Application of K-Means Clustering to Multidimensional Geo-Spatial Data and Clustering Algorithm for Handling Data Precision

Kiran M N[1] Dhanjai N[2] Mahendra Kumar .B[3]
[1,2]PG Scholars [3]Assistant Professor
[1,2,3]Master of Computer Application
[1,2,3]Dayananda Sagar College of Engineering, Bengaluru, India

Abstract— Clustering is a method of grouping a group of objects in such a way that data objects in same group called "CLUSTER" have similar properties to each other than to those in other cluster (group). The K-means algorithm is based on simple idea i.e. Given a set of initial clusters points, assign each point to one of items, then each cluster center is replaced by the mean point on the respective cluster. These two simple steps are repeated until convergence. A point is assigned to the cluster which is close in Euclidean distance to the point. In this paper we will discuss how K-means algorithm is used for clustering of data (objects) and solve problems like active learning intrusion detection, distribution of teachers, clustering of multidimensional geo-spatial data and handling data precision.

Key words: K-Means Clustering, Privacy Preserving, Anonymized Table, AES Encryption Algorithm, Multi-Spectral Data, Geo-Spatial Data, Machine Learning, Active Learning

I. INTRODUCTION

Data Mining is the technique used to visualize and analyse the data and drive some useful information from that data so that information can be used to perform any useful work. The main elements of data mining includes is to fetch the data, modify that data and then load the data into database. Data mining tasks can be predictive or descriptive.

Common Data Mining tasks are:

1) Classification: is a process where dataset is classified according to the needs for generating pattern for discovering useful patterns.
2) Clustering: Is the task grouping data objects into group which have similar properties using various clustering algorithms and techniques?
3) Regression: Is the task of aattempting to find a function which models the dataset with the minimum least error.
4) Summarisation: is providing more compact, understandable and easy representation of data including visualisation and report generation.

Clustering is an essential stage in data mining process which is used for the purpose to make groups or clusters of the given data objects in data set based on the similar property/similarity between them. K-means clustering algorithms is one of clustering algorithm used in data mining that performs the modelling process without supervise and also one of the methods using the method of grouping data in partition.

II. K-MEANS CLUSTERING

The K-means clustering is unsupervised algorithm used to group different data object into c1usters. A cluster is collection of data objects that have similar properties to each other within one cluster and dissimilar to data object in other cluster. A cluster of data objects can be treated collectively as one group and so clusters of data objects can be considered as a form of data compression. There are 'n' number of clustering algorithms are available but among all K-means clustering algorithm is used more often in which we divide the objects into K number of clusters. In this method, the entered data will be grouped into a certain class/group, with the data objects having no specific label or class.

A. Algorithm:

Input:

- k: the number of clusters,
- D: a data set containing 'n' data objects.

Output: A set of k clusters.

Steps:

1) arbitrarily choose 'k' objects from dataset 'D' as the initial cluster centres;
2) repeat
3) (re)assign each data object to the cluster to which the data object is the most similar, based on the mean value of the data objects in the cluster;
4) update the cluster means, i.e. calculate the mean value of the data objects for each cluster;
5) until no change;

The time complexity of the k-means Clustering algorithm is O (nkt), where 'n' is the total number of objects, 'k' is the number of clusters, and't' is the number of iterations. Normally, k<<n and t<<n. Therefore, the method is relatively scalable and efficient in processing large, data sets. Here we will discuss how k-means clustering algorithm is used in different scenarios and different problems for generating clusters.

III. REVIEW OF RESEARCH WORK

A. Implementation of K-Means Clustering Method to Distribution of High School Teachers

This problem deals with the unequal distribution of teachers in Indonesia, The purpose of this research is to apply k-means clustering algorithm to analyze distribution of high school teachers in Indonesia.

Fig. 1: Three stage of study

This research has three main steps, namely dataset selection, preprocessing data, and application of k-means clustering.

1) Data Selection:

In this stage the data is selected for the problem i.e. data of on the number of students and the number of teachers, namely the data of equalization of students entitled "Ratio of School

Students" and the distribution of teacher entitled "Ratio of School Teachers".

