ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 11, November 2018

A Cloud Based Intrusion Detection System Using BPN Classifier

Priyanka Alekar
Department of Computer Science & Engineering SKSITS, Rajiv Gandhi Proudyogiki Vishwavidyalaya
Indore, Madhya Pradesh, India

Abstract: **- The intrusion detection systems are developed in order to monitor the network activity and tack the malicious interruption in network. In this context the generated traffic data is need to be analyzed. In most of the IDS design for traffic data analysis machine learning and data mining algorithms are utilized. These algorithms are able to deal with the data for discovering the valuable or target patterns from data. But in network the data is generated and communicated in a small amount of time and the algorithms consumes a significant amount of time. In this context different feature selection approaches are implemented with the IDS systems which reduce the dimensions of the data. With the fewer amounts of data attributes learning algorithms are trained well but detection performance can be affected. On the other hand for efficient computing and dealing with large amount of data the cloud computing is used. The cloud computing provides scalable resources and efficient computing methodology for process the data is fewer amount of time. Therefore in this presented work both the computing techniques namely cloud computing and data mining both are used for improving the performance and working of traditional IDS systems. The proposed data model works in client server architecture where the server performs the computations and produces the decision by analysis of client submitted traffic samples. For training and testing of data model the SVM (support vector machine) and BPN (back propagation neural network) is used. Additionally for experimentation the KDD CUP data is consumed. The experimental results show the proposed technique reduces the overhead of client end computing efforts and enhance the decision capability of IDS systems.**

Keywords: - IDS (Intrusion Detection System), data mining, cloud computing, BPN (Back Propagation Neural Network), SVM (Support Vector Machine).

I. INTRODUCTION

Intrusion detection system can be hardware, software or both that monitors the network activity to distinguish the malicious behavior or violations of network protocols. The IDS systems report such kind of malicious activity to the administrator of network to keep in track the security of the entire network. In this presented work the data mining technique based intrusion detection system is proposed and implemented. There are a number of different kinds of techniques for intrusion detection is available but the data mining techniques are more profitable as compared to other techniques. In addition of that the analysis of network data is performed on cloud server end to reduce the cost of analysis and detection.

The proposed intrusion detection system works on a client server model. The client connect through the server database, and the server consist of the trained data mining algorithm which accept the network traffic pattern and then classify it to the predefined class labels. Basically the data mining techniques are used for analyzing the data and obtaining the application oriented patterns. These captured patterns after the data analysis can be used for decision making, prediction and other similar task. In addition of that the data mining algorithms are able to learn with some specific kind of data patterns therefore it can be used for pattern understanding and recognition. In this context the proposed data mining technique based model first learn from the training samples and then recognize the similar patterns when it appears. In order to perform this task the supervised learning techniques are used. The supervised learning techniques are efficient and accurate for recognition and prediction.

II. PROPOSED WORK

This chapter aimed to explain the proposed technique of intrusion detection system development. Therefore first the requirements of the cloud based intrusion system are explained and then the effort is made to explain the proposed working model of IDS system design.

A. System Overview

The IDS (intrusion detection system) is a security system that is developed on the basis of hardware or software or using both of the combination. The IDS system evaluates the network traffic and the hidden patterns for discovering the different patterns malicious or violation based patterns in the network traffic. But the network data analysis needs efficient algorithms by which the performance of network can be maintained and efficiently system can classify or distinguish the malicious activities in network. In this context the dimension of data played an important role because the higher dimensional data need more resources and processing time for classification. But the reduction of useful attributes from the network traffic data can impact on security of network systems. Therefore we need to develop an efficient system that works on large amount of data in less amount of time. In this context the cloud computing offers the scalable and efficient computing that deals with a large amount of data and with the less amount of time and other resources consumption. Therefore in this presented work a cloud based IDS model is proposed for design and implementation. The proposed model works on the concept of client server architecture. In this system the server system is responsible to collect the network traffic sample, process the network data and detect any malicious activity in the network. Therefore the server contains the machine learning algorithm that trained on the malicious and normal traffic samples. Additionally after training these algorithms are able to distinguish between malicious patterns and normal patterns. Therefore client end just collect and send the network samples to servers and take the appropriate decisions according to the requirements. In this section the basic overview of the required system is described. In next section the modeling of the required data model is provided and the detailed explanation is offered.

