**I J C T A**, 9(17) 2016, pp. 8473-8478
© International Science Press

# A Novel Neighborhood Density Based K-Means Algorithm for Clustering Quality Objective Function

**Gazal[*] and Pankaj Deep Kaur[*]**

**ABSTRACT**

Security of data is a major challenge incorporated by every enterprise in nowadays. With the rapid growth in the industrial environment, data has been increased exponentially. To protect the loads of data from unhealthy environment, unusual attacks requires quick attention. Big data can be defined as large amount of structured, unstructured and real time data which is difficult to classify, store, retrieve and manage increasing at an alarming rate. While transferring the information from one server to client, there is a huge opportunity available to intruders to steal or manipulate the data. Data mining techniques are gaining lot of attention in network intrusion detection system because of their ability to locate intrusions. These techniques has become liable with the increase in data and with the introduction of new attacks. Cluster analysis is a technique used to group the similar and dissimilar elements separately so as to locate the intrusions and also aids in the removal of overlapping of matched clusters.

***Keywords:*** K-Means Cluster, Big data, Density based clustering, Intrusion Detection system

## I. INTRODUCTION

Data mining has its own importance in the process of data analysis. The establishment and dynamic evolution of information technology had developed a large amount of data in all the fields. Earlier the researchers found, it was very difficult to manage, store, and get an information from such amount of data. In simple words, it is like finding a pin from the ocean. Data mining is a process of automatically mining worthy information or knowledge from large data sets. Data mining is also known as knowledge discovery process.

Commencement of big data comes in to our consideration with the alarming increase in the comprehensive data in various fields such as medical, astrological, business, economical, science etc. Big data with its development provides us new occasions for unearthing/discovering new values and gives us summons such as management issue, storage/organization issue and processing issue of data and help us to clinch/obtain high degree of understanding of hidden values. Big data has become a challenge in order to extract the knowledge because data grows beyond our range such as in zettabytes.

The Gartner definition of Big Data that is termed as 3 parts definition: "Big data is high- volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making [1]". A simple definition by Jason Bloomberg : "Big Data: a massive volume of both structured and unstructured data that is so large that it's difficult to process using traditional database and software techniques.[2]"

Cluster analysis is a technique of categorizing elements into number of clusters based on the information that describes elements and their relationships. Cluster analysis is used to group different elements which are either meaningful or useful or both. In some cases cluster analysis is just a useful opening point for other courses, such as data summarization. Clustering is also regarded as classification as it defines labeling

[*] Department of Computer Science and Engineering, Guru Nanak Dev University Regional Campus, Jalandhar, Punjab, India, *E-mail: gazalmittal9@gmail.com*

of the elements with cluster labels using available information. Cluster analysis also referred to as a supervised classification because it helps in assigning labels to new unnamed elements using a model evolved from elements with known cluster labels.

### *1.1. Cluster Types*---- There are numerous different notions of cluster

Well Separated Cluster: A cluster is group of elements in which all the elements in a cluster or a group must be satisfactorily close to this threshold (or to one another).

Prototype Based Cluster: Prototype cluster is generally called as center based clusters. A cluster is a set of elements in which each element is closely related to some set of rules that defines a particular cluster than to the rules of other clusters.

Graph Based Cluster: If the data is represented as graph, where the nodes are elements and the links denoted the relationships among nodes then a cluster can be defined as a linked component.

Shared Property Cluster: A cluster is said to be share property based cluster when the set of elements share some property. This cluster incorporates the meaning of all the other clusters.

Density Based Cluster: A cluster is an opaque area of elements that is bounded by an area of low opaque.

The rest of the paper is organized as follows. Related work is explained in section II. Proposed algorithm is explained in section III. Concluding remarks are given in section IV.

