Journal of Data Mining and Knowledge Discovery
ISSN: 2229-6662 & ISSN: 2229-6670, Volume 3, Issue 3, 2012, pp.-88-90.
Available online at http://www.bioinfo.in/contents.php?id=42

ON THE KDD'99 DATASET: STATISTICAL ANALYSIS FOR FEATURE SELECTION

TAISIR ELDOS*, MOHAMMAD KHUBEB SIDDIQUI AND AWS KANAN

College of Computer Engineering and Sciences, Salman bin Abdulaziz University, Saudi Arabia.
*Corresponding Author: Email- eldos@eldos.net

Received: July 15, 2012; Accepted: July 30, 2012

Abstract- We present a contribution to the network intrusion detection process using Adaptive Resonance Theory (ART1), a type of Artificial Neural Networks (ANN) with binary input unsupervised training. In this phase, we present a feature selection using data mining techniques, towards two dimensional dataset reduction that is efficient for the initial and on-going training. The well know KDD'99 Intrusion Detection Dataset (KDD'99 dataset for short) is tremendously huge and has been reported by many researchers to have unjustified redundancy, this makes adaptive learning process very time consuming and possibly infeasible. We intend to reduce the dataset both vertically and horizontally, numbers of vectors and number of features, such that nearly 10% of the training subset is used for the initial unsupervised training process and nearly 1% of the training subset is used for the on-going training, and that is only regarding the number of vectors. On top of that, only the significant features will be used yielding a highly reduced dataset, and this is the scope of this work.

Key words- Network Intrusion, Data Mining, Neural Networks, Feature Selection.

Citation: Taisir Eldos, Mohammad Khubeb Siddiqui and Aws Kanan (2012) On The KDD'99 Dataset: Statistical Analysis for Feature Selection. Journal of Data Mining and Knowledge Discovery, ISSN: 2229-6662 & ISSN: 2229-6670, Volume 3, Issue 3, pp.-88-90.

Introduction

The term data mining refers to the process of extracting useful information from large databases to find unsuspected relationship and to summarize the data in novel ways that are both understandable and useful to data owner. It typically deals with the data that have already been collected for some useful purpose other than data mining analysis.

Intrusion detection is defined to be the process of monitoring the events occurring in a computer system and detect computer attacks and misuse, and to alert the proper individuals upon detection [1]. In this paper, we use the Oracle Data Miner (ODM) for the purpose of statistical analysis and feature selection on the KDD'99 dataset.

The KDD'99 dataset was used for The Third International Knowledge Discovery and Data Mining Tools Competition, which was held in conjunction with KDD'99 dataset, the fifth International Conference on Knowledge Discovery and Data Mining [2].The competition task was to build a network intrusion detector. This database was acquired from the 1998 DARPA intrusion detection evaluation program. An environment was set up to acquire raw TCP/IP dump data for a local-area network (LAN) simulating a typical U.S. Air Force LAN, which was operated as if it was a true environment, but blasted with multiple attacks. There are totally 4,898,431 connections recorded, of which 3,925,650 are attacks. For each TCP/IP connection, 41 various quantitative and qualitative features were extracted.

The simulated attack fall in one of the following four categories[2]:

Denial of Service Attack (DOS): In this category the attacker makes some computing or memory resources too busy or too full to handle legitimate request, or deny legitimate users access to machine.

Users to Root Attack (U2R): In this category the attacker starts out with access to a normal user account on the system and is able to exploit some vulnerability to obtain root access to the system.

Remote to Local Attack: In this category the attacker sends packets to machine over a network but who does not have an account on that machine and exploits some vulnerability to gain local access as a user of that machine.

Probing Attack: In this category the attacker attempts to gather information about network of computers for the apparent purpose of circumventing its security.

Table 1- Attacks Categories

Attack Category	List of Attacks
DOS	'neptune', 'back', 'smurf', 'pod', 'land', 'teardrop'
U2R	'buffer_overflow', 'loadmodule', 'rootkit', 'perl'
R2L	'warezclient', ' multihop', ' ftp_write', 'spy' 'imap', 'guess_passwd', 'warezmaster', 'phf'
PROBE	'portsweep', 'satan', 'nmap', 'ipsweep'

Related Work

Our literature survey reveals many results; in [3], the authors proposed a real-time intrusion detection system based on the Self-Organizing Map (SOM); an unsupervised learning technique that is appropriate for anomaly detection in wireless sensor networks. The proposed system was tested using KDD'99 dataset. The system groups similar connections together based on correlations between features. A connection may be classified as normal or attack. Attacks are classified again based on the type of attack. It took the system 0.5 seconds to decide whether a given input represents a normal behavior or an attack.

