STEGANOGRAPHY USING KMEANS CLUSTERING ALGORITHM

[1]Ambe Nelsa Flora, [2]Dr. Umarani C.

[1]Student, [2]Assistant Professor

Masters of Computer Application - ISMS

Jain University, Bangalore, India

Abstract - The growth of high speed computer networks and particularly that of the internet has increased the ease of communication and transfer of data. The down sight of this discovery is the fear of getting ones data snooped or intruded during the time of transfer from one person or group of people to another. Many people forget that, just as the growth in internet, so has the growth is many attack techniques which are aimed at stealing and destroying these data. Data is an important aspect of every organisation so it needs to be protected at all cost. To combat these attacks and attackers, many security methods have been introduced. In all the many security measures employed, steganography stands out and is one of the most common methods of securing data to be transferred through the internet. Steganography is the method of stowing away classified information inside any media. Steganography involves hiding of text, image or any sensitive information inside another image, video or audio in such a way that an attacker will not be able to detect its presence. In this paper, we propose a technique where a sample data is first clustered using a clustering algorithm known as kmeans clustering algorithm and then the output of the clustering is then hidden using image steganography to provide a security layer to that data.

Keywords: Security, Steganography, Kmeans, Clustering.

I. INTRODUCTION

Millions of security problems and breaches happen on a daily basis and this has posed a problem to many organisations and individuals, who fear for their data. Over the years solutions to these problems have come up and one of which is steganography. Steganography was coined out from two words “steganos” which means “concealed” and “graphie” which means “writing”. Steganography is the process of hiding secret messages such that no one except the sender and the intended recipient can see them. The process of hiding data is used in many important applications in order to maintain confidentiality of important data, prevent unauthorized persons from identifying or understanding the confidential message, or add mark or a tag to the digital image to be used in order to identify the digital image ownership. What steganography essentially does is exploit human perception; human senses are not trained to look for files that have information inside of them. There are basically four formats in which steganography can be achieved and they include; text, audio, video and image steganography. Protocol steganography was recently added to the list. Protocol steganography refers to the technique of inserting or embedding information within messages and network control protocols used in network transmission.

Kmeans is an unsupervised clustering machine learning technique used to cluster data. Kmeans is a clustering algorithm whose main objective and aim is to group similar elements or data points into a cluster. The K in Kmeans represents the number of clusters in which the case dataset will be grouped into. What kmeans does is, it takes a dataset and then inputs a k. that K value is what is used to group the elements of the dataset. The elements of a centroid must not necessarily be the same but they have some key characteristics with each other. Kmeans is an iterative algorithm and can be applied to a numeric or continuous data with a smaller number of dimensions. It can be applied to any scenario where you want to make blocks of similar things from randomly distributed collection of things. For example, document classification where you want to group those documents into tasks, datelines, department and even contents of the documents.

In this paper, we are going to discuss how the combination of steganography and kmeans introduces a more safe and trusted method of securing data as it is been transported from one person to the other over the internet.

II. LITERATURE REVIEW

The exponential growth in technology has forced internet users to seek better and higher security techniques and methods to preserve their also growing data. Thus, throughout the ages, people have devised a means of hiding secrets from plain sight called steganography. Steganography has emerged as a glowing research area in which various methods have been proposed. Steganography has proven over time to be a trusted method for achieving the goal of safe and secure method in transferring data from the sender to the receiver. Almost all digital file formats can be used for steganography. Even though Image and audio files especially comply with this requirement, research and researchers have also uncovered other file formats that can be used for information hiding which includes protocol and text steganography. Images are the most popular cover objects used for steganography. When working with larger images of greater bit depth, the images tend to become too large to transmit over a standard Internet connection. The process of compression is employed to analyse and condense image data, resulting in smaller file sizes.

K-means algorithm is an iterative algorithm that divides the unlabelled dataset into k different clusters in such a way that each dataset belongs only to one group that has similar properties. The goal of the K-means algorithm is to find a group, assign each data to the cluster with the centroid closest to itself, and find minimum accumulated distance of each data point with the cluster center.

III. KMEANS AND STEGANOGRAPHY TECHNIQUES

There are several methodologies available to achieve steganography. It is very important to note that all the methods or techniques that are used, leads to data security or for transferring critical data from one person to another in a secure mode. These techniques include LSB, JPEG, and PNG and can be classified into the following;

Technical steganography: This method uses scientific methods to hide a message, such as the use of invisible ink or microdots and other size reduction methods.

Linguistic steganography: This method hides the message within the carrier in some nonobvious ways and is further categorized as semagramsor open codes.

Semagrams: This stenographic method hides information by the use of symbols or signs. A visual semagram uses innocent-looking or everyday physical objects to convey a message, such as doodles.

Open codes: This method hide a message within a legitimate carrier message in ways that are not obvious to an unsuspecting observer. The carrier message is sometimes called the overt communication, while the hidden message is the covert communication. This category is subdivided into jargon codes and covered ciphers. Jargon code uses language that is understood by a group of people but is meaningless to others while covered code, hides a message openly in the carrier medium so that it can be recovered by anyone who knows the secret for how it was concealed.

