HIGH PERFORMANCE INTRUSION DETECTION USING MODIFIED k-MEAN & NAÏVE BAYES

Anupam Sakhi[1], Bhawna Mallick[2]

[1]Department of Information Technology, Galgotia College of Engineering and Technology, Greater Noida, India
[2]Head of Department, Department of Computer Science and Information Technology, Galgotia College of Engineering and Technology, Greater Noida, India

Abstract

Internet Technology is growing at exponential rate day by day, making data security of computer systems more complex and critical. There has been multiple methodology implemented for the same in recent time as detailed in [1], [3]. Availability of larger bandwidth has made the multiple large computer server network connected worldwide and thus increasing the load on the necessity to secure data and Intrusion detection system (IDS) is one of the most efficient technique to maintain security of computer system. The proposed system is designed in such a way that are helpful in identifying malicious behavior and improper use of computer system. In this report we proposed a hybrid technique for intrusion detection using data mining algorithms. Our main objective is to do complete analysis of intrusion detection Dataset to test the implemented system.In This report we will propose a new methodology in which Modified k-mean is used for clustering whereas Naïve Bayes for the classification. These two data mining techniques will be used for Intrusion detection in large horizontally distributed database.

Keywords: *Intrusion Detection, Modified K-Mean, Naïve Bays*

----------***----------

1. INTRODUCTION

We integrated modified dynamic k-means (MDKM) clustering algorithm and Naïve Bayes classification technique, our aim is to improve the detection rate and decrease the false alarm rate. The MDKM algorithm filters the noise and isolated points on the data set then by calculating the distances between all sample data points, we obtain the high-density parameters and cluster-partition parameters, and using dynamic iterative process we get the k clustering center accurately [1], [3].

Naive Bayes classifier that is used to identify possible intrusions. Classification is a classic data mining technique based on machine learning. Classification is used to classify each item in a set of data into one of predefined set of classes or groups. Naïve Bayes is a commonly used classification supervised learning method to predict class probability of belonging. This report proposes a new method of Naïve Bayes Algorithm in which we tried to find effective detection rate and false positive rate of given data. We tested the performance of our proposed algorithm by employing KDD99 benchmark network intrusion detection dataset, and the experimental results proved that it improves detection rates as well as reduces false positives for different types of network intrusions [2], [4].

In this paper we proposed a method for arrangement of intruder in framework Intrusion detection is the real undertaking in data mining. There are such a large number of arrangement gave by the scientists to recognition of interloper in the information. Like, Pattern Matching, Measure Based method, Data Mining method and Machine Learning Method. Here we identified intrusion through data mining strategy by joining two data mining system Modified K mean and Naive Bayes grouping and shaped a half and half method. We consolidated these distinctive techniques for measured diverse parts of intrusions. Combined these guidelines discover the interloper attack all the more rapidly from the existing one. We tested the performance of our proposed algorithm by employing KDD99 benchmark network intrusion detection dataset, and the experimental results were better or comparable to other methods [2], [3]. In This report we will propose a new methodology in which Modified k-mean is used for clustering whereas Naïve Bayes for the classification. These two data mining techniques will be used for Intrusion detection in large horizontally distributed database.

2. RELATED WORK

There has been multiple researches towards the improvement of IDS technology as we are dealing with larger data chunks. The performance and efficiency are the two contradicting parameters for the Intrusion detection systems. Therefore trade-off need to be done for the development of system for the specific cases. For larger databases, performance takes the edge over efficiency as the turnaround time is very important and critical even if it is unable to detect few intrusion. The performance improvement is the goal of the project. The Objective of the project is to develop the complete Intrusion Detection system. One of the aim is to develop a java Based Platform (Software) for the Handling the workforce of a company. Then an Intrusion Detection system will be implemented specific for the developed system. The Testing methodology has been framed and implemented using various tools like SQL, Java, Shell Scripting. The Modified KDD'99 Datasets

will be used for the testing. The Datasets will be made suitable for our developed system. Finally, the testing results will be presented and compared with benchmarking dataset KDD'Cup-99. The aim of the project is to develop the complete flow so as this can be used at multiple places for reference.

Table 1 below summarizes the findings in the above presented literature. Extensive survey has been done to gather the progress in IDS in past 10 years and a concluding summary is presented below.

Table 1: Comparison Table on Literature Survey

Sr. No.	Approach	Pros	Cons
1	Hybrid PSO + C4.5, SNORT + ALAD + LERAD	High accuracy rate	Cannot be applied to real traffic, Increase complexity
2.	SNORT + PHAD + NETAD	more efficient than SNORT	Cannot detect behavioral attacks
3	Entropy + SVM classifier	Define network properties	Cannot process large data
4	K-Means, C4.5, K-Means + C4.5	High accuracy rate	Cannot process large database
5	CART, Bayesian Model, Artificial	Accuracy of CART is more	Do not support large database

3. PROPOSED SYSTEM ARCHITECTURE

This section describes about result analysis of the proposed scheme for intrusion detection. The parameters accuracy, time to build model, precision and recall are analyzed. I have proposed a system with Hybrid technology. Using Modified K-Mean (MDKM) with Naïve Bayes gives best result with large dataset. There are many system which is using k-means based algorithms but they have weakness for high dimensional data. Secondly, in spite of non-hyper spherical distribution of normal traffic in a feature space, these algorithms can only create hyper spherical clusters. These are the motivations behind using MDKM with Naïve Bayes. In this report we proposed a method for arrangement of intruder in framework Intrusion detection is the real undertaking in data mining. There are such a large number of arrangement gave by the scientists to recognition of interloper in the information. Like, Pattern Matching, Measure Based method, Data Mining method and Machine Learning Method. Here we identified intrusion through data mining strategy by joining two data mining system Modified K mean and Naive Bayes grouping and shaped a half and half method. We consolidated these distinctive techniques for measured diverse parts of intrusions. Combined these guidelines discover the interloper attack all the more rapidly from the existing one. We tested the performance of our proposed algorithm by employing KDD99 benchmark network intrusion detection dataset, and the experimental results were better or comparable to other methods

In this paper we proposed an algorithm that functions admirably with large datasets.

