A Machine Learning Approach to Detect Novelty in a Network using Patricia Tree

[1]Ananya Dey, [2]Hamsashree Reddy R, [3]Madhusoodhana Chari

***Abstract*: The internet has been one of the biggest revolutions in the history of mankind. Recent increase in the use of the internet has however led to an increase in the utilization of the network and hence increase in internet traffic. This has also led the network to become more vulnerable to attacks. In this paper, we have proposed a solution that considers the IP Addresses of a network. A Patricia tree is then constructed using these IP addresses and a novelty is determined in the network after analyzing the Patricia trees constructed with the help of the IP addresses. We then use K means clustering algorithm to determine the novelty in the network.**

***Index Terms*: Patricia Tree, K means clustering algorithm, ID3Algorithm, Prediction Probability algorithm.**

I. Introduction

The use of computers has been rapidly increasing in divergent fields such as education, business, banking and entertainment. Almost all work these days has been largely dependent on computers. However over the last few decades, their use has also led to a huge amount of data passing through the network. At any instance of time there are about a million IP addresses in a network. The increase in the network usage has led to an increase in the vulnerability of the network. Having said it is of utmost importance to provide users with a network that is free from any malicious attacks.

To explain our proposed solution, let us consider a scenario where the same devices are always connected to a network every day. Say for example, in an office where every day the employees come in with the same laptops and mobile devices. The IP addresses assigned to these devices remain almost the same over a given time period. However, one day if a new device is brought into the network, then it will be assigned a different IP address. Using this algorithm, we can determine in which day a new device was introduced and thus a novelty in the network.

In the previous research, some substantial work has been made to find out these anomalies and attacks based on the 'Prediction Probability algorithm'. However the approach presented here involves the novel method of creation of the Patricia tree to store the IP addresses in the network. This data structure is most commonly used to store words in a dictionary because it is based on the logic of common prefix. Likewise, this can also be used for storing the IP addresses of a network because most of the devices that are connected to a network have a similar subnet mask. This can be taken advantage of and can be used to store all the respective IP addresses having a common prefix into a Patricia tree. With the help of this, operations such as insertion and search become easier and more efficient. In this way we can improve the time and space complexity.

Once the Patricia tree is constructed, we must determine some parameters using which we give weights to each of these trees. This is done so that we can apply k means clustering to the dataset. The two parameters we have considered are x= number of leaf nodes that are present in the Patricia tree and y = total number of nodes that are present in the Patricia tree.

Finally K means clustering algorithm is applied on the (x,y) values of each of the trees. The value k denotes the number of clusters that are to be created. Once the clusters are created, the cluster that has one data point is the anomaly.

