Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 92 (2016) 188 – 198

Procedia Computer Science

2nd International Conference on Intelligent Computing, Communication & Convergence

(ICCC-2016)

Srikanta Patnaik, Editor in Chief

Conference Organized by Interscience Institute of Management and Technology

Bhubaneswar, Odisha, India

Feature analysis, evaluation and comparisons of classification algorithms based on noisy intrusion dataset

Jamal Hussain[a], Samuel Lalmuanawma[a*]

[a]*Department of Mathematics & Computer Science, Mizoram University, Aizawl, India, 796004*

Abstract

Various studies have been carried on an Intrusion Detection System (IDS) environment bycomparingthe performance of various Machine Learning (ML)based on a refined intrusion dataset with an error-free environment. However, the real-world network data deals with a large amount of noisy information on transmission, and the IDS have to work in such an environment frequently. Dealing with such noisy data is, therefore, a challenging issue in an IDS environment for detecting threads from network activities. In this paper, various Data Mining (DM) and ML algorithms are evaluated and compared by normal and noisy dataset prepared from KDD'99 and NSL-KDD dataset (10%-20% Noise). The empirical results demonstrate that NN (SOM) is far better compared to other tested algorithms regarding robustness tonoisy environment; however,JRip and J48 from the tree family outperform others regarding overall performance matrices. Feature dependency on datasets for a specific classifier is analyzedby Performance-based Method of Ranking (PMR). The evaluation results statistically proved that each classifier has a unique combination of a feature subset to results optimal performance. Empirical results demonstrate that evaluations of IDS based on NSL-KDD give more realistic results compared to theKDD'99 original dataset.

[a*] Corresponding Author.Tel:+919436353048
E-mail address: samuellalmuanawma@mzu.edu.in

doi:10.1016/j.procs.2016.07.345

Keywords: Intrusion Detection System, Machine Learning, anomaly detection, false alarm, ignored attack;

1. Introduction

The advancement of Information Technology (IT) raised numerous security breaches. Therefore, to secure valuable resources over the public network, it is essential to implement an Intrusion Detection System (IDS). IDS aimed to sort out various intrusive attempts on the computer network system based on the three important pillars of information security, i.e., confidentiality, integrity and availability of a resources[1]. It first gathers and analyze information from various sources within the computer network, triggers alarm to system administrators and blocks unauthorized access if an attack attempt is encountered.

Various recent studies in IDS are evaluated based ona refined intrusion dataset with an error-free environment. However, the real network information deals with a huge amount of noisy data, and the IDS have to work in such an environment repeatedly. Therefore, this paperinvestigates and evaluates on various data mining algorithms to study the performance of each classifier against various datasets,i.e., noise-free and noisy (10% & 20%) environment. We choose top-six classifier from various tested ML algorithms based on evaluating performance. Ranking of significance feature based on performance is done for each selected classifier to study and compare with various feature selection method used in recent research.

2. Theory and Algorithms

2.1. *Dataset organization:* In this studies, four types of datasets prepared from KDD'99 [2] and NSL-KDD[3] intrusion dataset are used to evaluate each classification algorithms. Details of the data preprocess are as follows:

2.1.1. KDD'99 Cup Dataset

This dataset is built and prepared by Stolfoet al.[4]based on the data captured in DARPA'98 Intrusion Detection System Evaluation program[5].The datasets contain a TCP-dump raw data of about 5 million connections collected from 7 weeks of network traffic records of training sets and 2 weeks records of test set data having around 2 million network traffic records. For each TCP/IP connection, 41 quantitative and qualitative features were extracted. For evaluation, the author used 10% of the original data. After folding the data onto 13 stratified folds, the first folds containing 39461 instances were used for final evaluation.

2.1.2. NSL-KDD Dataset

Tavallaeeetal.[3] proposed the NSL-KDD datasets thatare an enhanced edition of KDD'99 datasets. The KDD'99 dataset contains large records of redundant data, where 78% training dataset and 75% test dataset are duplicate which may direct classifier algorithm unreasonable towards the further repeated records. Redundant data found on the test dataset can also harm the evaluation performance into a higher degree of detection accuracy. The refined dataset in KDDtrain+.txt and KDDtest+.txt are combined, all the attack traffic in a dataset is grouped into one class named as an anomaly. The ratio of normal and anomaly instances is maintained to meet the preprocess requirement. After folding the data onto six (6) stratified folds, the first fold containing 27526 instances is used for evaluation.

2.1.3. Noisy Dataset (10% & 20%)

Since this studies focus on evaluating the robustness of various data mining algorithms in anoisy environment, the author used the NSL-KDD dataset for noise generation and added noisy data varying percentages of 10% and 20% to specific attributes using the KDD features. Noiseis added to the specific features after analyzing the dependency on thefeature using NSL-KDD. To evaluate and analyzed feature significance, GainRatio[6]and Info Gain[7], based on