Evaluation of Ensemble Learning algorithms for Intrusion Detection Systems

Mantek Singh (IIT2016007)
Gagan Ganapathy(IIT2016038)
Saurabh Mishra(IIT2016045)
Ridam Arora(IIT2016134)
Niharika Shrivastava(IIT2016501)

Ensemble Learning

- Ensemble learning is a machine learning paradigm where multiple learners are trained to solve the same problem. In contrast to ordinary machine learning approaches which try to learn one hypothesis from training data, ensemble methods try to construct a set of hypotheses and combine them to use.
- The generalization ability of an ensemble is usually much stronger than that of base learners.
- There are several reasons why ensemble learning can enhance the performance of a model.
- The first reason is that, the training data might not provide sufficient information for choosing a single best learner.
- The second reason is that, the hypothesis space being searched might not contain the true target function, while ensembles can give some good approximation
- Literature Review: https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/springerEBR09.pdf

Motivation

1. A Denial of Service (DoS) attack is a malicious attempt to affect the availability of a targeted system, such as a website or application, to legitimate end users.
2. In case of a Distributed Denial of Service (DDoS) attack, and the attacker uses multiple compromised or controlled sources to generate the attack.
3. Most common at the Network (layer 3), Transport (Layer 4), Presentation (Layer 6) and Application (Layer 7) Layers.
4. Hard solutions(Reduce Attack Surface Area, firewalls), easier solution : Accept only normal traffic

Literature Review

1. Evaluation of Machine Learning Algorithms for Intrusion Detection System - **Cornell University arXiv**
2. Research on Intrusion Detection Model using ensemble learning methods - **2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)**
3. An Ensemble Approach for Intrusion Detection System Using Machine Learning Algorithms - **2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence)**

Steps to perform

1. Retrieve the KDD set
2. Data cleaning (Using dummy variables, escape dummy variable trap, feature scaling, normalization, feature selection)
3. Implement Naive Bayes, Decision Trees, SVM, Random Forest etc.
4. Implement ensemble algorithms (Bootstrapping, AdaBoost etc.) with different bases.
5. Compare results with different metrics

Evaluation and results

1. Some models give high accuracy in a long time. Evaluate model prediction time and trade-offs
2. Evaluate false negatives, true positives
3. Understand which ensemble combination works best

Dataset features

Categories of Attack	Attack name	Number of instances
DOS	SMURF	2807886
	NEPTUNE	1072017
	Back	2203
	POD	264
	Teardrop	979
U2R	Buffer overflow	30
	Load Module	9
	PERL	3
	Rootkit	10
R2L	FTP Write	8
	Guess Passwd	53
	IMAP	12
	MulitHop	7
	PHF	4
	SPY	2
	Warez client	1020
	Warez Master	20
PROBE	IPSWEEP	12481
	NMAP	2316
	PORTSWEEP	10413
	SATAN	15892
normal		972781

Attributes	Type
Total duration of connections in second	continuous
Total number of bytes from sender to receiver.	continuous
Total number of bytes from receiver to sender	continuous
Total number of wrong fragments	continuous
Total number of urgent packets	continuous
Protocol type	discrete
Type of service	discrete
The status of the connection (normal or error)	discrete
Label (1) if the connection established from to the same host. Otherwise label (0)	discrete