B. Proposed Methodology

This section explains the proposed system using the architecture of the system and how the proposed model performs training and testing.

System architecture

The proposed system architecture for the cloud based intrusion detection system is demonstrated in figure 2.1. The proposed system architecture contains four major components all the components are explained as:

Remote network: that is the target network on which the monitoring required. Therefore either all the systems in network can access the IDS or a single machine by which the traffic goes in and out can access the IDS system. In this context the network nodes or node collect the traffic samples and forward to the server for finding the decision or server.

Cloud server: the cloud server is a remote server which accepts or connects with the outer network computers which are accessing the IDS services. Additionally the server accept the input traffic classify the network samples and develops the decisions for finding the malicious activity or not in the target network. Therefore the server machine contains the machine learning algorithm which train on the network samples and find the similar patterns in submitted traffic.

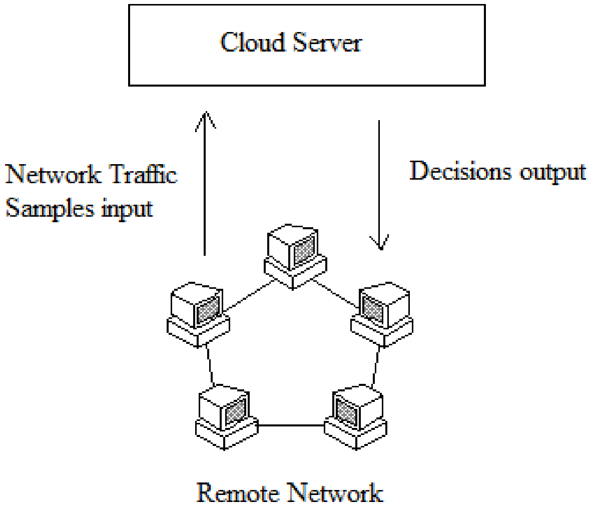

Figure 2.1 proposed system architecture

Network Traffic samples: in this presented demonstration the traffic samples are not collected in real time. For experimentation and simulation purpose the KDD CUP 99's Dataset is considered. That is a standard data set which is used for IDS system design. It is taken from UCI machine learning repository for training and testing of the proposed working model.

Decision outputs: after classification of network samples the server system generates the classification labels for each input traffic sample these class labels are returned as decision of server.

b. Server Training and Testing

In this section the information is reported for providing the functional description of the proposed system training and their testing with the input data. Figure 3.2 shows the functional scenario of the proposed model for training and testing.

Input Training samples: the training samples are taken from the UCI machine learning library which contains the KDD CUP 99's Dataset. Basically this dataset is a kind of network traffic samples which contains 41 features and one attribute as the class label. The class label of data contains the type of attack pattern according to the traffic sample attributes. In this context the KDD CUP dataset is provided as input to the system for initial learning and data processing.

Data Pre-processing: the data preprocessing is an essential step of data mining and machine learning. The preprocessing on data is performed for improving the quality of data in order to make it suitable for learning algorithm. Therefore the ambiguous nature of data instances are removed and filtered in this process. In addition of that those attributes are identified which are not suitable for learning. In this presented work the preprocessing of data is performed for removing those data instances which are containing the missing values or noisy in nature. In addition of that the attributes are mapped text values to the relevant numerical values. After preprocessing of data the entire dataset is partitioned in two parts testing dataset and training dataset.