## II. RELATED WORK

Christos Boutsidis, Anastasios Zouzias, Michael W. Mahoney, and Petros Drineas (2015) studied dimension reduction method having two approaches feature selection and feature extraction method [3]. A feature selection-based algorithm selects a small subset of the input features and then applies k-means clustering on the selected features. A feature extraction-based algorithm constructs a small set of new artificial features and then applies k -means clustering on the constructed features. The proposed algorithms are randomized and provide constant-factor approximation guarantees with respect to the optimal k-means objective value.

Combination of density based and partition based clustering algorithm-DBK-means was proposed by B.L. Krishna, P.Jhansi, Lakshmi, P.Satya Prakash [4]. In this algorithm they proposed DBK-means algorithm to eliminate the shortcomings of DBSCAN and k-means clustering and to increase the performance of handling clusters of circularly distributed data points and slightly overlapping clustering. The proposed algorithm is combination of density algorithm and partition based algorithm.

Manpreet kaur and Usvir Kaur in 2013 discussed the performance of the K-means algorithm in terms of execution time [5]. They observed that the standard k-means algorithm takes long time to execute. For the purpose of removing this limitation, they suggested two new methods Ranking method and Query Redirection in order to speed up the clustering process. When the performance was observed these methods provide fast results as compared to standard results.

In 2014 Dr. Urmila R. Pol recommended variation in K-means algorithm and proposed parallel K-means clustering algorithm [6]. The k-means algorithm is computationally very expensive. The proposed algorithm is found to be more accurate and efficient compared to the original k-means algorithm. The proposed algorithm produces the more accurate unique clustering results .We have contended that to make data mining practical for common people, data mining algorithms have to be efficient and data mining programs should not require dedicated hardware to run. On these fronts, we can conclude from that, Parallelization is a viable solution to efficient data mining for large data sets.

Adapting K-means for clustering in big data discussed the analysis of large amount of information in order to extract the knowledge using various data mining techniques [7]. Proposed approximate kmeans algorithm which is fast, accurate and highly efficient to minimize the drawbacks of k-means of undefined number of iterations by fixing the times of iterations without losing precision.

## III. PROPOSED ALGORITHM

Due to the exponential growth in the data in recent times, the threat of stealing and manipulating or hurting the data increased in a larger extent. In order to protect the large amount of data from intrusions which arises due to the great development and rapid increment in the data exchange techniques and system management administration, clustering, a data mining technique is used. While clustering the high dimensional data set, number of challenges regarding it comes across our way. Despite of numerous improvements in clustering techniques, overlapping of cluster which are similar in nature and identifying the intrusions are great hurdles. So we introduced a new algorithm which is a combination of k-means clustering, density based clustering and maximum likelihood estimation.

### 3.1. Design of Algorithm

Dataset Acquisition: The initial step of algorithm is collecting data or gathering dataset. The dataset used for this work is KDDCUP99. We collected this dataset from uci repository. 5 million connection records, each with about 100 bytes. The two weeks of test data have around 2 million connection records. KDDCUP99 dataset is the most commonly use dataset for the assessment of anomaly detection system which majorly include intrusion detection system. This dataset is generated by stlofo et al. and is derived from the information collected from DARPA (1998) dataset. Stlofo processed the DARPA (1998) dataset and derived the KDDCUP99 dataset. It involves large set of attacks which are

1. Denial of Service Attack (DoS): is a type of attack in through which intruder makes network or machine resources such as computing memory incomprehensible to its corresponding users in order to stop the services of a host linked to the internet temporarily for indefinite period of time.