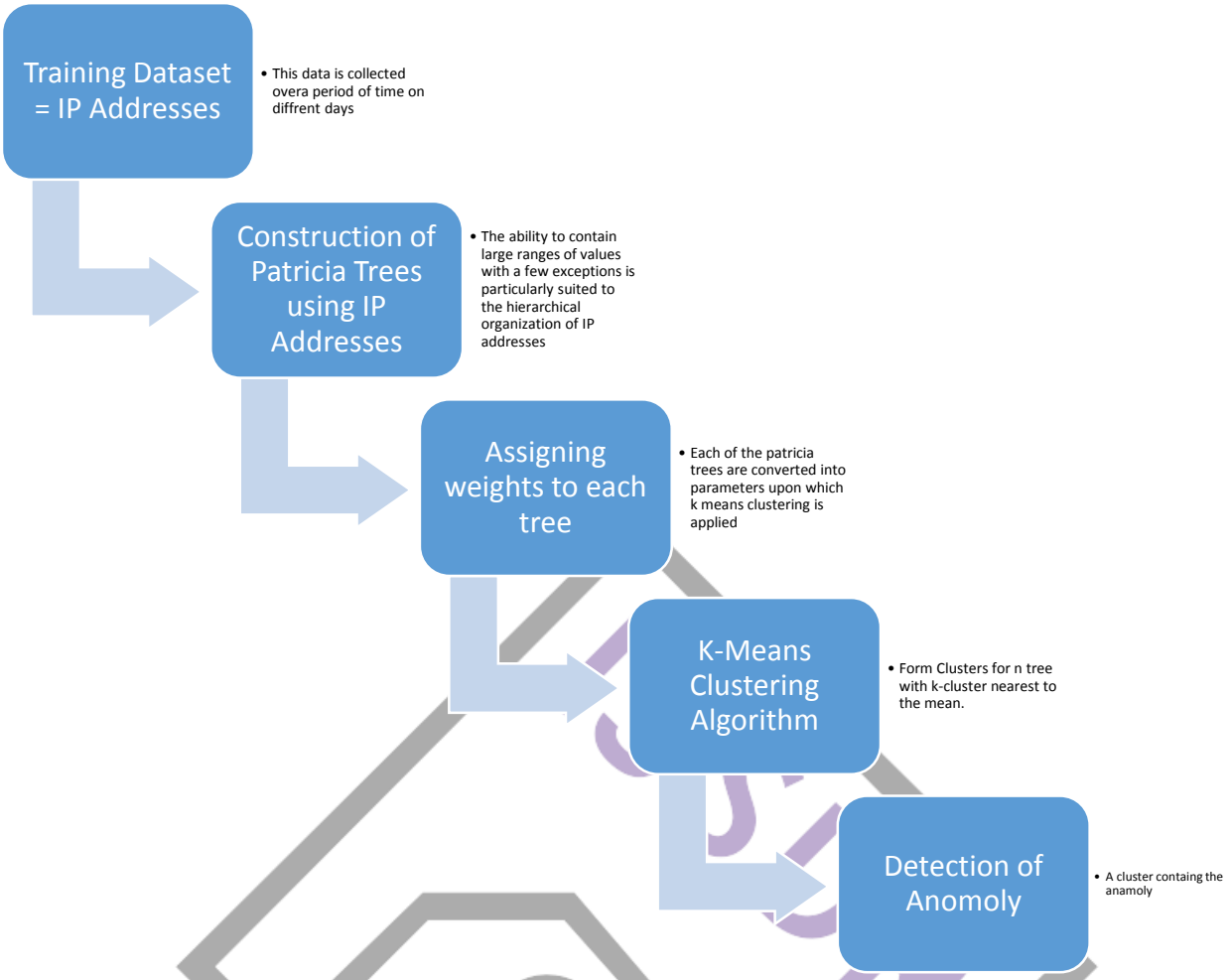

Figure 1 Proposed Solution

II. Materials and Methods

This section presents the materials and methods of this research work. It can be broadly classified into the following sections:

1. Collecting the IP addresses which will be used as the training dataset
2. Construction of Patricia trees and their analysis
3. K means clustering on converted parameters from Patricia trees

In the first step of collecting the data for the training dataset, we use the software 'myLanViewer' to get the network (IP) addresses of all the devices in the network. We have collected the data over a period of 7 days. Here we are presenting the sample screenshot for one particular day. For our test purpose, we have noted that on Day 2 of the dataset, a new device was introduced into the system.

