FALSE POSITIVE REDUCTION BY COMBINING SVM AND KNN ALGO

Sushil Kumar Mishra[1], Pankaj Bhatt[2]

[1]*PG Student, Computer Science Engineering, Graphic Era Hill University, Uttarakhand, India*
[2]*PG Student, Computer Science Engineering, Graphic Era Hill University, Uttarakhand, India*

Abstract

With the growth of information technology. There emerges many intrusion detection problem such as cyber security. Intrusion detection system provides basic infrastructure to detect a number of attacks. This research work focuses on intrusion detection problem of network security. The main goal is to detect network behaviour as normal or abnormal. In this research work, two different machine learning algorithm have been combined together to reduce its weakness and takes positive feature of both algorithm. Its experimental results generates better result than other algorithm in terms of performance, accuracy and false positive rate. These combined algorithm has been applied on KDDCUP99 dataset to find better result by improving its performance, accuracy and reducing its false positive rate.

Keywords: *Intrusion detection system, KDDCUP99 dataset, False positive rate.*

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

1. INTRODUCTION

In this century, Information security is a most menacing problem. For handling these problem, many intrusion detection method has been introduced but no one is perfect. Intrusion detection system can provide protection for a computer network from malicious files such as virus, spyware and torjan horse. In which many computers are interconnected. An intrusion detection system can monitor the behaviour of all files those are coming in that computer network. If any file is suspicious or malicious. So Intrusion detection system can detect that malicious file or virus. Intrusion detection system has created many clustering based models separate normal and abnormal files. Intrusion detection system can be used for neural network also to provide security for computer network. Neural network first uses trained dataset to recognize normal as well as abnormal activity. Intrusion detection system protects a network traffics from malicious files. It basically maintains confidentiality and integrity of computer network. Any unauthorized access of any personal data can not be made possibled. So secrecy of network traffic and information can be well maintained. Intrusion detection system can only takes preventive majors to protect a computer network. No intrusion detection system (IDS) is perfect to protect a computer network. A very deep research work is going on intrusion detection system to develop a such system that can fully provide protection for a network traffic or a computer network. In this research work, support vector machine (SVM) basically creates clustering model. Which contains normal as well as abnormal data. Which can monitor normal as well as malicious behaviour to protect a computer network from any malicious attack such as virus , worms, torjan horse, rootkits attacks.

Intrusion detection system has been divided into two parts.

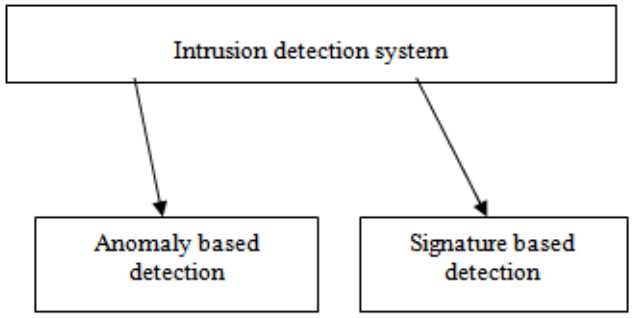

Fig. 1 Types of IDS

1.1 Anomaly Based Detection

Anomaly based intrusion detection system is based on a set of heuristic rule. Which basically monitors a normal as well as abnormal behaviour in a computer network. If any file is self replicating in nature or trying to damage any other file, such behaviours are detected by anomaly based detection. The main disadvantage of anomaly based detection system is higher false positive rate.

1.2 Signature Based Detection

Signature based intrusion detection system can detect only known computer virus in a computer network. The computer virus, those are discovered. Its signatures are created. These signatures are stored in database. If any file comes in a computer network. So its signatures are matched with all file. If file matches with virus signature so it is declared a computer virus otherwise a normal file. The main disadvantage of signature based intrusion detection system is that it can not detect a new computer virus.

2. EXPERIMENTAL PARAMETERS

There are many parameters such performance, accuracy and false positive rate, that can be calculated for intrusion detection system.

Performance : Performance deals with achieving a target in more efficient manner.

Performance = (True Positive)/(True Positive)+(True Negative)

Accuracy : Accuracy deals with achieving a goal more close to its actual value.

Accuracy = (True positive+True Negative)/(True Positive+True Negative+False positive+False Negative).

