International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 5 Issue: 5 559 – 562

Supervised Intrusions Detection System Using KNN

Mr Utpal Shrivastava
Department of Computer Science Engineering, Amity University, Gurgaon ,Haryana, India
Email: ushrivastava@ggn.amity.edu, utpalshrivastava@gmail.com

Abstract— This paper is on implementations of intrusion detection system using Knn algorithm using R language. The dataset used is the KDDcup 1999 a well know bench mark for IDS. The machine learning algorithm K nearest neighbor(Knn) is use for the detection and classification for the known attacks.
The experimental results are obtained using R programming language.

Index Terms—*Machine learning, Knn, attacks, R programming.*

*****

I. INTRODUCTION

In the present world there is a major problem of security of data from attack. Researchers are working for detection and prevention of attacks as it is coming up with a threat to commercial business as well as for personal use. Intrusion detection system is for monitoring the incoming and outgoing data in the network for determining the possibility venerable attack on a system or network.
It also monitor the network traffic for suspicious activity and alert the network or system administrator about those attacks when occurred. We have used KDDcup dataset which is made available from MIT's Lincoln Lab; a benchmark datasets. It was developed for Intrusion Detection System evaluations by DARPA. During the experiment, we have examine all the 23 attack explicitly the attack of four types, denial of service, user to root, root to local and probe, distinguish with normal.
The rest of the paper is organized as follows. Section 2 presents the related works using corresponding machine learning Algorithms for proposed model. Section 3 introduce about the our proposed model for AIDS. Section 4 described the KDD 99 intrusion detection cup dataset. Using those machine learning algorithms in our proposed system, which presented in Section 2, Section 5 describes the experimental results obtained by using R tool. Section 6 for conclusion for this paper.

II. RELATED WORK

A IDDM (Intrusion Detection using Data Mining Techniques) [24] is a real-time NIDS for misuse and anomaly detection. It applied association rules, Meta rules, and characteristic rules. Jiong Zhang and Mohammad Zulkernine [21] employ random forests for intrusion detection system. Random forests algorithm is more accurate and efficient on large dataset like network traffic. We also use this data mining technique to select features and handle imbalanced intrusion problem. The most related work to ours is done also by them [19]. They use Random Forests Algorithm over rule-based NIDSs. Thus, novel attacks can be detected in this network intrusion detection system. In contrast to the previously proposed data mining based IDSs, we employ random forests for anomaly intrusion detection. Random forests algorithm is more accurate and efficient on large dataset like network traffic. We also use the data mining techniques to select features and handle imbalanced intrusion problem.[16] Random Forest (RDF) also intend to handle new instances that are not considered in all current supervised machine learning techniques[21], And k-Nearest Neighbor(k-NN) algorithm, is one of those algorithms that are very simple to understand but works incredibly well in practice. k-NN method was used as a supporter method for multi-class classification [22][25].

III. DATASETS DESCRIPTION

There are around 494020 records in the dataset having 41 features. The features are as follow:

Table 1: Feature KDDcup of dataset

1	duration	continuous.
2	protocol_type	symbolic.
3	service	symbolic.
4	flag	symbolic.
5	src_bytes	continuous.
6	dst_bytes	continuous.
7	land	symbolic.
8	wrong_fragment	continuous.
9	urgent	continuous.
10	hot	continuous.
11	num_failed_logins	continuous.
12	logged_in	symbolic.
13	num_compromised	continuous.
14	root_shell	continuous.
15	su_attempted	continuous.
16	num_root	continuous.
17	num_file_creations	continuous.
18	num_shells	continuous.
19	num_access_files	continuous.
20	num_outbound_cmds	continuous.
21	is_host_login	symbolic.
22	is_guest_login	symbolic.

 ISSN: 2321-8169
XXX – YYY

23	count	continuous.
24	srv_count	continuous.
25	serror_rate	continuous.
26	srv_serror_rate	continuous.
27	rerror_rate	continuous.
28	srv_rerror_rate	continuous.
29	same_srv_rate	continuous.
30	diff_srv_rate	continuous.
31	srv_diff_host_rate	continuous.
32	dst_host_count	continuous.
33	dst_host_srv_count	continuous.
34	dst_host_same_srv_rate	continuous.
35	dst_host_diff_srv_rate	continuous.
36	dst_host_same_src_port_rate	continuous.
37	dst_host_srv_diff_host_rate	continuous.
38	dst_host_serror_rate	continuous.
39	dst_host_srv_serror_rate	continuous.
40	dst_host_rerror_rate	continuous.
41	dst_host_srv_rerror_rate	continuous.