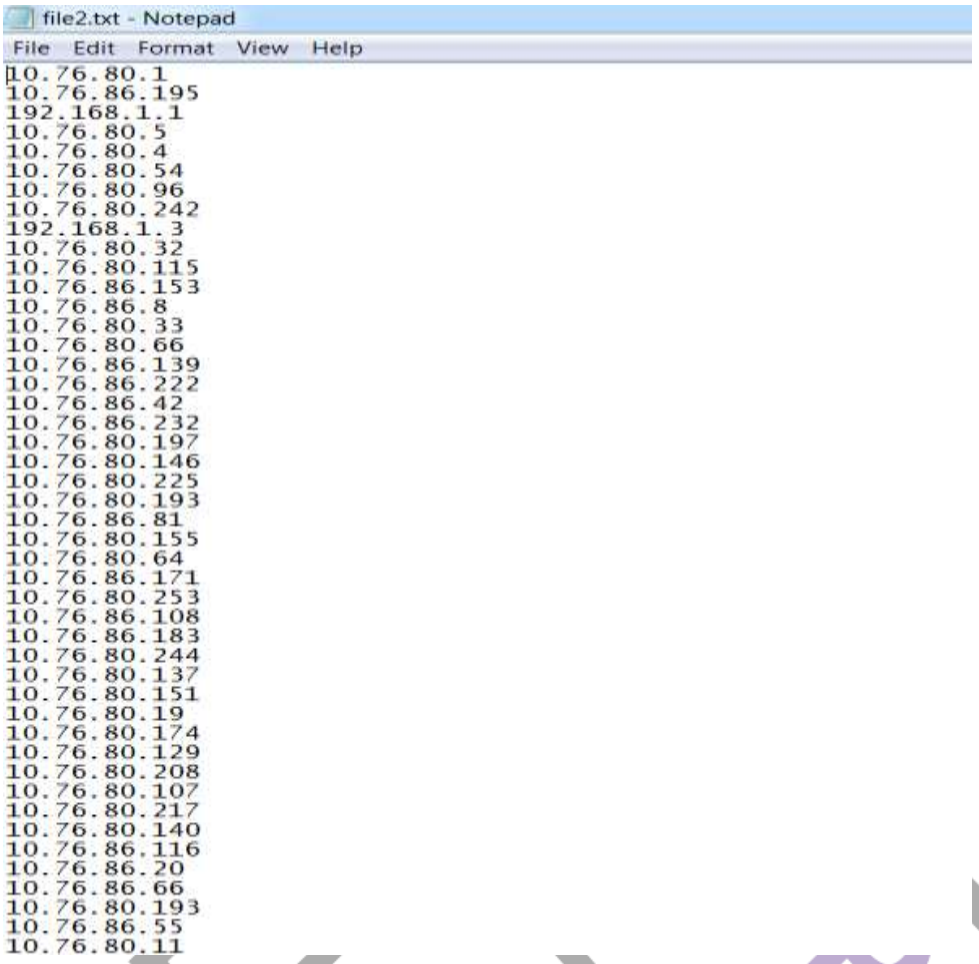

Figure 2 Training Dataset Snapshot

We notice that the subnet mask of the IP addresses remain almost the same. This can be taken advantage of in the construction of Patricia tree which looks for a common prefix.

The algorithm proceeds by reading these files and tokenizing the IP Addresses (strings) using the delimiter character as '.'. For example if the IP address is 15.213.91.33, on tokenizing we get 4 different strings as 15, 213, 91 and 33. Since the algorithm involves the use of array index as the value of the tokenized string, these strings have to be converted into integer datatype after tokenizing. A Patricia tree is then created based on these tokenized strings. We create a Boolean array of size 256 since the maximum value of each component of the subnet can be at most 2^8=256. The tokenized values are then traversed through and for each of the integer values obtained, the value at the index value is set to true.

The next step is to convert the constructed Patricia tree into two parameters, x and y, so that k-means clustering can be applied on the data. As mentioned before the X parameter is the number of leaf nodes that are present in the tree and the Y parameter is the total number of nodes that are present in the Patricia tree. X is calculated by counting the number of true values in the array and Y is calculated by counting the number of IP addresses present in the dataset as each of them correspond to a leaf node.