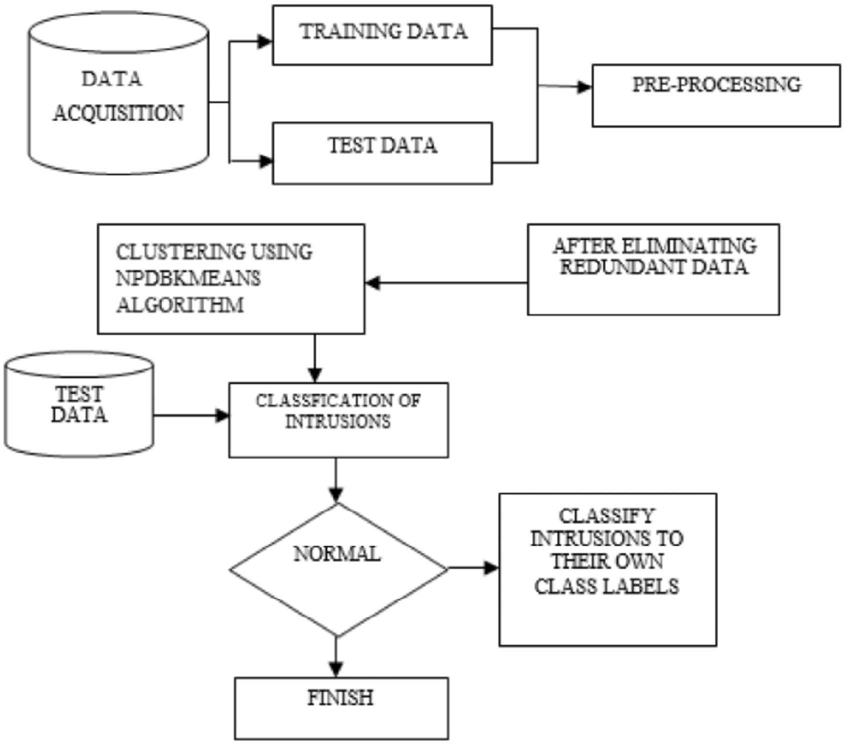
DATA ACQUISITION
TRAINING DATA
TEST DATA
PRE-PROCESSING
CLUSTERING USING NPDBKMEANS ALGORITHM
AFTER ELIMINATING REDUNDANT DATA
TEST DATA
CLASSFICATION OF INTRUSIONS
NORMAL
CLASSIFY INTRUSIONS TO THEIR OWN CLASS LABELS
FINISH

**Figure 1: Architecture of NPDBKMEANS Algorithm**

2. User to Root Attack (U2R): U2R means unauthorized access to root privileges. U2R is a type of intrusion in which intruders firstly attack the user personal account on the system to get access of the account (by hacking passwords class of exploit in which the attacker starts out with access to a normal user account on the system (perhaps gained by sniffing passwords, a dictionary attack, or social engineering) and is able to exploit some vulnerability to gain root access to the system.
3. Remote to Local Attack (R2L): occurs when an attacker who has the ability to send packets to a machine over a network but who does not have an account on that machine exploits some vulnerability to gain local access as a user of that machine.
4. Probing Attack: Probe stands for port scanning. Probe is an attempt to gather information about a network of computers for the apparent purpose of circumventing its security controls. It is important to note that the test data is not from the same probability distribution as the training data, and it includes specific attack types not in the training data which make the task more realistic. Some intrusion experts believe that most novel attacks are variants of known attacks and the signature of known attacks can be sufficient to catch novel variants.

Segregation of dataset: The dataset is segregated into two dataset which are training dataset and test dataset. A training dataset is processed to generate results and test data is executed to confirm the obtained results are correct. The last results are evaluated on the basis of implementation on the testing dataset.

The datasets contain a total number of 24 training attack types which are listed in table.

**Table I**
**Class Labels**

| Class Level | Intrusion |
|---|---|
| Normal | Normal |
| Denial of Service | back, teardrop, pod, smurf, neptune , land |
| User to Root | Buffer_overflow, loadmodule, perl, rootkit |
| Root to Local | Guess_passwd, ftp_write, imap,phf, multihop, warzmaster, warzclient, spy |
| Probe | Portsweep probe, nmap probe, ipsweep probe, satan probe |

Preprocessing and removal of redundant data: In feature selection process, we firstly filtered out the required and non-required attributes such that numeric and non-numeric respectively from dataset. The attributes either contain distinct values or non-distinct values. The attribute values in the dataset are not scaled. Some of the attributes values are scaled very high and some of the values are scaled very low. In order to remove the duplicate values of the dataset, normalization is used.

