Good morning. I’ve really been looking forward to meeting you.

I´m Ronak Agrawal. I have completed my Engineering in Computer science, while my graduation I have worked on the project “Distributed Intrusion detection system using Bayesian learning and Apache Mahout”. Apart from that I also have worked on Apache Hadoop, Linux OS and shell script. In cybage I have worked as a buffer on “Online Book Store” project.

Well briefing about my project “Distributed Intrusion detection system using Bayesian learning and Apache Mahout”. The Problem was to detect the intrusions in the distributed network. The objective behind making this application was to make the systems more secure on the basis of the intrusion detected by any system can be shared as an experience with other systems and this can be used to be aware of it next time.

With the growing era of internet, the network security has become the key foundation for a lot of financial and business web applications. Imperfectness of intrusion detection systems (IDS) has given an opportunity for data mining to make several important contributions to the field of intrusion detection. Here, we propose a new approach by utilizing data mining techniques such as neuro-fuzzy and Bayesian learning algorithm for helping IDS to attain higher detection rate. The proposed technique has four major steps: primarily, k-means clustering is used to generate different training subsets. Then, based on the obtained training subsets, different neuro-fuzzy models are trained. Subsequently, the probability computation of the intrusions based on the fuzzy rules is performed using Bayesian algorithm to detect intrusion has happened or not. And then these probabilities are passed to Apache Mahout for the collaborative filtering and recommendations for the next time depending on this experience.

Intrusion is nothing but an attempt made to harm the system or data remotely or malicious activity or policy violations, there are many different types of intrusions. In our project we mainly focused on three of them, they are DOS, DDOS, and R2L (Remote-to-Local). The technologies and algorithms used in developing this application are k-means clustering algorithm, Neuro-fuzzy logic, Bayesian Probability Algorithm, Apache Mahout. This project is developed in Java using NetBeans IDE and the database used is MySQL. It took us 6 months to design this application. While developing this application we also published 2 white papers related to the studies. The 1st paper “Distributed Intrusion Detection System using Bayesian learning and Apache Mahout” which was published in IJARCCE Journal is based on the research and the survey we did to choose the algorithms and technologies for our application. The 2nd paper “Heuristic Framework for Network Attack Detection Using Fuzzy ANN and Apache Mahout” which was published in IJARIIE Journal contains all the information of the developed application likewise architecture diagram, algorithm, hardware requirements, mathematical model, and results.

Now let me explain you the architecture and the working of the system in detail:

For detection of the intrusions we need a dataset which can be provided to our system as an input for further processing. For these we were having two options, one was to use standard KDD dataset which can be downloaded from the web easily, it is the dataset which contains all the network properties and some intrusions also, which can be used for train the smart systems. And the other option was to use the real time router data which we have to collect from the routers, in these we can analysis the specific or the actual network traffic which is incoming through the router. So we used real time router data over the KDD dataset, for making our system more efficient.

The dataset is available in the excel format, now we need to convert this data so that it can be used for the development. We used JXL API to convert the excel data to the Java objects, in this process each row of the dataset was converted to an object and stored in vector. The dataset was having 21 different fields (columns) of the network and such 50,000 records were taken.

After this we have to divide the data into different clusters so that it can be analysed more accurately. For this we used K-means clustering algorithm, initially we kept the number of clusters static that was 5 till the developments completes. But after that we made it dynamic input, which is the number of clusters was the user’s choice. The reason why we used K-means is one of the simplest algorithm which uses unsupervised learning method to solve known clustering issues. It works really well with large datasets and it took less time to form the precise clusters as compared to other algorithm (sequence clustering). There was one more reason as we have to use ANN (Artificial Neural Network) so for that it was required to use K-means clustering.

In the initialization state of K-means clustering as per the number of clusters (K) entered by the user the data points are randomly selected. After this the distance between each point and the data point is calculated and then depending on the mean of these distances the clusters are formed. After this the data points are shifted and again the mean distance is computed, these process goes on until the precise clusters are not form.

These clusters are then converted to the neurons for the interest ratio and the ANN processing. The number of neurons was set to 2 per cluster. That is total number of neurons is twice the number of clusters. After this the interest ratio was calculated per neuron. The interest ratio is nothing but the number of destination IP’s the source has communicated. Its output helped us to recognise the target, which the source was communicating.

Later the Fuzzy Ranges were set. We used 5 set of Rules starting from 0 to 100 on the range of 20 each set and then ANN was used, in this the neurons were passed through the entire rule base and the data from one neuron is shifted to the other if the similar interest ratio is found.

Then the Bayesian Probability of each object was calculated neuron wise, taking into consideration the priori probability of it from the dataset occurrences count. Further the obtained result was filtered on the basis of number of occurrences and also the objects whose probabilities are greater than 1 is optimised to 1. Now we get the filtered optimised Bayesian Probability.

The result of the Bayesian Probability shows the probability of the intrusion occurring in the network and the level of harm which can be made by it to our system. And then it is provided to the Apache Mahout, which we are using for the recommendations based on the past experiences. Apache Mahout is built on the top of the Apache Hadoop and is used for the Machine learning and scalable computation. We have used “Taste algorithm” of Apache Mahout for project. The Taste is collaborative-filtering recommender component of Mahout was originally a separate project and can run stand-alone without Hadoop.

Apache Mahout stores all the results to the database (here we are using MySQL database) so that it can be shared with all the users in the network and can be saved for the later use by the system. The Apache Mahout filters out the source IP’s which are causing intrusions in the network and stores it.

The intrusions detected were of three different kinds that are DOS, DDOS, and R2L. The source IP’s for all the different intrusions were bifurcated and can be seen in the graph as well as in the list format.

It was very nice to meet you; I hope we meet again soon.