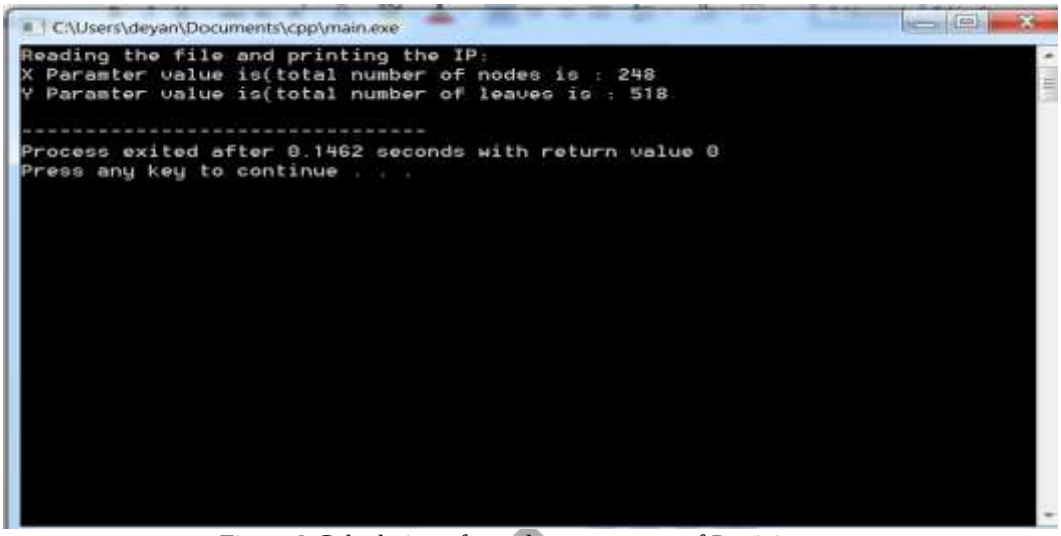

Figure 2 Calculation of x and y parameters of Patricia tree

The third step involves gathering the x and y values of the seven Patricia trees created and storing then in a csv file. This is then passed through the K means clustering algorithm.

Table 1 Table of X and Y values of each tree

DataPoint (Day of the week)	X (Number of Leaf Nodes)	Y (Total Number of Nodes)
1	248	518
2	135	138
3	234	505
4	228	477
5	199	329
6	211	358
7	196	301

On applying K means clustering with K=2 and number of iterations n= 10

Table 2 Table after applying K means clustering

DataPoint (Day of the week)	X (Number of Leaf Nodes)	Y (Total Number of Nodes)	Cluster Number
1	248	518	2
2	135	138	1
3	234	505	2
4	228	477	2
5	199	329	2
6	211	358	2
7	196	301	2

III. Results and Analysis

In this section we discuss the results and analysis. The x values of the Patricia trees are plotted along the X axis with the corresponding y values along the Y axis. We notice that there are two clusters obtained. The green data points denote cluster 2 as represented in table 2 while the red denotes cluster 1. This cluster contains only one data point and is hence the anomaly in the network.