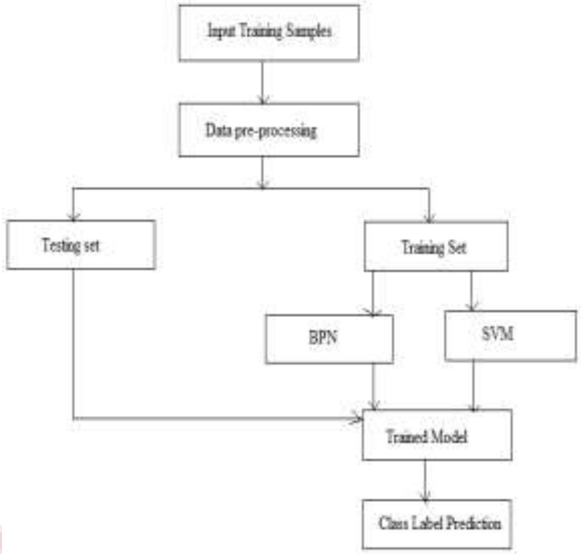

Figure 2.2 proposed classification system

Training set: training set is collection or set of data instances which are randomly selected from the entire dataset. The 70% of data instances are randomly picked from initial input samples which are used with the learning model.

Testing set: after creating the training set the remaining 30% of data samples are used for testing of the system or trained classifier.

BPN: the BPN (back propagation neural network) is a machine learning model which is used for both the purpose i.e. classification and prediction purpose. The neural network developed with the tree layer architecture the first layer is known as the input layer. The input layer contains the similar number of neurons as the number of input attributes. The second layer is termed as hidden layer. The hidden layer is used for computation and the final layer is the output layer. The computed outcomes of the neural network are collected using the output layer. In output layer the error of output is computed and corrected using weight adjustment of input layer and hidden layer.

In our proposed model the 42 total attributes are available in dataset, among 41 attributes are the patterns of the data and the 42th attribute is the output of sample. That is also used for correction of weights which is used for training. Additionally after training the class labels are need to be predicted.

SVM: SVM (support vector machine) is also works for classification and prediction both. But the key issue of this model is that, it is just used for binary classification. In other words the SVM can be used for classifying the data samples in two major groups. The SVM data model considers each data object as a point in space. The space can be n-dimensional and depends upon the number of attributes available in dataset. The SVM algorithm tries to fit a hyper plane between the scattered points in space in such manner by which the two different types of samples can be distinguishable based on the partitioning of entire space. In our model the 41 attributes means there are 41 dimensions in space and in the each object in dataset are considered as the point in space. During the training the hyper plane is optimized to place in space for make separable these points optimally.

Trained model: basically when the data is processed using the machine learning algorithms the weights are adjusted or the distance is computed. Based on these computations the mathematical demonstration of model is prepared. Those mathematical models accept the similar kinds of pattern and fit on the model computation to get the possible class labels from data.

Class label prediction: the input samples are produced to the train model and based on training the model predict the class labels for each input samples. These class labels are the final outcome of the IDS system. Using the computed classes of the input data objects the accuracy and error rate of the system is described.

C. Proposed Algorithm

This section provides the summary of the proposed system design in form of algorithm steps. The table 2.1 contains the algorithm steps:

```
Input: Training Samples Tr, Test samples Ts
Output: class labels C

Process:
    1. $R_n = readTrainingSet(T_r)$
    2. $P_n = preProcessData(R_n)$
    3. $Train_{model} = BPN.Train(P_n)$
    4. $for(i = 1; i \leq T_s.lenght; i++)$
           a. $C = Train_{model}.Predict(Ts_i)$
    5. $end\ for$
    6. Return C
```

Table 2.1 proposed algorithm

III. RESULT ANALYSIS

This chapter provides the discussion about the evaluation of the proposed cloud based IDS system. During evaluation of the system the different parameters are obtained and reported in this chapter.

A. Time Consumption

The amount of time consumed for accepting the input test data for predicting the class labels for data is termed here as the time consumption of the proposed system. That is difference between the algorithm start time and end of time of data processing. Additionally that can also be computed using the following formula:

Time consumed=algorithm end time-start time

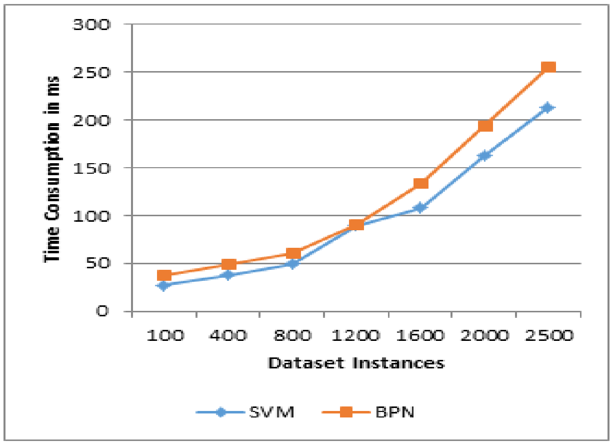

Figure 3.1 time consumption

Dataset Size	SVM	BPN
100	27.4	38.1
400	38.2	49.4
800	49.5	60.6
1200	89.8	91.2
1600	108.3	133.7
2000	162.9	194.1
2500	213.2	255.8

