Application of k-Nearest Neighbour Classification Method for Intrusion Detection in Network Data

S.Venkata Lakshmi
Department of Computer Science
Loyola College, Chennai

T.Edwin Prabakaran, Ph.D
Department of Statistics
Loyola College, Chennai

ABSTRACT

In today's world, increasing number of networks connected to the Internet poses a great challenge on security issues. Many defensive mechanisms exist and one such higher level mechanism is network intrusion detection system. Intrusion detection system is a process of intelligently monitoring the events in an individual system or network, analysing them for signs of violation of security policy. Two major classifications of intrusion detection systems are misuse and anomaly intrusion detection systems. Misuse detection system refers to detection of intrusions that follow well defined intrusion patterns. Anomaly detection model refers to detection performed by detecting changes in the behaviour of the system. Many data mining techniques like k-Nearest Neighbour (kNN), Association Rule Mining etc., have been applied to intrusion detection. This paper aims at application of kNN to a subset of records from the KDD Cup 1999 dataset for classification of connection records into normal or attacked data. The paper also applies kNN to the subset of records with the selected features proposed by **Kok-Chin-Khor et al [5]** to compare the classifications.

General Terms

Intrusion Detection, Feature Selection, k-Nearest Neighbour

Keywords

Attacks, connection records, normal, testing dataset, training dataset

1. INTRODUCTION

Intrusion is defined as any set of action that can compromise the integrity, confidentiality and availability of system resources[1][8].

There are two types of intrusion detection namely, misuse detection and anomaly detection. Misuse detection refers to the identification of well known intrusion patterns in data. Known attack patterns are easily identified using misuse detection models. Anomaly detection refers to unknown intrusion patterns in data. This could be used to identify known and unknown attacks [2],[9],[10].

Abundant data is collected from networks and it poses a great challenge to identify the intrusion in it. The raw data should be summarized into some records which comprises of lot of features [3],[6]. Selecting the relevant features from the given records actually simplifies the classification of data as well as improves its efficiency.

In this paper, a subset of records from the intrusion dataset is taken for application of kNN. Also the proposed feature sets [5],[7] of the network data is taken and then the kNN is applied for effective and efficient classification of connection

records into normal or attack. There would be also a comparison of the results as follows:

- The classification of data obtained when the entire features are used by the k-NN
- The classification of data obtained when only proposed selective features are used by the k-NN.

The paper is organized as follows: Section 2 gives the introduction to the intrusion dataset used. Section 3 gives a brief explanation of kNN and Section 4 focuses on feature selection and the five different feature sets proposed by [5]. Section 5 represents the application of kNN on the subset of records taken into consideration with entire features and application of kNN on the same subset of records with the different feature sets and the experimental results are plotted. Section 6 presents the conclusion.

2. INTRUSION DATASET

The **KDD Cup dataset [4]** is used for intrusion detection problems and is also used in this paper. This dataset is considered to be the benchmark data in Intrusion detection. The dataset was a collection of simulated raw TCP dump data over a period of nine weeks on a local area network. The known attack types are those present in the training dataset while the novel attacks are the additional attacks which are not present in the training dataset [1]. There are various attacks like Buffer overflow, Perl, Portsweep, Neptune, Smurf, Teardrop, Guess password, IPSweep etc., The training dataset consists of 4,94,021 records. The testing dataset consists of 3,11,029 records. In each connection record there are 41 attributes describing different features of the connection.

3. k-NEAREST NEIGHBOUR

The k-Nearest Neighbour (k-NN) is a method for classification of objects based on the closest training examples in the feature space. k-NN is a type of instance based learning or lazy learning. The k-NN is one of the simplest of machine learning algorithms. An object is classified by a majority vote of its neighbours, with the object being assigned to the class most common amongst its k nearest neighbours (k is a positive integer, typically small). If k=1, then the object is simply assigned to the class of its nearest neighbour[3].

3.1 Algorithm for kNN :

```
Input: KDD99 Testing dataset U, KDD99 Training dataset V
for each connection u from U
    for each connection v from V
        calculate dist=d(u,v)
    next
    sort v according to dist
    select top k connections
    classify u as the majority of classes of the selected
    connections
next
```

4. FEATURE SELECTION

Since the network data is voluminous, it becomes difficult for analysis and extra features can make detection of intrusions very hard and difficult. There could also be complex relationships within the features and there could be redundant features representing the same information. Hence, it becomes necessary to reduce the amount of data for effective intrusion detection. Feature selection improves classification by searching for the best subset of features which classifies the data in an improved way.

The dataset comprises of 41 attributes for every connection record. The feature sets proposed by [5] are mentioned below:

Table 1. The features of the five feature sets

Feature	Selected features	No. of features
FS1	service, dst_bytes, logged_in, root_shell, count, srv_diff_host_rate, dst_host_count, dst_host_srv_diff_host_rate	8
FS2	service, src_bytes, dst_bytes, logged_in, count, dst_host_srv_count, dst_host_diff_srv_rate, dst_host_error_rate	8
FS3	service, dst_bytes, logged_in, root_shell, count, srv_diff_host_rate, dst_host_count, dst_host_srv_diff_host_rate, src_bytes, dst_host_srv_count, dst_host_diff_srv_rate, dst_host_error_rate	12
FS4	service, dst_bytes, logged_in, count	4
FS5	service, dst_bytes, logged_in, count, dst_host_count, root_shell, dst_host_error_rate	7

5. RESULTS

The following plot represents the application of kNN algorithm on the subset of records with the entire set of attributes of the connection records. The training dataset comprises of 220 records and the testing dataset consists of 270 records.