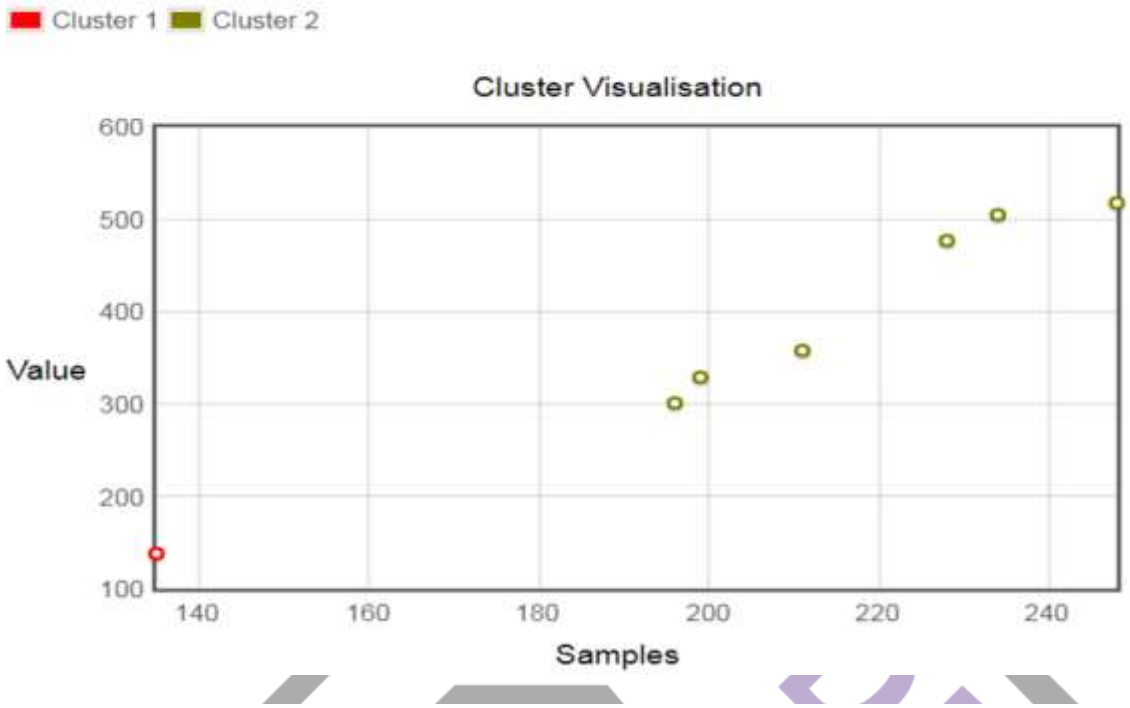

Figure 4 K mean clustering with k=2

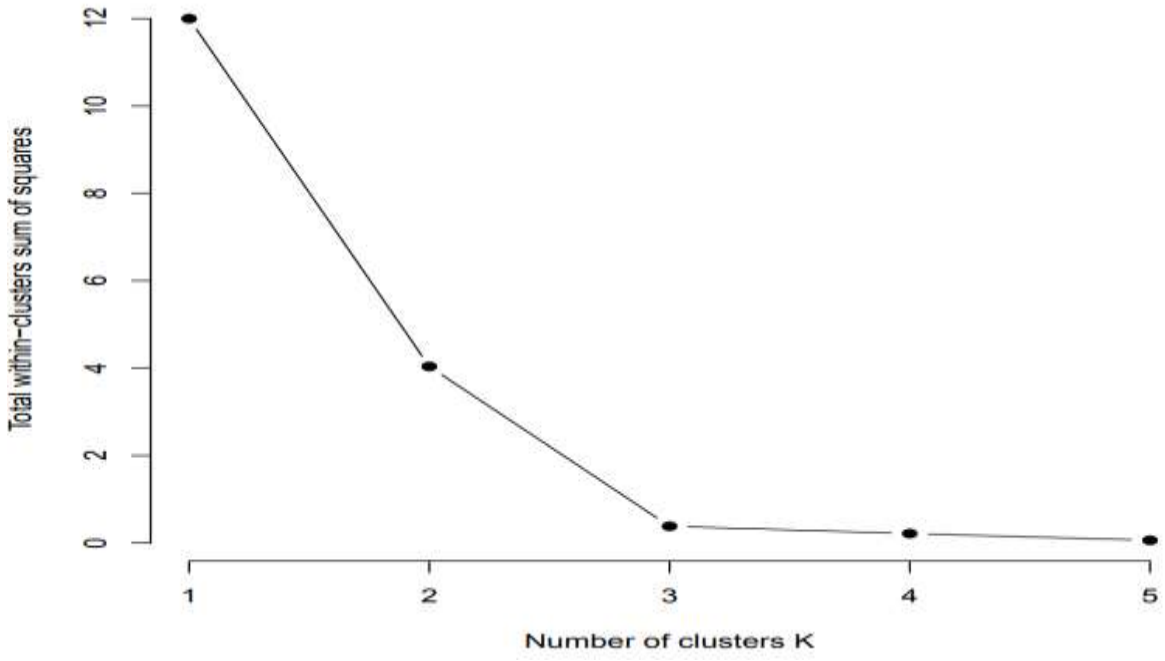

Figure 5 Elbow curve

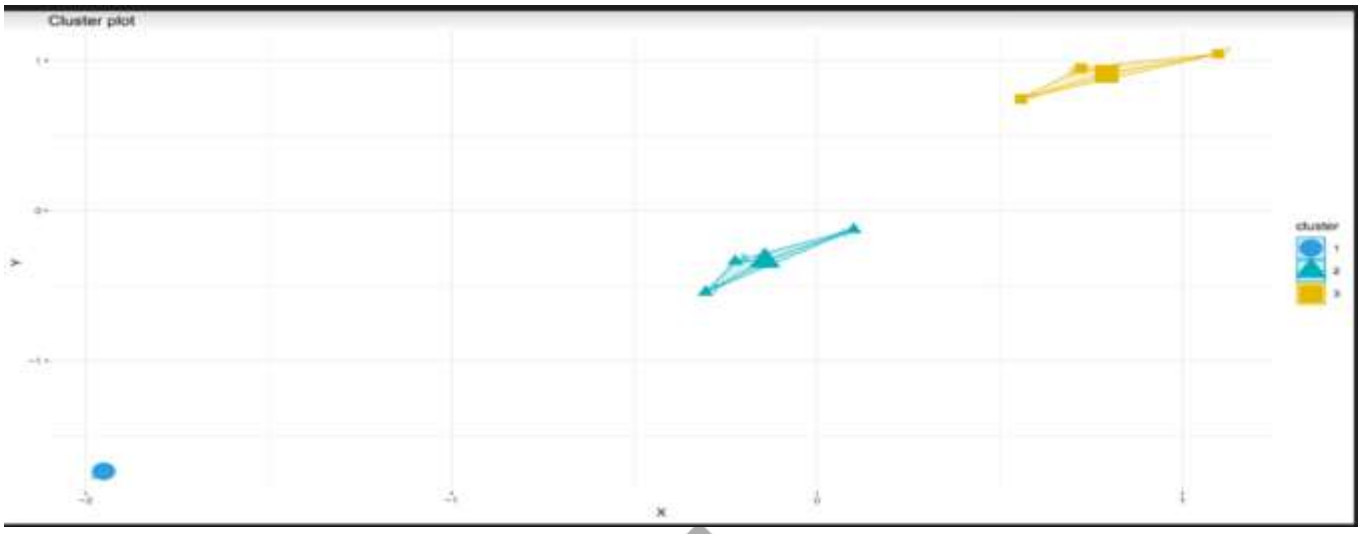

Figure 6 K means clustering with k=3

Based on this curve, we infer that using k=3 will give optimal results.

There are three clusters of data points and the first cluster denotes the data point 2 which is considered as the anomaly.

IV. Conclusion

Previous research has been done on Novelty detection in a network where ID3 decision tree learning methods has been used. This is a new method as it involves the concept of representation of IP addresses in a Patricia tree. The space and time complexity have been put into check for a large data set. Along with K means clustering technique, we have come to a conclusion of accuracy of 87 %. This is obtained by considering three more datasets of a seven day period and applying this technique.

References

[I].A novel unsupervised classification approach for network anomaly detection by k-Means clustering and ID3 decision tree learning methods Yasser. Yasami, Saadat Pour Mozaffari.

[II].Novelty Detection in Learning Systems. Stephen Marsland, Division of Imaging Science and Biomedical Engineering, Stopford Building, The University of Manchester, Oxford Road, Manchester M13 9PL, UK

[III].A One-Class SVM Based Tool for Machine Learning Novelty Detection in HVAC Chiller Systems. A. Beghi ∗ , L. Cecchinato ∗∗ , C. Corazzol ∗ , M. Rampazzo ∗ , F. Simmini ∗ , G.A. Susto ∗