A Proposed Architecture Of Anomaly Detection Using K-Means Clustering Algorithm

Krishna Prajapati[1], Hitesh Patel[2]
[1]Research Scholar, Computer Engineering, Hasmukh Goswami Collage of Engineering, Ahmedabad, Gujarat, India
[2]Head, department of Information Technology, Hasmukh Goswami Collage of Engineering, Ahmedabad, Gujarat, India

Abstract

Data Mining is an efficient data analysis process which is used to find the patterns and relationship of a large database. Clustering is a popular technique of data mining for unsupervised learning in which labels are not defined previously. Anomaly detection is a problem of finding unexpected patterns in a dataset. Unexpected patterns can be defined as those that do not conform to the general behavior of the dataset. Anomaly detection is important for several application domains such as financial and communication services, public health, and climate studies.

Keywords: *k-means 2-tier clustering Algorithm*

INTRODUCTION

1. Introduction to anomaly detection

Anomaly detection is a problem of finding unexpected patterns in a dataset. Unexpected patterns can be defined as those that do not conform to the general behavior of the dataset. Anomaly detection is important for several application domains such as financial and communication services, public health, and climate studies.

1.1 Data mining:

Data Mining is an analytic process designed to explore data in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. In other way, it is the extraction of hidden predictive information from large databases. It is a powerful technology with great potential to help companies focus on the most information in their data warehouses. Data mining tools predict future trends and behaviors, allowing. Data mining has been very interesting topic for the researchers as it leads to automatic discovery of useful patterns from the database.

2. Background Theory

Anomaly detection is the process of finding the patterns in a dataset whose behavior is not normal on expected. These unexpected behaviors are also termed as anomalies or outliers. The anomalies cannot always be categorized as an attack but it can be a surprising behavior which is previously not known.

It may or may not be harmful. The anomaly detection provides very significant and critical information in various applications, for example Credit card thefts or identity thefts.

When data has to be analyzed in order to find relationship or to predict known or unknown data mining techniques are used.

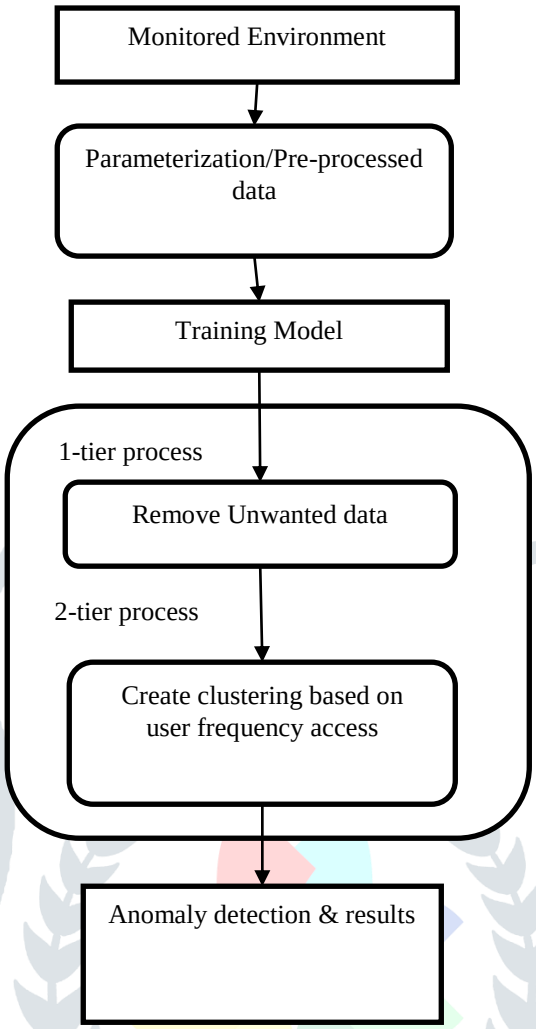

Figure 2.1: Methodology of Anomaly Detection

3. Related Work:

K-means 2-tier clustering Algorithm

K-means clustering is a cluster analysis method where we define k disjoint clusters on the basis of the feature value of the objects to be grouped. Here, k is the user defined parameter. There has been a Network Data Mining (NDM) approach which deploys the K-mean clustering algorithm in order to separate time intervals with normal and anomalous traffic in the training dataset. The resulting cluster centroids are then used for fast anomaly detection in monitoring of new data.