NPDBK-means clustering algorithm: NPDBK-means clustering algorithm is a combination of DBSCAN and K-mean clustering and Maximum Likelihood Estimator. This algorithm performs better than other traditional algorithm when handling clusters of circularly distributed data points and slightly overlapped clusters. The criteria for splitting or joining a cluster can be decided based on the number of expected points in a cluster or the expected density of the cluster (derived by using the number of points in a cluster and the area of the cluster). There is lot of scope for the NPDBK-means clustering algorithm in different application areas such as medical image segmentation and medical data mining. Basically NPDBK-means clustering algorithm overcome the drawbacks of DBSCAN and K-means clustering algorithms.

Figure 2. Describe basic flow chart of Neighborhood Probability Density Based K-means Algorithm. In very first step, calculate the Euclidean distance “eps” of all neighbors’ points to the starting point from the given dataset. Marked starting point as visited if neighbor ‘N’ is greater than or equal to ‘minPts’, other marked it as noise. Then new unvisited points are recall until all points are marked as visited. After that we

have ‘m’ clusters and then find cluster centers, ‘Cm’ by taking the mean find the total number of points in each cluster. To achieve K clusters with ‘Ck’ centers by joining two or more cluster based upon density and no. of points, if ‘m’ cluster is greater than ‘K’ clusters, otherwise select a cluster based on density and number of points split it using k-means clustering algorithm.

Let X = {$x_1$, $x_2$, $x_3$... $x_n$} be the set of data points in Dataset D with n points, Euclidean ”å” (eps) be Euclidean neighborhood threshold value.

K is the number of clusters to be found.

minPts is a minimum number of neighbors required in å neighborhood to form a cluster.

N is a set of points in ε neighborhood.

X can be any point.

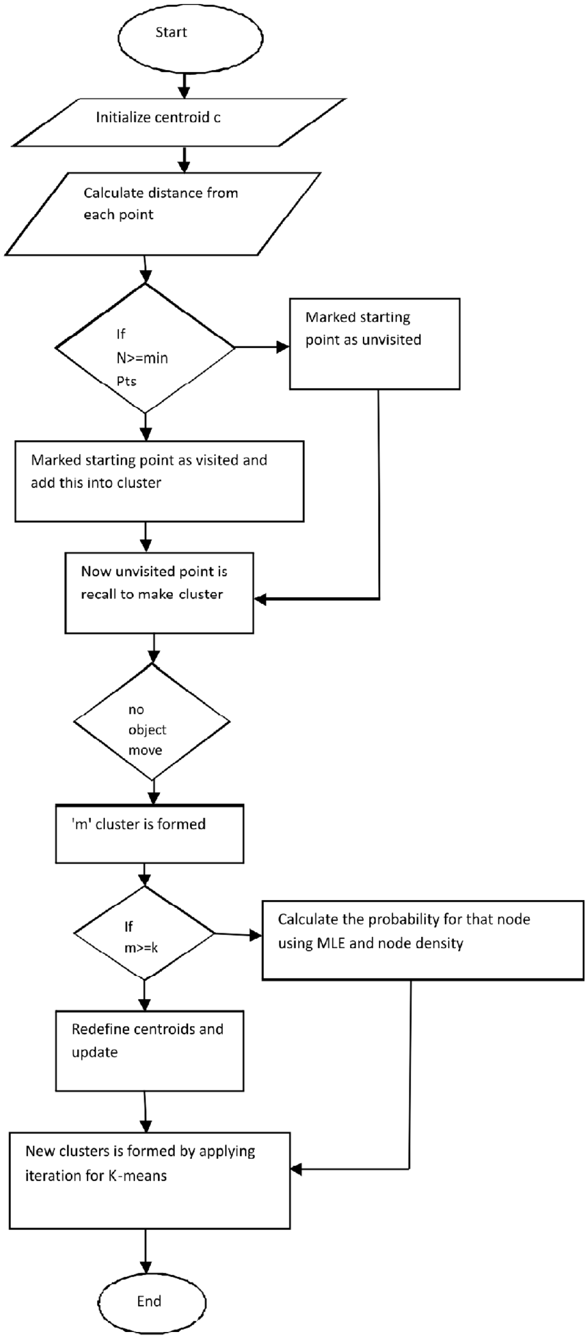
Start
Initialize centroid c
Calculate distance from each point
If N>=min Pts
Marked starting point as unvisited
Marked starting point as visited and add this into cluster
Now unvisited point is recall to make cluster
no object move
'm' cluster is formed
If m>=k
Calculate the probability for that node using MLE and node density
Redefine centroids and update
New clusters is formed by applying iteration for K-means
End