Table 3.1 time consumption

The time consumption of both the models namely SVM (support vector machine) and BPN (back propagation neural network) is described using figure 3.1 and table 3.1. In this diagram the comparative performance of both the model is represented. The X axis of diagram contains the dataset size used for experimentation and the Y axis contains the respective time consumed. The time consumption of the algorithms is calculated here in millisecond (ms). According to the obtained results the experiments the proposed BPN based classification technique utilizes higher amount of time as compared to the SVM based classification model. The higher amount of time is required to properly train the developed model.

B. Memory usages

System allocates an amount of memory to running processes. The running process consumes the amount of allocated memory resource is termed here as the memory usages or memory consumption. The amount of memory required to execute the given task is known as the space complexity. Using the JAVA technology it can be calculated using the following formula:

Memory Usage=Total allocated memory-total free memory

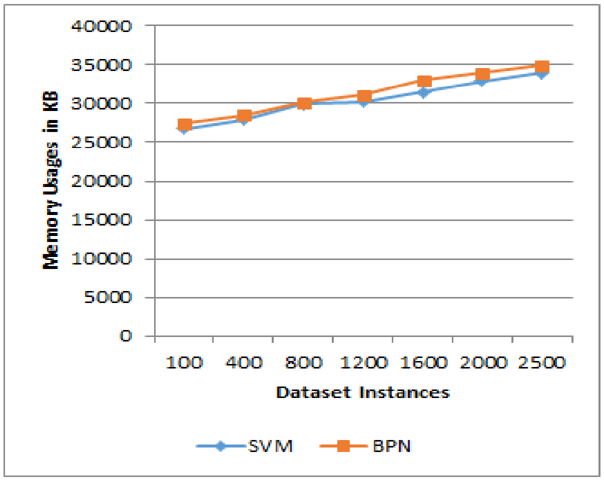

Figure 3.2 memory usages

Dataset Size	SVM	BPN
100	26748	27473
400	27947	28471
800	29948	30119
1200	30184	31084
1600	31483	32994
2000	32948	33975
2500	33957	34929

Table 3.2 memory usages

The memory usages of two machine learning models for intrusion detection are compared using figure 3.2 and table 3.2. The table contains the values obtained from different set of experiments and with the increasing amount of data. Additionally the diagram contains the graphical representation of obtained values from experiments. The X axis of this diagram contains the dataset instances and the Y axis provides the memory consumed with the given algorithms for the same set of data instances. According to the obtained results the SVM classifier consumes less amount of memory as compared to the proposed BPN based technique for training and testing.

C. Accuracy

The accuracy is the measurement of correctness of a data mining or machine learning system. The accuracy of demonstrate how accurately a data mining model classify the input unlabelled data after training. The accuracy of a data model can be computed using the following formula:

$$Accuracy = \frac{Total\ Correctly\ classified\ instances}{total\ instances\ to\ be\ classify} X100$$

The accuracy of classification for both the data models are described in figure 3.3 and table 3.3. The figure contains the line graph for demonstration of the proposed system performance in terms of accuracy. The accuracy of the system is measured here in terms of percentage. The X axis of the given diagram contains the number of data instances in data set for experimentation and the Y axis shows the performance in terms of accuracy of both the models. According to the obtained experimental results the proposed BPN based intrusion detection system provides improved accuracy as the amount of data is increases. Additionally the SVM based model provides the similar patterns of results but that is less then BPN classification model.