There are two clustering algorithms introduced .1-Tier and 2-Tier .In 1-Tier algorithm unwanted data like .doc, 404 page not found ,image should be removed from data source and in 2-Tier algorithm finding pattern clustering after removing duplicate request from data source.

Algorithm

1) ONE TIER CLUSTERING ALGORITHM
    1. Read N no of records from data source DS
       For i=1 to i<=N
       Next
    2. For each records R find unwanted or error data item from data source DS
    3. If R with .jpg, .png, 404 status then remove
    4. Original Record from data source DS
    5. Else maintain record in data source DS
    6. End if
    7. Next record
2) TWO TIER CLUSTERING ALGORITHM
    1. Read N no of records from data source DS
       For i=1 to i<=N
       Next
    2. For each records R from data source DS find pattern request data
    3. Read pattern request data using specified address from data source DS.
    4. If requested records from data source DS with specified pattern then
    5. Collect and save in pattern data source FDS.
    6. Repeat request then put FLAG=1
    7. Make two level cluster in pattern data source PDS.
    8. Else not select that records.

Conclusion

K-means clustering is a clustering method where we define k clusters on the basis of the feature value of the objects to be grouped.

We applied K-means 1-tier and 2-tier algorithm for creating clustering and detecting anomalies in dataset.

So it increases the performance and decreases the complexity in database.

Acknowledgements

I have taken efforts in this in this dissertation Preliminaries. However, it would not have been possible without the kind support and help of many individuals. I would like to extend my sincere thanks to all of them.

I am highly indebted to Prof. Hitesh Patel, for his valuable guidance and supervision regarding my topic as well as for providing necessary information regarding the dissertation.

I would like to thank Prof. Ashvin Prajapati also for his precious suggestions for the dissertation.

I would like to express my gratitude towards my lovely Parents for their kind co-operation and encouragement which help me in completion of this dissertation Report.

My thanks and appreciations also go to my friends who helped me out with their abilities.

References

[1] Shikha Agrawal, Jitendra Agrawal *"Survey on Anomaly Detection using Data Mining Techniques"* 2015 ELSEVIER.

[2] Naila Belhadj Aissa, Mohamed Guerroumi *"A Genetic Clustering Technique for Anomaly-Based Intrusion Detection Systems"*2015 IEEE 978-1-4799-8676-7.

[3] Zongwen Fan, Raymond Chiong, Zhongyi Hu, Yuqing Lin *"Investigating the effects of varying cluster numbers on anomalies detected in mining machines"* 2017 IEEE 978-1-5386-0765-7.

[4] LI Han *"Research and Implementation of an Anomaly Detection Model Based on Clustering Analysis"* 2010 IEEE 978-1-5090-6167-9.

[5] V. Jyothsna, V. V. Rama Prasad *" A Review of Anomaly based IntrusionDetection Systems"* 2011 IEEE.

[6] Chunyong Yin, Sun Zhang, Jin Wang *"An Improved K-Means Using in Anomaly Detection"* 2015 IEEE 978-1-4673-8600-5.

[7] J.James Manoharan1, Dr. S. Hari Ganesh2 Ph.D., Dr. J.G.R. Sathiaseelan *"Outlier Detection Using Enhanced K-Means Clustering Algorithm And Weight Based Center Approach"* 2016 IJCSMC.

[8] "Data Mining", *http://www.zentut.com/Data Mining*.

[9] Osmar R. Zaane, "Introduction of Data Mining", *Principles of Knowledge Discovery in Databases, University of Alberta-1999.*