There are four types of attacks in the dataset a) Denial of Service Attack (DoS): is an attack in whichthe attacker makes some computing or memory resource toobusy or too full to handle legitimate requests, or denies legitimate users access to a machine.(b) User to Root Attack (U2R): is a class of exploit inwhich the attacker starts out with access to a normal user account on the system (perhaps gained by sniffing passwords, a dictionary attack, or social engineering) and is able to exploit some vulnerability to gain root access to the system. (c) Remote to Local Attack (R2L): occurs when an attacker who has the ability to send packets to a machine over a network but who does not have an account on that machine exploits some vulnerability to gain local access as a user of that machine. (d) Probing Attack: is an attempt to gather information about a network of computers for the apparent purpose of circumventing its security controls.

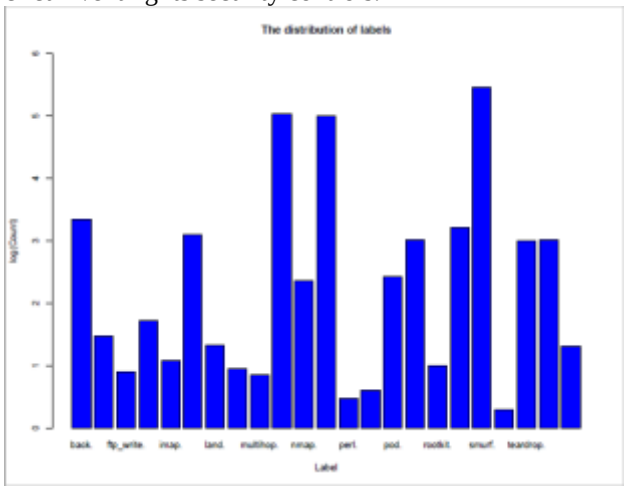

Figure 1: Distributions of labels

Figure 1 shows the distributions of different types of attack by label in the dataset.

.

Table 2: Label and count

	label	count
1	back.	2203
2	buffer_overflow.	30
3	ftp_write.	8
4	guess_passwd.	53
5	imap.	12
6	ipsweep.	1247
7	land.	21
8	loadmodule.	9
9	multihop.	7
10	neptune.	107201
11	nmap.	231
12	normal.	97277
13	perl.	3
14	phf.	4
15	pod.	264
16	portsweep.	1040
17	rootkit.	10
18	satan.	1589
19	smurf.	280790
20	spy.	2
21	teardrop.	979
22	warezclient.	1020
23	warezmaster.	20
23	warezmaster.	20

IV FEATURES SELECTION

The cleaning of data is required as there are many redundant data in the data set. There are some features which has less importance and that can be removed. Cleaning process also helps is reducing the size of the data base and hence saving the memory and processing time of data during training and testing. We have first removed the redundant records form the dataset and we got finally 145585 records out of 494020 records in the KDDcup. Then the data with the zero variance is removed from the dataset and remove the non numeric data from the dataset. we got 19 feature selected out of 41 features in dataset the selected feature are as follows:

```
$ logged_in
$ count
$ srv_count
$ serror_rate
$ srv_serror_rate
$ rerror_rate
$ srv_rerror_rate
$ same_srv_rate
```

```
$ diff_srv_rate
$ srv_diff_host_rate
$ dst_host_srv_count
$ dst_host_same_srv_rate
$ dst_host_diff_srv_rate
$ dst_host_same_src_port_rate
$ dst_host_srv_diff_host_rate
$ dst_host_serror_rate
$ dst_host_srv_serror_rate
$ dst_host_rerror_rate
$ dst_host_srv_rerror_rate
```

The above 19 feature is used to train the model and test the model.