**Figure 2: Flow Chart of NPDBKMEANS Algorithm**

## IV. CONCLUSION

A novel neighborhood probability density based k-means algorithm using cluster quality objective function for intrusion detection system is proposed which restricts overlapping of cluster and also improve the accuracy of an algorithm to detect the intrusions attacking the data. Many algorithms which are combination of density and k-means algorithm are existing but in proposed algorithm we also combined these two with maximum likelihood estimator which solves our issue. It also addresses various attacks and normal queries by working on the dataset using the novel neighborhood probability density based k-means algorithm using cluster quality objective function for intrusion detection system.

## ACKNOWLEDGMENT

The following is trivial attempt to offer my heartfelt gratitude to the various persons who have been of immense help in one way or the other for the successful completion of this dissertation. I am highly grateful to the authorities of Guru Nanak Dev University, Regional campus, Jalandhar for providing this opportunity to carry out the dissertation work. My first, and most earnest, acknowledgment must go to my advisor Dr. Pankaj Deep Kaur to thank her for great patience, constructive criticism and myriad useful suggestions apart from invaluable guidance to me.

## REFERENCES

[1] Gartner's Big Data Definition Consists of Three Parts, Not to Be Confused with Three "V"s, By Svetlana Sicular, Gartner, Inc. 27 March 2013.[online] http://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-datadefinition-consists-of-three- parts-not-to-be-confused-with-three-vs/.

[2] The Big Data Long Tail. Blog post by Bloomberg, Jason. On January 17, 2013.[online] http://www.devx.com/blog/the-big-data-long-tail.html.

[3] Christos Boutsidis, Anastasios Zouzias, Michael W. Mahoney, and Petros Drineas." Randomized dimensionality reduction for k-means clustering", in IEEE transactions on information theory, vol. 61, no. 2, pages 1045-1058, February 2015

[4] B.L. Krishna, P.Jhansi Lakshmi, P.Satya Prakash. "Combination of Density Based and Partition Based Clustering Algorithm-DBK Means". Published in International Journal of Computer Science and Information Technology. vol3(3), 2012 ISSN 0975-9646

[5] Manpreet kaur and Usvir Kaur," A survey on clustering principles with K-means clustering algorithm using different methods in detail". In International Journal of Computer Science and Mobile Computing, ISSN 2320–088X, Vol. 2, Issue. 5, May 2013, pg.327 –331

[6] Dr. Urmila R. Pol , "Enhancing K- means Clustering Algorithm and Proposed Parallel K-means Clustering for Large Data Sets". In International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, Volume 4, Issue 5, May 2014, © 2014.

[7] Mugdha Jain, Chakradhar Verma. "Adapting Kmeans for Clustering in Big Data", Published in International Journal of Computer Applications (0975 – 8887) Volume 101– No.1, September 2014.