Kmeans requires k as an input and doesn’t learn it from data. In the cluster-predict methodology, we can evaluate how well the models are performing based on different K clusters since clusters are used in the downstream modelling. Two metrics that may give us some intuition about k are as follows;

Elbow method: Elbow method gives us an idea on what a good *k* number of clusters would be based on the sum of squared distance (SSE) between data points and their assigned clusters’ centroids. We pick *k* at the spot where SSE starts to flatten out and forming an elbow.

Silhouette analysis: Silhouette analysis can be used to determine the degree of separation between clusters. For each sample:

- Compute the average distance from all data points in the same cluster (ai).
- Compute the average distance from all data points in the closest cluster (bi).

- Compute the coefficient:

$$\frac{b^i - a^i}{max(a^i, b^i)}$$

The coefficient can take values in the interval [-1, 1].

- If it is 0 –> the sample is very close to the neighbouring clusters.
- If it is 1 –> the sample is far away from the neighbouring clusters.
- If it is -1 –> the sample is assigned to the wrong clusters.

IV. PROPOSED SYSTEM

In this project, we propose a method which takes a dataset, clusters that data using Kmeans clustering algorithm and then hides the result of the clustered data behind an image using image steganography. For the clustering technique, we created a user interface which makes it easy for the user to input data at any given point and time. It is important to note that this data must be in MS excel format. Once the file is selected, the user is asked to enter the number of clusters in which that data would be grouped into. Once the number of clusters has been entered, then the code runs and sends an output which includes the characteristics of each cluster or centroid as well as a plotted representation of the data. Once we get this output, the image is put in a file and that file is then encoded using steganography to improve on the security of that data. For the hiding/encoding the user selects the source of the file, inputs the name of picture to be used and finally inputs the destination of output after encoding. For the decoding/retrieving method, the user inputs only the source of the image to be decoded. Below is the proposed design for this project.

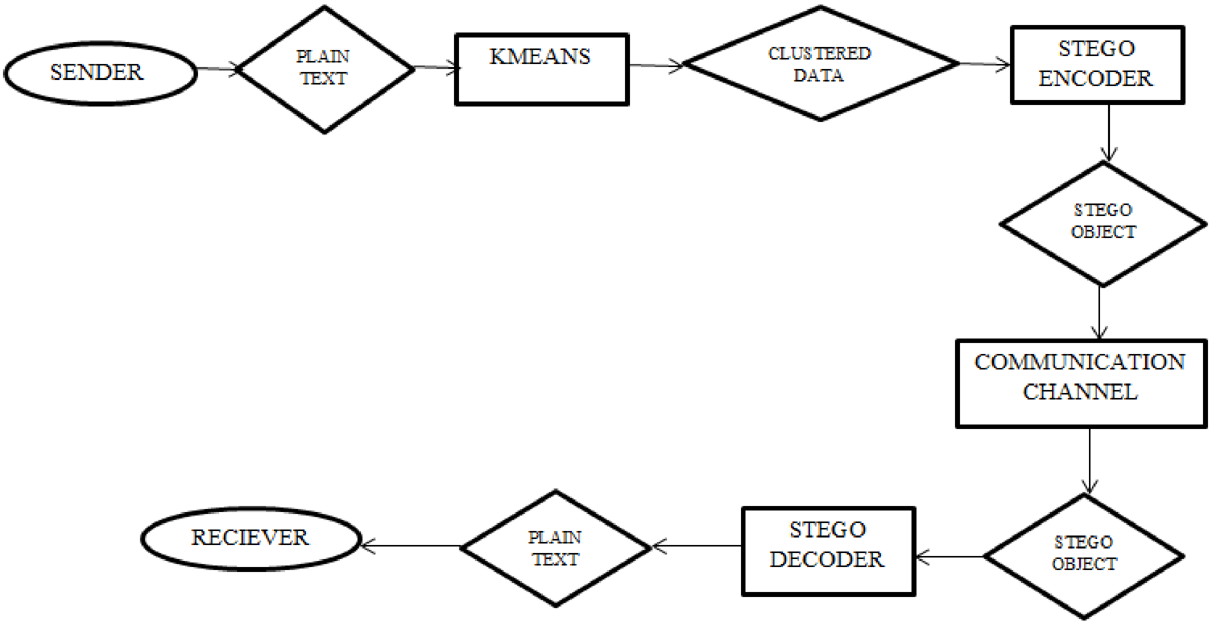

Fig: Proposed Design

V. MODULES

- Cluster
- Cluster classification
- Hide
- Retrieve

CLUSTERING

To create clusters from the input data, we have used k-means clustering algorithm. K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The algorithm initially have empty set of clusters and updates it as proceeds. For each record it computes the Euclidean distance between it and each of the centroids of the clusters. The instance is placed in the cluster from which it has shortest distance. Assume we have fixed metric M, and constant cluster Width W. Let di(C, d) is the distance with metric M, Cluster centroid C and instance d where centroid of cluster is the instance from feature vector.
Input: The number of clusters K and a dataset for intrusion detection
Output: A set of K-clusters Algorithm:
1. Initialize Set of clusters S. (randomly select k elements from the data)
2. While cluster structure changes, repeat from 2.
3. Determine the cluster to which source data belongs Use Euclidean distance formula. Select d from training set. If S is empty, then create a cluster with centroid as d else add d to cluster C with min (distance (C, d)) or distance (C ,d)<=distance(C1, d).
4. Calculate the means of the clusters. Change cluster centroids to means obtained using Step 3