2) Preprocessing data:

Before the dataset undergoes clustering, data needs to be preprocessed and the following steps are mentioned below

a) Data Selection:

In this step, Selection of data in dataset is done to obtain information required by the user. Selection of such data is used to assist the clustering and further processes. The initial attribute selection is done by eliminating unused attributes which are not used for clustering process.

b) Data Integration:

Data Integration is step where data which has been collected from various data sources are combined so that data set required for clustering process is available.

c) Attribute Addition:

In this step additional required attributes will be added to the data set for clustering process.

d) Data Conversion:

In this step, data conversion is used to help data adapt to the program. It may include conversion of one data type to another

i.e. nominal to numeric and vice-versa.

e) Replace Missing Value:

In this step empty or Missing data will be filled. The Missing value is filled with the mean of the attribute in the data set.

3) K-Means algorithm:

After the data has been pre-processed ,K-means clustering process is done. The steps for clustering with the K-means method are:

1) Select the number of clusters k
2) Initialize k cluster center randomly. Cluster centers are given a random initial value.
3) Allocate all existing object data on the closest cluster. The similarity of two objects can be determined by the distance between the two objects.
4) Then recalculate the cluster center with the current cluster membership.
5) Repeat the steps of cluster processing before, so there is no change.

After the k-means clustering process, this is done by grouping as many as 12 'k' clusters. The clustering process will result in 12 clusters that will be re-grouped into 3 clusters as fewer teachers, enough teachers and more teachers based on results.

Based on the calculation of k-means clustering, it can be used for solving the distribution of teachers to region that has less teacher, so as to solve the issue of uneven distribution of teachers.

B. Analysis of K-means clustering for Human Capital Trends

In the fast moving world, employees of the company should get familiar with newer technology which helps the company to move forward and be more productive. In this paper we will analyze the challenges faced by organizations. Many of the organizations are facing a large amount of requirement of deciding and auditing the huge organizational practices which include reducing the work/task and reconstruct the task to help people to adjust to the demanding working culture. The following challenges are faced by organizations are mentioned below

1) Culture and engagement
2) Leadership
3) Learning and development
4) Reinventing HR
5) Workforce on demand
6) Performance management
7) HR and people analytics
8) Simplification of work
9) Machines as talent
10) People data everywhere

In this study we are using K-Means clustering algorithm to understand all the issues that are found in organizations by clustering the all the major trends to their priority.

Issues	Global	Consumer Business	Energy & resources	Financial services	Life sciences & health care	Manufac turing	Professional services	Public sector	Technology media & telecommuni cation
Culture & engagement	78	81	76	81	79	75	78	75	77
Leadership	78	80	79	81	78	77	77	78	77
Learning & development	74	74	73	74	71	70	76	77	75
Reinventing HR	71	74	71	72	73	70	69	72	71
Workforce capability	70	69	72	70	74	68	72	68	73
Performance management	68	70	66	69	67	66	69	67	71
HR & people analytics	66	66	66	69	65	64	67	64	67
Simplification of work	63	64	60	64	63	62	64	64	63
Machines as talent	55	53	54	55	50	53	57	56	57
People data everywhere	50	49	49	50	50	48	50	48	52

Table I: Important trends in various organizations

Table I Gives a brief information about the various trends in different organizations. K-means algorithm is a partitioning based algorithm which divides the data objects into 'K' clusters (CI, C2, C3 ...CK), denoted by their cluster means generated considering their similarities with other data objects belonging to the same cluster. The mean associated with the cluster for every iteration evaluates the final cluster head so formed.

Challenges	Cluster assignment
Culture & engagement	1
Leadership	1
Learning & development	1
Reinventing HR	2
Workforce capability	2
Performance management	2
HR & people analytics	3
Simplification of work	3
Machines as talent	3
People data everywhere	3

Table II: Shows clusters drawn from the acquired data which clustered the ten trends into three clusters as shown in the table

By analyzing the information from numerous organization we explored the challenges faced by those organizations and readiness for the new technologies. Few challenges were vital in both small and large organizations while few challenges where affecting more on large organizations. Few challenges did not affect large organizations but affected small organizations.