Data set instances	SVM	BPN
100	87.5	89.3
400	88.3	90.1
800	89.1	90.9
1200	90.8	91.7
1600	90.4	92.4
2000	91.2	93.9
2500	92.6	94.6

Table 3.3 accuracy of both models

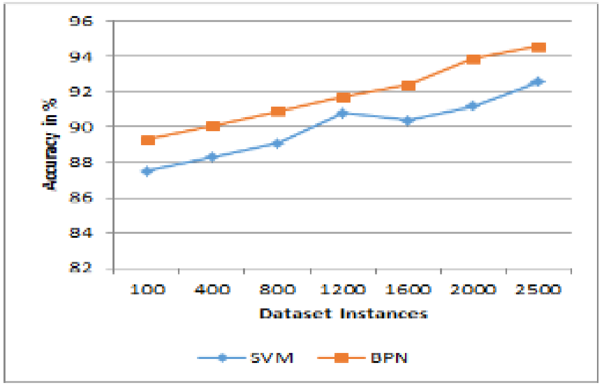

Figure 3.3 accuracy of IDS

D. Error Rate

Error rate is the measurement of incorrectly recognized data instances. Therefore the amount of data which are not correctly recognized using the trained model is termed here as error rate of the system. The error rate of a data mining model can be computed using the following formula:

$$Error\ rate = \frac{total\ misclassified\ data\ instance}{total\ samples\ to\ classify} X100$$

Or

$$error\ rate = 100 - accuracy\ \%$$

Dataset Size	SVM	BPN
100	12.5	10.7
400	11.7	9.9
800	10.9	9.1
1200	9.2	8.3
1600	9.6	7.6
2000	8.8	6.1
2500	7.4	5.4

Table 3.4 Error rate

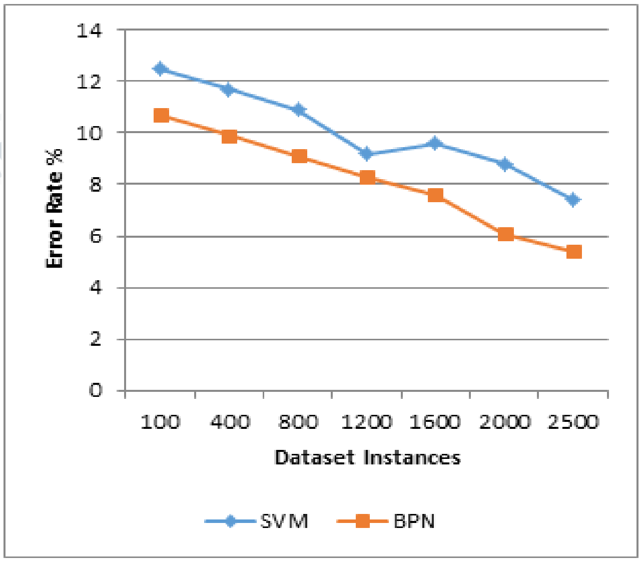

Figure 3.4 Error rate percentage

The percentage error rate obtained from experiments is given in figure 3.4 and table 3.4 for both the classification models. According to the diagram representation the X axis contains the amount of data instances in increasing manner. Similarly the Y axis contains the observed error rate percentage values for the amount of dataset instances. According to the obtained results the proposed BPN based technique produces less amount of error rate as compared to the SVM based technique. Thus the proposed data model is acceptable for application usages.

IV. CONCLUSION

This chapter provides the overall conclusion of the made effort for successfully design and implementation of cloud based intrusion detection system. The conclusion includes the facts about the experiments and observations. Similarly the future work is also provided in this chapter.

A. Conclusion

Data mining and their techniques enable computers to compute the data patterns automatically for discovering the trends in data. These trends can be used for different intelligent task which are normal not suitable to be perform using pen and paper. On the other hand the cloud computing is a new generation computing technology that makes computation easy and efficient as required by the applications. Cloud is a scalable, efficient and secure for working with large amount of data. Additionally that is capable to host and process the significant amount of information in a little amount of time. If both the concepts are combined each other then revolution in commuting technology can be obtain. In this context the proposed work is intended to develop a machine learning based IDS (intrusion detection system) that works on cloud servers for better efficiency and accuracy.