IV. Machine Learning Algorithms

The 19 feature from extracted form the dataset is further normalized. Normalization is important as too much data variance of the data compare to other attribute can effect the data classification. We have used Knn algorithm for the classification. k-NN classification is an easy to understand and easy to implement classification technique[22]. Despite its simplicity, it can perform well in many situations. k-NN is particularly well suited for multi-modal classes as well as applications in which an object can have many class labels. For example, for the assignment of functions to genes based on expression profiles, some researchers found that k-NN outperformed SVM, which is a much more sophisticated classification scheme[2]. The 1-Nearest Neighbor(1NN) classifier is an
important pattern recognizing method based on representative points [23]. In the 1NN algorithm, whole train samples are taken as representative points and the distances from the test samples to each representative point are computed. The test samples have the same class label as the representative point nearest to them. The k-NN is an extension of 1NN, which determines the test samples through finding the k nearest neighbors. type of IDS, for example, network-based IDS will analyze
network related information such as packet destination IP address, logged in time of a user, type of protocol, duration of connection etc. It is not known which of these features are redundant or irrelevant for IDS and which ones are relevant or essential for IDS. There does not exist any model or function that captures the relationship between different features or between the different attacks and features. If such a model did exist, the intrusion detection process would be simple and straightforward. In this paper we use data mining techniques for feature selection. The subset of selected features is then used to detect intrusions.

V Implimentation and Results

We have used R for the implementation for model. 70 percent of the data is used for training the model and 30 percent of data is used for the testing the model. We have used caret package in R for implementing KNN algorithm.

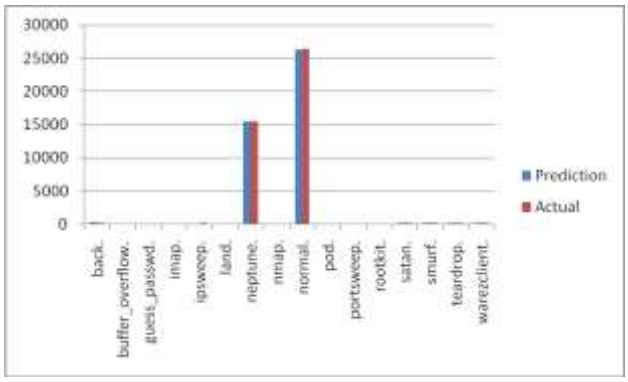

Figure 2: Bar chart for prediction and actual result

Accuracy	0.9933814
Kappa	0.9870008
AccuracyLower	0.9925757
AccuracyUpper	0.9941205
AccuracyNull	0.6034352
AccuracyPValue	0.0000000

Table 3: Result of KNN in R

V Conclusion

Recent researches employed decision trees, artificial neural networks and a probabilistic classifier and reported, in terms of detection and false alarm rates, but it was still high false positives and irrelevant alerts in detection of novel attacks. This paper has presented a implementation of the KNN algorithm in R data mining techniques that have been implemented for detections of venerable attacks in intrusion detection systems. And, we applied the classification methods for classifying the attacks (intrusions) on KDDcup dataset. The results showing the performance of the KNN is giving good result.

Acknowledgment

The authors wish to thank various authors and researchers whose research work has proved a great source of help for this paper. Authors also thank Google for bringing quality and relevant results.

References

[1] W. Lee and S. J. Stolfo, “Data Mining Approaches for Intrusion Detection”, the 7th USENIX Security Symposium, San Antonio, TX, January 1998.

[2] K.T.Khaing and T.T.Naing, “Enhanced Feature Ranking and Selection using Recurisive Featue Elemination and k-Nearest Neighbor Algorithms in SVM for IDS”, Internaiton Journal of Network and Mobile Technology(IJNMT), No.1, Vol 1. 2010.

[3] M. Bahrololum, E. Salahi and M. Khaleghi, "Anomaly Intrusion Detection Design using Hybrid of Unsupervised and Supervised Neural Network", International Journal of Computer Network & Communications(IJCNC), Vol.1, No.2, July 2009.

[4] L. Breiman, “Random Forests”, Machine Learning 45(1):5–32, 2001.

[5] V. Marinova-Boncheva, "A Short Survey of Intrusion Detection System" , 2007.

[6] Tamas Abraham, “IDDM: Intrusion Detection Using Data Mining Techniques”, DSTO Electronics and Surveillance Research Laboratory, Salisbury, Australia, May 2001.

[7] M. Mahoney and P. Chan, “An Analysis of the 1999 DARPA/Lincoln Laboratory Evaluation Data for Network Anomaly Detection”, Proceeding of Recent Advances in Intrusion Detection (RAID)-2003, Pittsburgh, USA, September 2003.