C. *Using K-Means Clustering Algorithm for Handling Data Precision*

In this research, we are dealing with security of sensitive data such as patient details, student details data..etc. In the world of internet and digital data data privacy plays an important role.To prevent the misuse of sensitive data, privacy preservation mechnaism which can be carried using anonymization tables.This system states that the original patient data is converted to anonymized table which is identifiable only by the administrator. And k-means algorithm is used to form cluster of anonymized tables to differnet clusters on the basis of similar data object which is based on anonymized data attribute.After the data has been clustered it is protected using AES algorithm. The aim of this paper is to provide more accuracy and better level of privacy of sensitive attribute.

The proposed system implemented with the following modules:

1) Role Management
2) Anonymization
3) Clustering Analysis
4) Privacy preserving Techniques

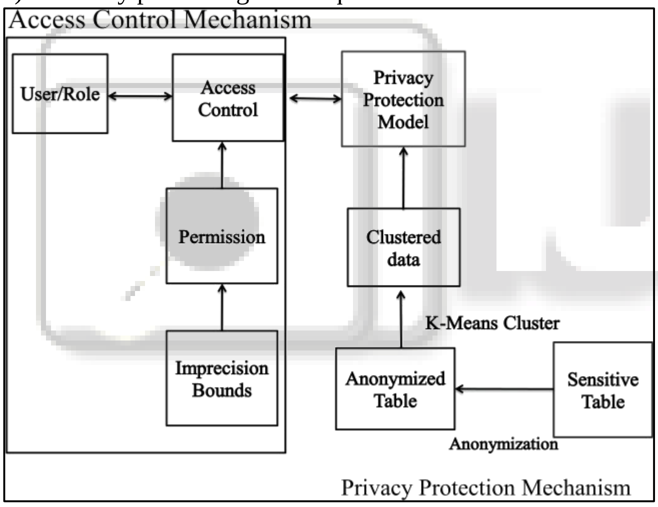

Fig 2: System Description

1) *AES algorithm:*

1) Step1: Get the Anonymized Table which have non sensitive data.
2) Step2: The activation code is created for each entity in the anonymized table. The key is only generated by the system/database administrator.
3) Step3: RBAC request to admin when view the anonymized table.
4) Step4: Admin accept or reject the request. Encrypt the data using shared secret key by AES algorithm.
5) Step5: Compare the key, if the key match means, the anonymized table are displayed to RBAC.

Otherwise

Reject the Request of RBAC.

This paper deals with protecting sensitive data using encrypted anonymized tables and protecting those tables using AES algorithm so that the sensitive data could not be misused The privacy preserving techniques anonymizes the data to acquire requirements. In this study, we use cryptography based anonymization which ensure that the resulting details should be anonymous.

D. *Application of k-means clustering to multidimensional geo-spatial data*

This study deals with clustering of multi-dimensional geo spatial data. Due to fast growing world the amount of geo spatial data that is being captured and stored is increasing day by day. So it is now necessary that the data has to processed and result should be delivered in time. With advanced technologies like parallel and distributed framework it can be done. In this paper we have extended the k means algorithm such that it is used with Hadoop by performing clustering on multiple dimensions (multiple spectrums) simultaneously over a distributed processing framework, Apache Hadoop, our approach allows detailed clustering.

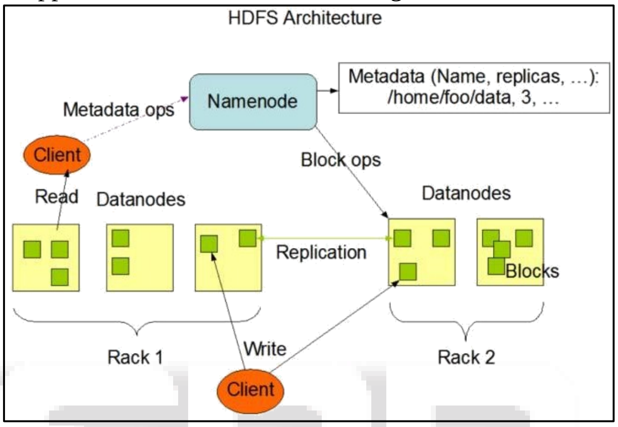

Fig. 3: Hadoop Distributed File System

Hadoop is an open-source framework for large-scale data storage and data processing that is compatible with commodity hardware. A distributed file system is a file system that can store large files spread across the nodes of a cluster. Spatial Hadoop] is a extension which deals with Hadoop for supporting for Multi-dimensional geo-spatial vector data over Hadoop. It has 4 main layers which are Language, storage MapReduce and Operations layer