In [4], the authors addressed the main drawback of detecting intrusions by means of anomaly (outliers) detection, which is the high rate of false alarms when a behavior that has never been seen before is presented. In their work, they added a new feature to the unknown behaviors before they are considered as attacks, and they claim that the proposed system guarantees a very low ratio of false alarms, making unsupervised clustering for intrusion detection more effective, realistic and feasible.

In [5], the authors introduced an intrusion detection system based on Adaptive Resonance Theory (ART) and Rough Set theory. The ART was used to create raw clusters that were refined using Rough Set. As a preprocessing stage, symbolic-valued attributes of the dataset were mapped to numerical values. The proposed system was able to detect not only known attacks, but also new unknown attacks.

In [6], the authors conducted a statistical analysis of this dataset a KDD'99 dataset, the most common dataset widely used to evaluate intrusion detection systems, and found some issues that would result in poor systems evaluation. A new dataset (NSL-KDD) have been proposed. This dataset consists of selected records from the original dataset to overcome those shortcomings.

Our Approach

Intrusion detection systems can be signature based or anomaly based. Signature based systems are built on known connections features and are capable of detecting only known intrusions, while the anomaly type employ methods that are heuristic based and hence capable of detecting unknown intrusions as well. Anomaly based solutions require training systems on large datasets representing connections of various types. Practically, those systems have false predictions; false positives and false negatives that vary from system to system [7].

We intend to design an ART1 ANN for this purpose, and to minimize the miss prediction rate, we offer a continuous training approach. This approach requires a relatively small dataset which contradicts the main idea of having larger dataset for better training.

In this paper, we introduce the first phase of our project, which targets the horizontal dimension of the dataset; the number of attributes in the KDD'99 dataset, and hence the vector length, while the second stage will focus on the vertical dimension; to reduce the number of vectors.

The KDD'99 dataset has 41 attributes for each record. Some of these attributes are irrelevant and redundant [9]. Irrelevant attributes simply add noise to the dataset and affect the accuracy of proposed models using it. Another important point that has to be considered is the computation cost. Using a dataset with a large number of attributes results in a lengthy training and detection processes, and hence degrading the performance of an intrusion detection system.

We use the “Attribute Importance” mining function available in Oracle Data Mining that ranks the attributes in a dataset based on their significance using the Minimum Description Length (MDL) algorithm [8].

Experimental Results

Oracle Data Miner supports supervised learning techniques (classification, regression, and prediction problems), unsupervised learning technique (clustering technique and feature selection problem), and attribute importance technique. The availability of these algorithms provides all the necessary tools required in gathering information from a dataset. The main advantage of using Oracle Data Miner is that all data mining processing occurs within the oracle database [8].

Attribute Significance

Fig. 1 shows the result of applying the Attribute Importance function to the KDD'99 dataset. The tool ranks the attributes based on their significance, with the attribute of rank 1 being the most important attribute and all attributes having an importance less than or equal to zero have the same rank and considered as noise [8].

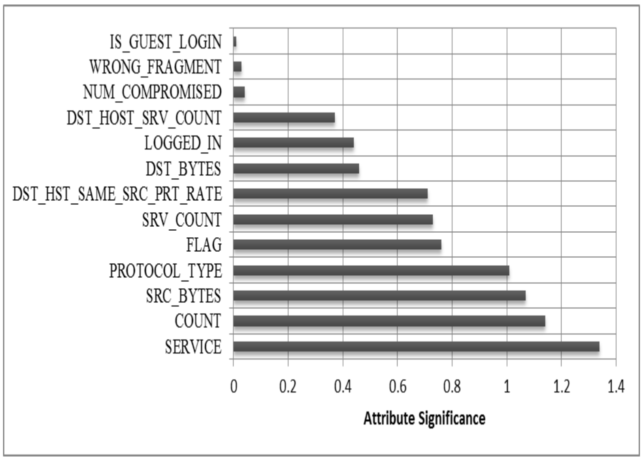

Fig. 1- Attribute Significance

It is clear from this figure that 13 attributes out of the 41 attributes of the KDD'99 dataset have an importance value above zero, and the rest have an importance of zero and hence not shown in the plot. We will use these attributes in the ART1 ANN learning process. We expect this to be more accurate having only 7 features, while keeping the total size low through the variable code length.