Table 2. Entire set of attributes: 41

NOR-MAL	BUF-FER	PERL	PORT-SWEEP	NEP-TUNE	SMU-RF
191	4	1	34	14	13

TEAR-DROP	GUESS-PWD	IP-SWEEP	ERROR
2	3	3	5

Fig 1: Attacks with entire set of attributes

The following plots represent the application of kNN on the same subset of records but with different feature sets.

Table 3. Feature Set 1 Comprising of 8 Attributes:

NOR-MAL	BUF-FER	PERL	PORT-SWEEP	NEP-TUNE	SMU-RF
102	1	0	59	13	71

TEAR-DROP	GUESS-PWD	IPSWEEP	ERROR
3	5	7	9

Fig 2: Attacks with Feature Set 1 attributes

Table 4. Feature Set 2 Comprising of 8 Attributes:

NOR-MAL	BUF-FER	PERL	PORT-SWEEP	NEP-TUNE	SMU-RF
236	3	0	4	3	1

TEAR-DROP	GUESS-PWD	IPSWEEP	ERROR
1	4	10	8

Fig 3: Attacks with Feature Set 2 attributes

Table 5. Feature Set 3 Comprising of 12 Attributes:

NOR-MAL	BUF-FER	PERL	PORT-SWEEP	NEP-TUNE	SMU-RF
224	3	0	10	20	2

TEAR-DROP	GUESS-PWD	IPSWEEP	ERROR
0	3	1	7

Fig 4: Attacks with Feature Set 3 attributes

Table 6. Feature Set 4 Comprising of 4 Attributes:

NOR-MAL	BUF-FER	PERL	PORT-SWEEP	NEP--TUNE	SMU-RF
112	2	0	122	16	4

TEAR-DROP	GUESS-PWD	IPSWEEP	ERROR
0	2	10	2

Fig 5: Attacks with Feature Set 4 attributes

Table 7. Feature Set 5 Comprising of 7 Attributes:

NOR-MAL	BUF-FER	PERL	PORT-SWEEP	NEP-TUNE	SMU-RF
81	0	0	49	26	83

TEAR-DROP	GUESS-PWD	IPSWEEP	ERROR
12	5	7	7

Fig 6: Attacks with Feature Set 5 attributes

6. CONCLUSION

In this paper the performance of k-Nearest Neighbour on the subset of intrusion detection dataset is presented with whole set of attributes as well as on the dataset with 5 different feature subsets of attributes. The classification done with the whole set of attributes identified portsweep attack records effectively. With the feature set 1 attributes, Smurf and Portsweep attacks were found out. With FS2 and FS3 attributes, identification of attacks were not efficient. Using FS4, the identification of portsweep attacks were effective. It has been identified that the feature set (FS5) consisting of 7 attributes yields the best results in terms of identifying the maximum number of attacks when compared to the whole set of attributes as well as the other feature sets (FS1, FS2, FS3 and FS4).

The application of k-NN with the feature set 5 identified the number of three different attacks namely, Smurf, Portsweep, Neptune as 83,49 and 26 respectively out of the 270 testing dataset records. The feature set 5 (FS5) also has fewer number of attributes and hence the required computational cost in processing the network data could also be reduced. In future, the feature set which could identify the maximum number of attacks in classification of network data shall be analysed and found out. This kNN technique could also be applied for different networks and real time data.

7. REFERENCES

- [1] Adebayo O. Adetunmbi*, Samuel O. Falaki, Olumide S. Adewale and Boniface K. Alese, "Network Intrusion Detection based on Rough Set and k-Nearest Neighbour", *International Journal of Computing and ICT Research*, Vol. 2, No.1, pp.60-66.,2008, <http://www.ijcir.org/volume1number2/article7.pdf>.
- [2] Ganapathy.S, Jaishankar.N, Yogesh.P and Kannan.A, "An Intelligent Intrusion Detection system using Outlier Detection and Multiclass SVM, *Int. J. on Recent Trends in Engineering & Technology*, Vol.05,No.01,166-169, 2011.
- [3] Hossein M.Shirazi, "Anomaly Intrusion Detection using Information Theory,k-NN and KMC Algorithms", *Australian Journal of Basic and Applied Sciences*,3(3):2581-2597,2009.
- [4]KDD CUP 1999 DATASET:
<http://kdd.ics.uci.edu/databases/kddcup99/>
- [5] Kok-Chin Khor, Choo-Yee Ting and Somnuk-Phon Amnuaisuk, "From Feature Selection to Building of Bayesian Classifiers: A Network Intrusion Detection Perspective", *American Journal of Applied Sciences* 6(11),1948-1959,2009.
- [6] Lee.W, Stolfo.S.J, Mok.K.W, "Algorithms for Mining System Audit Data", of *Proc.KDD*,1999.
- [7] Neveen I.Ghali , "Feature Selection for Effective Anomaly Based Intrusion Detection", *International Journal of Computer Science and Network Security*, Vol.9 No.3 March 2009.
- [8] Norbik Bashah, Idris Bharanidharan Shanmugam, Abdul Manan Ahmed, "Hybrid Intelligent Intrusion Detection System", *World Academy of Science, Engineering and Technology*, pp 23-26,2005.
- [9] Stolfo.S.J, Hershkop.S, Wang.K, Nimeskern.O, and Hu.C.W, "Behaviour Profiling of Email", *First NSF/NIJ,ISI*,2003.
- [10] Victor-Valeriu Patriciu, Liviu Rusu, Iustin Priescu, "Data mining approaches for Intrusion Detection in Email system Internet-Based",*Military Technical Academy*, 144-147,2003.