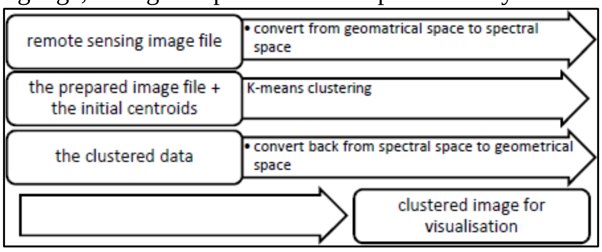

Fig. 4: Workflow of the proposed geo-processing k-means clustering model

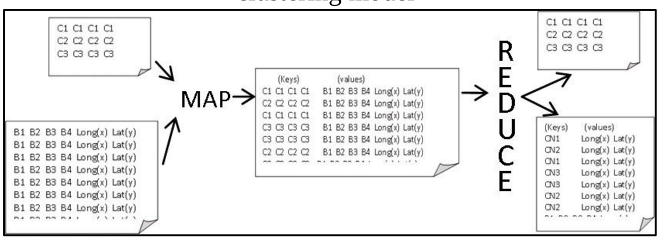

Fig. 5: Multi-spectral k-means clustering depicting transformation of values from Map to Reduce phases

After the mapper and reduces phase the final output of proposed model will be two main files which are the final centroid set, and the second file will contain the coordinates of each point in the input dataset along with the cluster number which that point belong to as shown in Fig. 5.

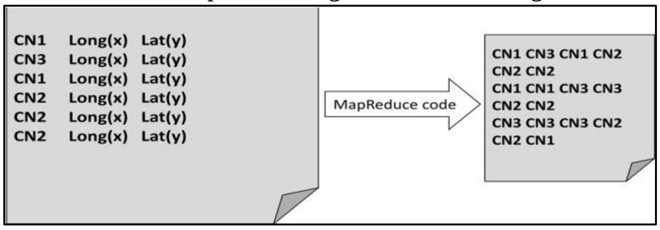

Fig. 6: Transformation for Cluster visualization from MapReduce output

Geo-spatial data comes through various different sources it is necessary that those data is stored and processed through which huge amount of hidden knowledge can be extracted despite of their size ,complexity..etc. By using Hadoop, MapReduce technique, distributed framework. The proposed model supports n-number of bands for processing the data directly from HDFS. After processing, the final output can be visualized using the visualization interface and tools.

E. Active Learning Intrusion Detection using k-Means Clustering Selection

In the world of data and internet, Cyber-attacks on government, private organizations are common and increasing day by day. To identify attacks and compromises of computer systems we use Intrusion detection system. The challenge of intrusion detection is further complicated because the increasing interconnection of devices on the internet. This research states that until now there's not an proper, effective algorithm(Machine learning algorithm) that minimizes human interaction to train and evolve the learner to adapt to changing cyber-attacks. This study implemented the Active Learning Intrusion Detection System (ALIDS) prototype as shown in Figure 6.