Statistical Analysis

ART1 neural network [10] is an unsupervised learning model that is used to recognize binary patterns. In this section, we provide a statistical analysis for the selected attributes and rely on these statistics to assign binary codes to these attributes. Table 2 shows the minimum value, maximum value, number of distinct values, and the number of bits assigned for each attribute.

Table 2- Significance Based Code Length, Total Number of Bits = 106

Attribute Name	Min	Max	Number of Distinct Values	Number of bits
SRC_BYTES	0	693375640	3300	30
DST_BYTES	0	5155468	10725	23
SRV_COUNT	0	511	470	9
COUNT	0	511	490	9
DST_HOST_SRV_COUNT	0	255	256	8
DST_HST_SAME_SRC_PRT_RATE	0	1	2	1
LOGGED_IN	0	1	2	1
WRONG_FRAGMENT	0	3	3	2
NUM_COMPROMISED	0	884	23	10
IS_GUEST_LOGIN	0	1	2	1
PROTOCOL_TYPE	-	-	3	2
FLAG	-	-	11	4
SERVICE	-	-	66	7

The proposed code length for representation is based on the minimum number of bits to express the largest value of each attribute, resulting in 106 bits per vector. However, we expect shorter code can be used by considering the number of distinct values of each attribute; which allow using 16 bits (instead of 30) for SCR_BYTES, 14 bits (instead of 23) for DST_BYTES, and 5 bits (instead of 10) for NUM_COMROMISED, making the total 78 bits. This is to be tested in the second stage.

Conclusion

Data mining application to the original KDD'99 dataset lead to reducing the original 41 attributes, with several hundreds of bits to represent each vector, to a hundred or even less bits per vector. This process dramatically reduces the space and time requirement of the implementation and improves the offline and online performance, in particular when continuous adaptive training is to be adopted to improve the detection accuracy. This reduction is consistent with an early work where only 7 attributes are used, without jeopardizing the richness of content, which could be the result of eliminating too many attributes.

Future Work

The resulting dataset will be used to train an ART1 ANN using the optimized attributes and reduced data set, and the performance in terms prediction accuracy and raining time will be benchmarked against other methods. Also, the architecture capability to improve its performance over time will be monitored.

Acknowledgment

This research was funded by a grant from the Deanship of Scientific Research in Salman bin Abdulaziz University in the Kingdom of Saudi Arabia.

References

[1] Srinoy S., Kurutach W., Chimphlee W., Chimphilee S. (2005) *World Academy of Science, Engineering and Technology*, 9, 140-144.

[2] KDD'99 dataset, University of California, Irvine (1999) *http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html*.

[3] Hayoung Oh, Inshil Doh and Kijoon Chae (2009) *International Journal of Computer Science and Applications*, 6(3), 20-32.

[4] Goverdhan Singh, Florent Masseglia, C´eline Fiot, Alice Marascu and Pascal Poncelet (2009) *The 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining* (*PAKDD CUP*), Thailand.

[5] Kanok Prothives and Surat Srinoy (2009) *Proceedings of the International Multi Conference of Engineers and Computer Scientists*, 1.

[6] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu and Ali Ghorbani (2009) *The 2nd IEEE Symposium on Computational Intelligence Conference for Security and Defense Applications* (*CISDA*).

[7] Lazarevic A., Kumar V. and Srivastava J. (2005) *Managing cyber threats: issues, approaches, and challenges*, 330.

[8] *http://download.oracle.com/docs/html/B10698_01/toc.htm*.

[9] Neveen I. Ghali (2009) *IJCSNS International Journal of Computer Science and Network Security*, 9(3).

[10] Carpenter G.A. and Grossberg S. (2003) *The Handbook of Brain Theory and Neural Networks*, Cambridge, MA, 87-90.