False positive rate : Falsely detect a normal file as abnormal file.

False positive rate =(False Positive)/(False Positive+True Negative).

3. EVALUATION DATA SOURCES

False positive rate was calculated by the standard data set KDDCUP99 given by the MIT laboratory. In this data set, there are different types of attacks. Those may categorize normal as well as abnormal data.

MIT Lincoln laboratory basically establishes a computer network. About 7 days, monitors network traffic. Which contains normal as well as abnormal data.

KDDCUP99 data set basically contains normal, denial of service, buffer overflow, guess_passwd(53) and probe attacks.

Denial of service : Denial of service (DOS) intrusion is an intrusion. In which , legitmate information can not be make available to legitmate receiver. DOS intrusion also slows down computer system.

User to Root(U2R) : In this type of attack, attacker accesses client's password in unauthorized manner and can access personal information or secret information from computer system by using stolen password.

Remote to User(R2U) : In this attack, attacker can transmit a packet over network. Which is not legitmate for that network. Which increases network traffic. Remote to user(R2U) can adversly affacts performance of that computer network and can slow down computer system or can restart a computer system again and again.

Probe : In this attack, attacker monitors all information. Which are being sent in that network and can access it.

4. COMBINING SVM AND KNN ALGORITHM

Support vector machine(SVM) is a supervised learning method for classification. In which, a hyperplane is created through which a normal as well as abnormal data is separated from each other. Support vector machine(SVM) basically contains two phases-

1- Training phase
2- Testing phase

1-Training phase : Support vector machine(SVM) is able to learn a huge set of pattern from dataset. In the dataset, there are various kind of homogeneous pattern and heterogeneous pattern of data . That can provide better classification between normal and abnormal data.

2-Testing phase : By using training phases, Testing can be done by support vector machine. Support vector machine can evaluate accuracy, performance etc.

Support vector machine can evaluate false positive rate but it generates very high false positive rate.

K nearest neighbor algorithm is basically a machine learning algorithm. Which can be used to solve traveling salesman problem.

By using K nearest neighbor algorithm, false positive rate can be evaluated but it gives higher false positive rate.

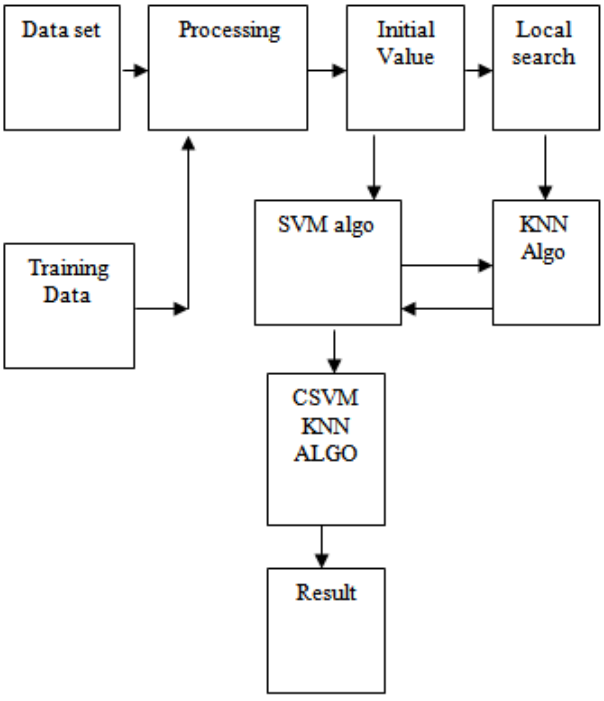

Fig. 2: Intrusion detection system using CSVMKNN

Support vector machine(SVM) basically uses support vectors to create a hyperplane. Hyperplane is used to separate normal and abnormal data. Knn algorithm is used to find new data added to training data set.

so here, Support vector machine(SVM) and K nearest neighbor (KNN) algorithms are combined together to evaluate false positive rate is known as **COMBINED SUPPORT VECTOR K NEAREST NEIGHBOR** (**CSVMKNN**) algorithm. CSVMKNN algorithm is a mixture of support vector machine (SVM) and K nearest neighbor (KNN) algorithm. These two algorithm works together in CSVMKNN algorithm. In which, support vector machine (SVM) uses training data set to learn something from data set. If any new is added to its dataset. so it is updated by K nearest neighbor (KNN) algorithm.