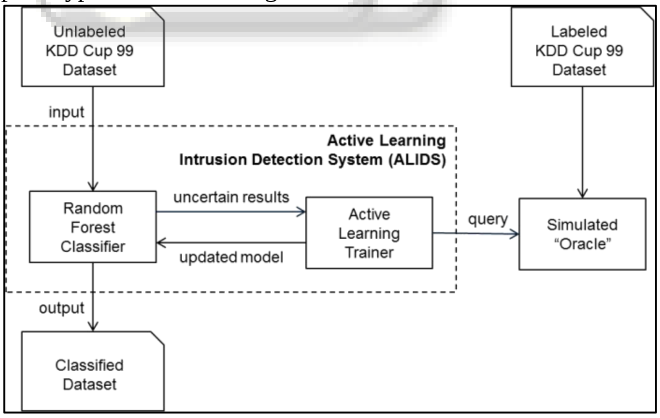

Fig. 7: ALIDS overview

This study uses Active Learning Intrusion Detection System (ALIDS).In this system Unlabelled data is given as input to ALIDS system where it has KDD Cup 99 dataset which has been manually labelled by human subject matter experts. The labels were excluded for use with ALIDS, Random forest classifier is used for evaluating classification results and Random forest is highly effective

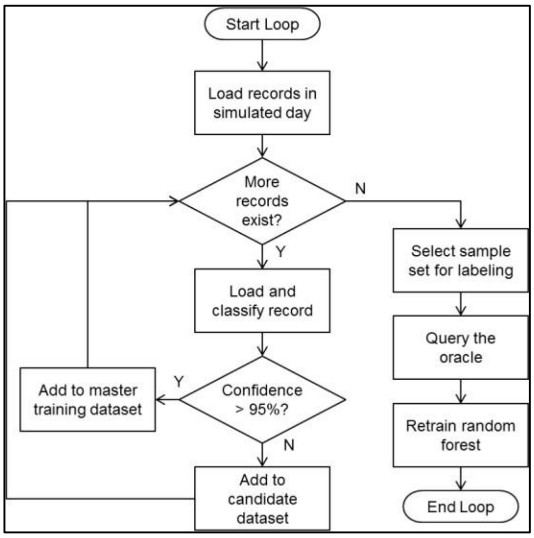

Fig. 8: Algorithm of each simulated day

Figure 7 depicts the ALIDS algorithm which is used in this system which is used for each simulated day .This algorithm Classifies each record, evaluates the confidence of the classification then ALIDS adds the data of each simulated day to master training data set for high confidence results or for candidate dataset. Since Random forest is efficient and fast the testing can be done at very less time.

This study has shown the usability of using active learning for intrusion detection to reduce the human effort and the algorithm adopts to threats that are against machine learning algorithms .Using K-mean clustering effective sampling of unlabelled data was done and reducing manual labelling of data by human.

F. Classification of Moving Vehicles using K-Means Clustering

In the fast growing economy of world, the numbers of vehicles have gradually increased and it's increasing day by day. As number of vehicles increases more complex gets the surveillance, more traffic congestion and more accidents. This study deals with importance of variety of traffic congestion, Surveillance and accident prevention etc. Sequences of traffic video clips are captured using stationary NI cameras. Then the captured clips are processed in LABVIEW and data is clustered into two clusters. LabVIEW is graphical programming software that allows for instrument control, data acquisition, and pre/post processing of acquired data.

The vehicle classification system was implemented in following steps:

1) Use NI smart camera and image acquisition technique to grab traffic video clip from a road
2) From the video clip obtain image frames in LabVIEW environment
3) Moving vehicle detection on road in LabVIEW
4) Image processing and vehicle attributes feature extraction with vision assistant in LabVIEW
5) Vehicle classification by WEKA data mining toolkit

The data such as size, length, height, width is extracted from the clips using LABVIEW then the data is clustered into two clusters using WEKA tool. WEKA tool is open source tool used for clustering which has GUI. Clusters generated are two clusters where one cluster contains all the data of small vehicles according to the data obtained from clips and other cluster with data of larger vehicles. The vehicle classification was90% accurate when compared to manually classification. And this method is more efficient than the manual method. The vehicle classification system can be used to automate classification of vehicles on road in a traffic monitoring framework.