[8] KDD’99 datasets, The UCI KDD Archive, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html , Irvine, CA, USA, 1999.

[9] KDD Cup 1999. Available on: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, December 2009.

[10] Lan Guo, Yan Ma, Bojan Cukic, and Harshinder Singh, “Robust Prediction of Fault-Proneness by Random Forests”, Proceedings of the 15th International Symposium on Software Reliability Engineering (ISSRE'04), pp. 417-428, Brittany, France, November 2004.

[11] Ting-Fan Wu, Chih-Jen Lin, and Ruby C. Weng, “Probability Estimates for Multi-class Classification by Pairwise Coupling”, The Journal of Machine Learning Research, Volume 5, December 2004.

[12] Yimin Wu, High-dimensional Pattern Analysis in Multimedia Information Retrieval and Bioinformatics, Doctoral Thesis, State University of New York, January 2004.

[13] Bogdan E. Popescu, and Jerome H. Friedman, Ensemble Learning for Prediction, Doctoral Thesis, Stanford University, January 2004.

[14] Eleazar Eskin, Andrew Arnold, Michael Prerau, Leonid Portnoy, and Salvatore Stolfo. “A Geometric Framework for Unsupervised Anomaly Detection: Detecting Intrusions in Unlabeled Data.” Applications of Data Mining in Computer Security, 2002.

[15] WEKA software, Machine Learning, http://www.cs.waikato.ac.nz/ml/weka/, The University of Waikato, Hamilton, New Zealand.

[16] Leo Breiman and Adele Cutler, Random forests, http://statwww.berkeley.edu/users/breiman/RandomForests/cc_home.htm, University of California, Berkeley, CA, USA.

[17] David J. Hand, Heikki Mannila, and Padhraic Smyth, Principles of Data Mining, The MIT Press, August, 2001.

[18] MIT Lincoln Laboratory, DARPA Intrusion Detection Evaluation, http://www.ll.mit.edu/IST/ideval/,MA, USA.

[19] J.Zhange and M. Zulkerline, “Network Intrusion Detection using Random Forests”,2011.

[20] T. Lappas and K. Pelechrinis Data Mining Techniques for (Network) Intrusion Detection Systems”.

[21] J. Zhang and M. Zulkernine, ”Anomaly Based Network Intrusion Detection with Unsupervised Outlier Detection”, Symposium on Network Security and Information Assurance Proc. of the IEEE International Conference on Communications (ICC), 6 pages, Istanbul, Turkey, June 2006.

[22] S. Thirumuruganathan , “A Detailed Introduction to K-Nearest Neighbor (KNN) Algorithm”, World Press, May 17, 2010.

[23] X Wu, V Kumar, J Ross Quinlan, J Ghosh, “Top 10 Data mining Algorithm”, Knowledge and Information Systems, Volume 14, Issue 1, pp 1-37 ,2008 – Springer

[24] S. Mukkamala, A.H. Hung and A. Abraham, “Intrusion Detection Using an Ensemble of Intelligent Paradigms.” Journal of Network and Computer Applications, Vol. 28(2005), 167-182.

[25] S. Chebrolu, , A. Abraham, and J.P. Thomas, “Feature Deduction and Ensemble Design of Intrusion Detection Systems.” International Journal of Computers and Security, Vol 24, Issue 4,(June 2005), 295- 307

[26] A.H. Sung and S. Mukkamala, “The Feature Selection and Intrusion Detection Problems.” Proceedings of Advances in Computer Science - ASIAN 2004: Higher- Level Decision Making. 9th Asian Computing Science Conference. Vol. 321(2004) , 468-482.

[27] S. Mukkamala, A.H. Sung and A. Abraham, “Modeling Intrusion Detection Systems Using Linear Genetic Programming Approach.” LNCS 3029, Springer Hiedelberg, 2004, pp. 633-642.

[28] A. Abraham and R. Jain, “Soft Computing Models for Network Intrusion Detection Systems.” Soft Computing in Knowledge Discovery: Methods and Applications, Springer Chap 16, 2004, 20pp.

[29] A. Abraham, C. Grosan, and C.M. Vide, “Evolutionar Design of Intrusion Detection Programs.” InternationalJournal of Network Security, Vol. 4, No. 3, 2007, pp. 328-339.