CLUSTER CLASSIFICATION

If cluster width is chosen properly then after clustering each cluster contains instance of same type. The major task is to determine which clusters are normal and intrusive in case of intrusion detection. Here we assume that maximum numbers of records are normal from the training set. Then it is highly possible that the cluster with maximum numbers of instances contains normal records and other contains attack records. We have used 75% as threshold percentage value for labelling the normal cluster. The other clusters are labelled as anomalous.

HIDE

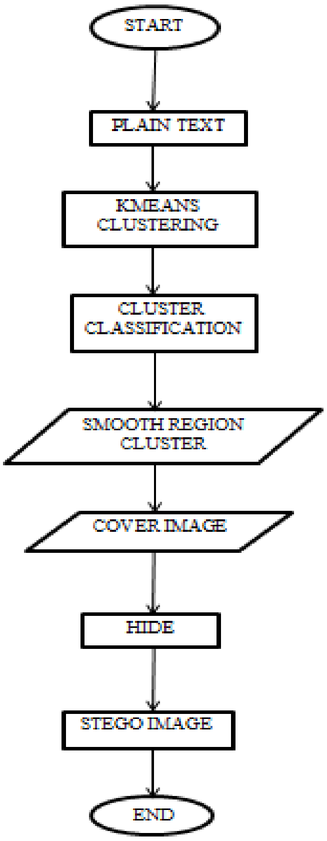

Fig: Encoding phase

It takes the filename and the message. It opens the image library where it gives the filename as an input. Then it converts the message from string format to binary format.
It checks if the image in rgb format or not. If not in rgb format, it converts it into rgb
It takes each and every bit if it in proper format, to see if the actual bit of the secret message can fit into it.
It replaces the bit
If the file doesn't exist, it returns a message saying "incorrect image mode couldn't hide".

RETRIEVE

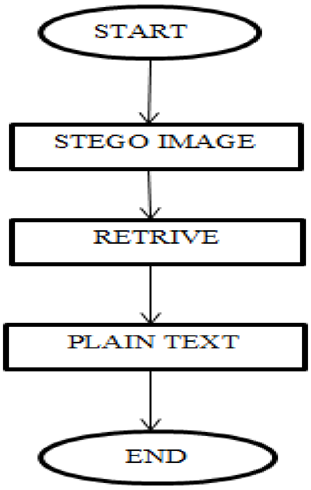

Fig: Decoding phase

It takes the filename of the data location, checks if the data is in rgb format and if not, it converts the data into rgb format. After conversion, it extracts the data, retrieves all the 0's and 1's until it finds the delimiter. Once it reaches the end of the data, it sends a message "successful" else it sends "incorrect image mode, couldn't retrieve".

VI. FUTURE SCOPE

In this project, our focus was based on clustering data using kmeans before performing image steganography. Kmeans is very limited in that, it is a lossy method and the data after being cluttered cannot be decluttered. Hence, as future scope, it would be great to work on methods of decluttering the transferred data upon reaching the receiver.

VII. CONCLUSION

This project is based on steganography with a twist incorporating the kmeans clustering algorithm which goes ahead to ensure that data is more secure as it leaves from point A to point B of the communication channel. With information security being a trivial and sensitive issue in this day and time, it is a breadth of fresh air that more techniques such as this, are continuously worked on, to improve the security of our day to day data. Our results shows that data successfully left from point A to point B through a communication medium safely and still maintained integrity and confidentiality which are 2 main features pf the CIA triad.

VIII. ACKNOWLEDGMENT

I would like to acknowledge God almighty for the strength and knowledge to write this paper. I would also like to acknowledge my above mentioned supervisor for her encouragement and guidance.

IX. REFERENCES

[1] Swati Gill,1 and Rajkumar Yadav1 “A New Method of Image Steganography Using 7th Bit of a Pixel as Indicator by Introducing the Successive Temporary Pixel in the Gray Scale Image”. 2018

[2] An efficient k-means clustering algorithm: analysis and implementation - Pattern Analysis and Machine Intelligence, IEEE Transactions on (umd.edu)

[3] A.A. Lubis, R. Purba and I. A. Pardosi, "Combination of Steganography with K Means Clustering and 256 AES Cryptography for Secret Message," in 2019 Fourth International Conference on Informatics and Computing (ICIC), Semarang, Indonesia, 2019.

[4] https://www.sciencedirect.com/topics/computer-science/steganographic-technique

[5] https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks

[6] https://digifors.cs.up.ac.za/issa/2005/Proceedings/Full/098_Article.pdf

[7] https://www.researchgate.net/publication/335080736_Application_of_K-means_Algorithm_in_Image_Compression

JETIR