IV. COMPARATIVE ANALYSIS

In this section we are comparing the methods which are explained earlier in this paper

	Keywords	Applications	Methodology
PAPER -1	k-means clustering; student ratio; teacher ratio	Distribution of School teachers	K-means Clustering
PAPER -2	developmen t; challenges; clustering capital;	Used in Organization s	K-means Clustering
PAPER -3	Privacy preserving, Anonymize d Table, Clustering, AES encryption algorithm.	Used in organization s where sensitive data is used like schools, hospital,.etc.	Anonymizatio n (Generalizatio n) and Suppression (Algorithm),K-means Clustering
PAPER -4	multi-spectral data; geo-spatial data processing; distributed geo-processing; k- means clustering; spatial processing;	Used in Space Research organization s, Geology department and Defense Departments	Hadoop, Map-reduce, K-means Clustering
PAPER -5	machine learning, intrusion detection, random forest, active learning, k-means clustering,	Used in Data Bases, data warehouses	Active Learning Intrusion detection, Machine learning, K-means Clustering
PAPER -6	Feature extraction, image processing, LabVIEW, data mining, k-means Clustering, WEKA	Used in Traffic managing Systems, Surveillance systems	Training dataset,K-means Clustering

V. SCOPE /IDEAS OF IMPROVEMENT

A. Best method

The best method is implementation of K-means clustering with Hadoop which supports distributed processing framework. This approach of K-means with Hadoop allows detailed efficient clustering and guarantees faster processing and delivering results in less time than the normal conventional method

B. Improvisation Ideas

As every method successfully achieves its aim and works perfectly fine most of the time, but it fails in some situations due to failures or lack of performance. So improvement should be done for them to work perfectly every time. My idea in School teachers distribution scenario is to use schools which has less teachers and more teachers with the regions as mentioned in paper which allows us to distribute teachers to the schools where teachers are required

VI. CONCLUSION

In the teacher’s distribution problem, by applying k-means clustering, we can solve the uneven distribution of teachers to the entire region. Based on the calculation of k-means clustering, it can be used as a reference to perform the distribution of teachers to region that are less teacher, so as to solve the issue of uneven distribution of teachers. In handling data precision, sensitive data is protected using encrypted anonymized tables. And k-means algorithm is used to form cluster of anonymized tables to different clusters on the basis of similar data object which is based on anonymized data attribute. After the data has been clustered it is protected using AES algorithm so that the sensitive data could not be misused or altered or damaged, such that only authorized person can do the modification to the sensitive data. In ALIDS study k- means clustering is used for effectively sampling of unlabeled data which reduces manual labelling of data by humans. In the case of geospatial data, proposed k-means model is developed in such a way that it works with MapReduce environment and it’s designed specifically for clustering the multi-dimensional spatial data. The final output of our proposed model will be two main files which are the final centroids set, and the second file will contain the coordinates of each point in the input dataset along with the cluster number which that point belong. In vehicles classification scenario data from NI cameras(clips) are taken, and using LABVIEW the vehicles properties like size, width, height, perimeter are extracted and k-means clustering is applied. Two clusters are formed in which one cluster consists of data which are of small vehicles and other one with data of vehicles of larger vehicles. By this data we can manage the traffic congestion in the metropolitan cities and also can be surveillance system.

REFERENCES

[1] Triyanna Widiyaningtyas, Martin Indra Wisnu Prabowo, M. Ardhika Mulya Pratama" Implementation of K-Means Clustering Method to Distribution of High School Teachers", Yogyakarta, Indonesia, 19-21 September 2017
[2] Gamini Sharma, Manish Kumar Sharma, Dakshata Sharma" Analysis of K-means clustering for Human Capital Trends" Indore, India, April 2017
[3] P.Suganthi, K.Kala, Dr.C.Balasubramanian," Using K-Means Clustering Algorithm for Handling Data Precision", Kovilpatti, India, October 2016
[4] Mazin Alkathiri, Jhummarwala Abdul, Dr. M.B. Potdar," Application of k-means clustering to multidimensional geo- spatial data" Indore, India,,February 2018
[5] Steven McElwee," Active Learning Intrusion Detection using k-Means Clustering Selection", Charlotte, NC, USA, May 2017
[6] Suresh Babu Changalasetty, Lalitha Saroja Thota, Ahmed Said Badawy, Wade Ghribi," Classification of Moving Vehicles using K-Means Clustering", Coimbatore, India, August 2015