The cloud computing is enable us to perform the computation efficiently from the remote host. In this context a user provide instructions from remote end and server process the client computational data and produces the results directly to the client. Using this concept the large amount of task can be processed in less amount of resource consumption for the end client. Therefore the proposed work developed as a client server based approach that works for intrusion detection. The server system contains the intelligent algorithm namely SVM (support vector machine) and the BPN (back propagation neural network). Using the training samples both the algorithms is trained. After training of these machine learning algorithms the client system forward the traffic details as test set to the server system. The server system accepts the input traffic pattern and classifies them to identify the attack pattern or normal traffic pattern.

The implementation of the proposed intrusion detection system using data mining technique is performed with the help of JAVA technology. After implementation the experiments with the system is conduced and based on the experimental analysis the performance summary is prepared which is reported in table 4.1.

S. No.	Parameters	SVM	BPN (Proposed)
1	Memory usages	26748 KB - 33957 KB	27473 KB – 34929 KB
2	Time consumption	27.4 MS – 213.2 MS	38.1 MS – 255.8 MS
3	Accuracy	87.5 % - 92.6 %	89.3 % - 94.6 %
4	Error rate	7.4 % - 12.5 %	5.4 % - 10.7 %

Table 4.1 performance summary

According to the obtained performance as listed in table 4.1 the proposed technique of the intrusion detection using BPN algorithm is acceptable for accurate classification of the attacks. But it is costly in terms of resource consumption. Therefore near future need to minimize the time and space complexity of the proposed model server side.

B. Future Work

The main aim of the proposed work is to design and develop a cloud based intrusion detection model using data mining techniques which is accomplished in this work. For the future extension of the proposed work the following directions are suggested.

1. The proposed model implements the data mining classifiers in near future the ensemble learning technique is proposed for more accurate classification of attack samples
2. The proposed system works on the basis of traditional data mining models therefore extension of the given work can be performed using new machine learning techniques i.e. deep learning.

REFERENCES

1. Suraj Pandey, Kapil Kumar Gupta, Adam Barker, Rajkumar Buyya, “Minimizing Cost when using Globally Distributed Cloud Services: A Case Study in Analysis of Intrusion Detection Workflow Application”, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.151. 6055& rep=rep1&type=pdf

2. Foster, Z. Yong, I. Raicu, and S. Lu, "Cloud Computing and Grid Computing 360-Degree Compared," in Grid Computing Environments

Workshop, 2008. GCE '08, 2008, pp. 1-10

3. “Introduction to Cloud Computing”, Dialiogic, available online at: https://www.dialogic.com/~/media/products/docs/whitepapers/12023-cloud-computing-wp.pdf

4. Carroll, Mariana, Alta Van Der Merwe, and Paula Kotze, "Secure cloud computing: Benefits, risks and controls", In Information Security South Africa (ISSA), 2011, pp. 1-9. IEEE, 2011.

5. Introduction to Data Mining and Knowledge Discovery, Dunham, M. H., Sridhar, S., “Data Mining: Introductory and Advanced Topics”, Pearson Education, New Delhi, 1st Edition, 2006.

6. Phridvi Raj MSB., GuruRao CV (2013) Data mining – past, present and future – a typical survey on data streams. INTER-ENG Procedia Technology 12, pp. 255 – 263

7. Veepu Uppal and Gunjan Chindwani, “An Empirical Study of Application of Data Mining Techniques in Library System”, International Journal of Computer Applications (IJCA), Volume 74– No.11, July 2013.

8. Delmater R and Hancock M, Data Mining explained-a manager’s guide to customer-centric business intelligence (Digital Press, Boston) 2002.

9. Aakanksha Bhatnagar, Shweta P. Jadye, Madan Mohan Nagar” Data Mining Techniques & Distinct Applications: A Literature Review” International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 9, November-2012

10. Industry Application of data mining, available online at: http://www.pearsonhighered.com/samplechapter/0130862711.pdf

11. Jiban K Pal, “Usefulness and applications of data-mining in extracting information from different perspective”, Annals of Library and Information Studies, Vol-58, March 2011, pp. 7-16.

12. D. Denning, An intrusion-detection model. Journal of Graph Theory, SE- 13(2): pp. 222–232, 1987.