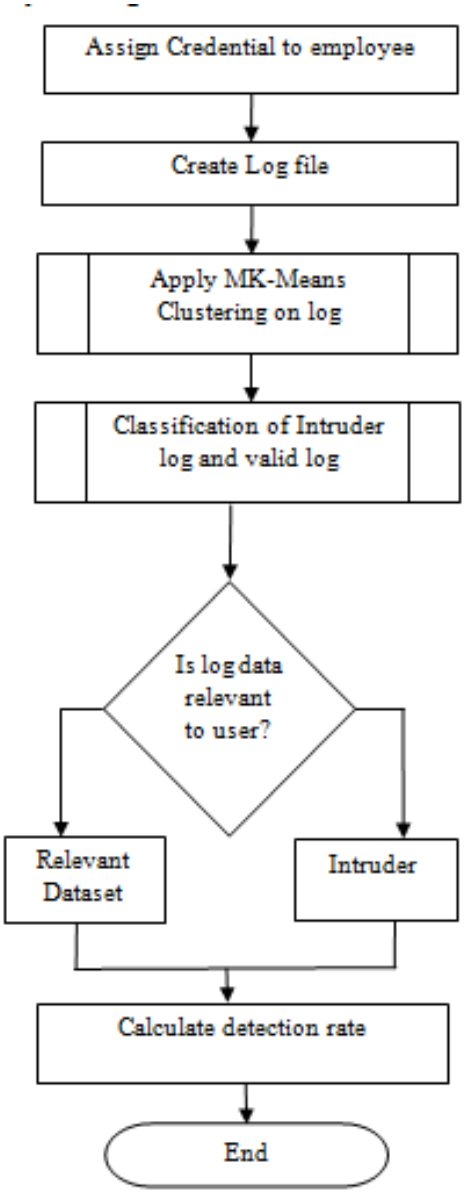

Fig 1: Flow Diagram of the Algorithm

4. RESULTS & DISCUSSION

The KDD99 cup dataset was used in the 3rd International Knowledge Discovery and Data Mining Tools Competition for building a network intrusion detector, a predictive model capable of distinguishing between intrusions and normal connections. As stated earlier, performance improvement has been the major goal of the project. To evaluate the results of classifier, we have used standard metrics such as confusion matrix, true-positive rate, false positive rate, and classifier's accuracy. The Confusion Matrix is given below:

Table 2 : Confusion Matrix

	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Recall = \frac{TP}{TP + FN}$$

$$FNR = \frac{FN}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

Table 3: Accuracy Results for Training Dataset

Total Number Of Instances	**10000**	**% of Accuracy**
Correctly Classified Instances	9913	99.13
Incorrectly Classified Instances	87	0.87

Table 4: Confusion Matrix for Subset Dataset

A	**B**	**Classified as**
9087	**23**	**Normal**
56	**8913**	**Attack**

Table 5: Detailed Accuracy Results for Subset Dataset

TP rate	**FP Rate**	**Precision**	**Recall**
0.993875	0.002525	0.978375	0.99563

5. CONCLUSION

In this paper we proposed a method for classification of intruder in system Intrusion detection. Here we detected intrusion through data mining method by combining two data mining technique Modified K means and Naive Bayes classification and formed a hybrid technique. We combined these different methods for measured different aspects of intrusions. Combined these rules find the intruder attack more quickly from the exiting one. We have successfully implemented the Employee task assignment interface as well as Interface for the employee to simultaneously work on the assigned file. Multiple security measures has been implemented to reduce the threat of corrupting the stored data on the system. A unique key is assigned to each user for each session and therefore assigned file can be modified only if user provide the security key correctly. This gives the system first level of security in horizontally distributed database system. The system has been made immune to the major attacks by developing the Intrusion detection system using two well know data mining algorithm Modified k-mean and Naïve bayes Algorithm. We have obtained excellent results for the implemented system and Dataset.

ACKNOWLEDGEMENT

I want to express my sincere gratitude towards my supervisor Dr. Bhawna Mallick. I would like to thank her for giving me the opportunity to work on this project. I am extremely grateful to my supervisor for being more than a mentor. She always provided encouragement and needful advice, which took me out of confusions and always inspired me. The authors would like to express their gratitude to the "Galgotia college of engineering and Technology", for providing such an excellent infrastructure & facilities.

REFERENCES

[1]. Praveen P Naik, Prashantha S J,"An Approach for Building Intrusion Detection System by Using Data Mining Techniques", International Journal of Emerging Engineering Research and Technology Volume 2, Issue 2, May 2014, PP 112-118,

[2]. Richa, Saurabh Mittal, "Data Mining Approach IDS K-Mean using Weka Environment", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 8, August 2014 ISSN: 2277 128X,

[3]. Shubhangi S. Gujar and B. M. Patil, "INTRUSION DETECTION USING NAÏVE BAYES FOR REAL TIME DATA", International Journal of Advances in Engineering & Technology, May, 2014.

[4]. Ravi Ranjan and G. Sahoo, "A New Clustering Approach for Anomaly Intrusion Detection", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.4, No.2, March 2014,

[5]. Gitanjali J, C.Ranichandra, M.Pounamba, "APRIORI algorithm based medical data mining for frequent disease identification", IPASJ International Journal of Information Technology (IIJIT), Volume 2, Issue 4, April 2014 ISSN 2321-5976,