CSVMKNN algorithm can be used as support vector machine (SVM) and K nearest neighbor (KNN) algorithm to evaluate false positive rate or false alarm rate. False positive rate evaluated by using CSVMKNN algorithm, Can produce better result. CSVMKNN algorithm is applied on KDDCUP99 data set. This data set contains several type of attack such as buffer overflow, Denial of service (DOS) etc. CSVMKNN algorithm generates false positive rate. Which is better than Support vector machine (SVM) and K nearest neighbor (KNN) algorithm.

5. CSVMKNN ALGORITHM

Algorithm1 : SVM with KNN clustering
Input: Use training data set containing normal and abnormal data (Class type).
Output: Generate SVM classifier.
1 start
2 select data from different class;
3 Separate normal and abnormal data by SVM classifier;
4 While number of iteration to add data to data set
5 Use support vector to create hyperplane;
6 Hyperplane separate normal and abnormal data;
7 Apply KNN clustering
8 KNN clustering classified normal and abnormal cluster.
9 If new data added to data set
10 update dataset;
11 else
12 Continues it as it;
13 end.

After this algorithm, SVM learning process is applied on data set. Its main goal is to randomly choose data points from KDDCUP99 data set. Hyperplane is used to separate normal and abnormal data points. So there must be a separate hyperplane between each training data points. So it can provide a better selection method for each data points. Support vector machine (SVM) training phase should be introduced. In which. Hyperplane can allocate between each data points. KNN clustering phase is introduced to separate normal data and abnormal data. If new data is added to training data set. So by using K nearest neighbor (KNN) clustering phase, these new added data can be updated to training data set. So these strategy is carried out in next algorithm.

Algorithm2:
Input: Training data set (KDDCUP99).
Input: S1-Number of iteration.
Input: S2-Maximum detection rate.
Input: S3-Minimum detection rate.
Output: Support vector machine(SVM) and K nearest neighbor (KNN) Classifier.
1 **Start**
2 initialize the data;
3 Let S2 is maximum detection rate, initially zero;
4 Let S3 is minimum detection rate, initially Zero
5 While S3<S2
6 initialize i=0;
7 for i=1,……..,….S1
8 Training phase :
9 Support vector machine (SVM) training phase;
10 Clustering Phase :
11 K nearest neighbor (KNN) clustering phase;
12 end
13 Use Support vector machine(SVM) Classifier;
14 Use hyperplane to separate normal and abnormal data;
15 if new data is added to data set ;
15 Use Knn algorithm to update S2;
16 Update learning process;
17 else
18 continue it as it:
19 end
20 end

The KNN clustering phase is used for better selection strategy. False positive can be decreased by using CSVMKNN algorithm. If new added data is declared as normal. Otherwise, it increases its true positive rate. Which basically adversely affacts performance and accuracy. In SVM training phase, if new data is declared as abnormal but in KNN clustering phase, it is declared as normal. So such new data is declared a new kind of intrusion. In SVM training phase, if new data is added to training data set , declared as normal and in KNN clustering phase, it is again declared as normal. So such data decreases false positive rate or false alarm rate. It increases performance and accuracy of that machine learning algorithm.

Combined support vector machine k nearest neighbor (CSVMKNN) algorithm basically provides better selection strategy than support vector machine (SVM) and K nearest neighbor (KNN) algorithm. CSVMKNN algorithm takes positive features of support vector machine (SVM) algorithm and K nearest neighbor (KNN) algorithm and avoids weakness of Support vector machine (SVM) algorithm and K nearest neighbor (KNN) algorithm. CSVMKNN algorithm reduces false positive rate of its algorithm by using better selection strategy and improves performance of machine learning (CSVMKNN) algorithm. So, CVMKNN algorithm generates lesser false positive rate than support vector machine (SVM) algorithm and K nearest neighbor algorithm (KNN) algorithm. CSVMKNN algorithm can produce higher performance and accuracy than support vector machine (SVM) and K nearest neighbor (KNN) algorithm.

