evidence Accumulation, *in IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27 (6), pp. 835-850.