6. RESULTS

Support vector machine (SVM) algorithm, KNN nearest neighbor (KNN) algorithm and CSVMKNN algorithm are applied on training data set (KDDCUP99). Through which , false positive rate can be calculated. These false positive rate will be compared to determine. Which algorithm has generated lesser false positive rate

Support vector machine (SVM) classifier: SVM classifier **is** used to create a hyperplane between different data points by using support vector **.** These hyperplane is used to separate normal and abnormal data. On the basis of this, we can evaluate performance, accuracy, false positive rate.

Class	Normal	Denial Of service	User To Root	Remote To User	Probe
Normal	900	7	8	1	0
Denial Of service	3	345	0	2	11
User To Root	400	0	0	0	10
Remote To User	345	0	41	34	0
Probe	127	100	0	10	0

Fig-3 SVM classifier

K nearest neighbor (KNN) classifier is used to discover new data added to training data set. KNN classifier also determines that new added data is normal or abnormal. KNN algorithm is applied on KDDCUP99 data set to evaluate performance, accuracy and false positive rate.

Class	Normal	Denial Of service	User To Root	Remote To User	Probe
Normal	928	1	5	0	1
Denial Of service	0	45	0	200	1
User To Root	4	3	6	5	0
Remote To User	0	0	412	234	15
Probe	1	4	0	0	23

Fig-4 KNN classifier

CSVMKNN classifier basically contains feature of both algorithm support vector machine (SVM) and K nearest neighbor (KNN) algorithm. CSVMKNN algorithm is applied on KDDCUP99 dataset to generate its performance, accuracy, false positive rate.

Class	Normal	Denial Of service	User To Root	Remote To User	Probe
Normal	100	0	8	9	70
Denial Of service	30	35	0	0	89
User To Root	0	0	0	50	0
Remote To User	0	0	0	24	0
Probe	1	4	0	0	0

Fig-5 CSVMKNN Classifier

Evaluation Measure	SVM	KNN	CSVMKNN
False positive Rate	12.00	11.00	6.00
False Negative Rate	26.00	6.00	0.89
Performance	8.00	9.00	14.50
Accuracy	7.50	3.50	16.00

Fig-6 Comparison of false positive rate

CSVMKNN algorithm generates lesser false positive rate than Support vector machine (SVM) and K nearest neighbor (KNN) algorithm.

7. CONCLUSION

In this research work, Support vector machine (SVM) algorithm, K nearest neighbor (KNN) algorithm and CSVMKNN algorithm have been applied on KDDCUP99 data set separately. In which CSVMKNN algorithm has generated lower false positive rate than SVM and KNN algorithm. CSVMKNN algorithm has enhanced performance , accuracy and higher detection rate than other machine learning algorithm. Still , there is area of improvement in this algorithm until we are not getting zero false positive rate.

REFERENCES

[1]. pgale, Robert, Sheodoor schote, rengin and Christopher kruegel.”A Literature analysis on automated malware analysis technique”

[2]. Pargas, Rob Jonathan jarcy, Eleazar Aguirre Anaya , Samon Galeana Huerta and Alba Felix Moreno Hernandez,"Security controls for Android" In Computational Aspects of Social Networks (CASoN), 2012 Fourth International Conference on, pp.212-216,IEEE,2012

[3]. Blasing, Thomas, Leonid Batyuk, A-D.Schmidt, Seyit Ahmet Camtepe, and Sahin Albayrak." An android application sandbox system for suspicious software detection" In Malicious and Unwanted Software (MALWARE), 2010 5th International Conference on ,pp. 55-62 IEEE, 2010.

[4]. Johnson Ryan, Zhaohui Wang , Corey Gagnon and Angelos Stavrou." Analysis of Android Applications' Permissions. " In Software Security and Reliability Companion(SERE-C),2012 IEEE Sixth International Conference on, pp. 45 - 46.IEEE,2012.
[5]. Susan M. B. and Rayford B.V. (2000). Intrusion detection via fuzzy data mining, Proceedings of the 12th Annual Canadian Information Technology,Ottawa, Canada, June 19-23, 2000, PP.109-122.
[6]. A Detailed Analysis of the KDD CUP 99 Data Set, Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A.

BIOGRAPHIES

Sushil kumar Mishra is a M.tech student and doing research work in computer security

Pankaj Bhatt is pursuing M.tech and doing research work in computer security.