Intrusion Detection Systems using ensemble learning

Mentor - Prof. O.P. Vyas

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

What is 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.[4]
- Incorporate dynamic updates,such as selective removal or addition of classifiers, and diversity of models[5]
- The generalization ability of an ensemble is usually much stronger than that of base learners.[6]

How and why it started

Problems with existing learning algorithms :-

- Statistical
- Computational
- Representational[7]

Strong learners are difficult to obtain. Weak learners is relatively easy to develop and can effectively be boosted into a strong learner as long as they are trained and combined strategically.[5]

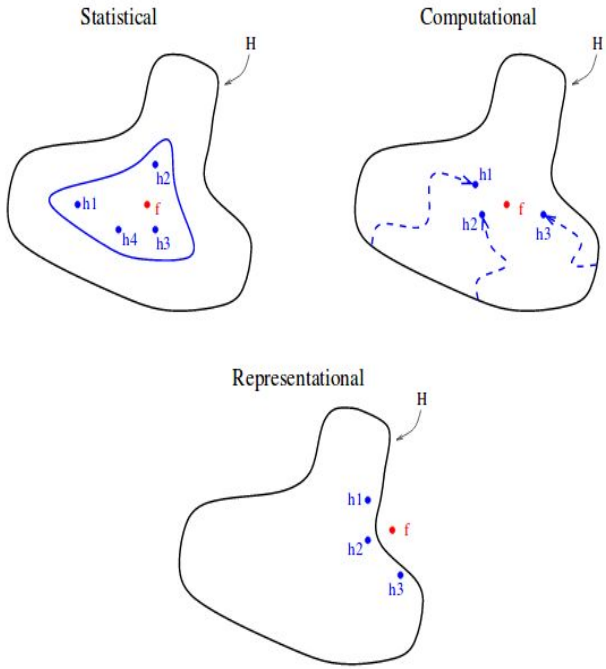


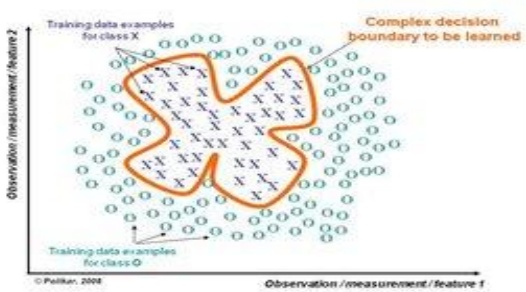
Fig. 2. Three fundamental reasons why an ensemble may work better than a single classifier

History of Ensemble

- Foundation:- Dasarathy 1979 Paper - partitioning data set and combining results[8]
- Hansen 1990 paper- Combining similarly configured networks
- Scapire 1990 - Boosting technique introduction
- Late 1990s and early 2000s - Classifier selection and fusion
- 2009 Netflix Competition leading to commercial fame

Benefits of Ensemble Learning

1. **Model Selection:** Weighted average of diverse models [5]
2. **Too much/little data**: Large datasets partitioned and separate classifiers made and ensembled together for ease
3. **Data Fusion**: Collect data from multiple resources
4. **Time complexity**
5. **Divide and conquer:**



Intrusion Detection System

- An intrusion detection system (IDS) inspects all inbound and outbound network activity and identifies suspicious patterns that may indicate a network or system attack from someone attempting to break into or compromise a system. [1]
- Focus : Using ML algorithms to classify attacks on KDD dataset and use ensemble to get better accuracy to predict an attack based on some attributes.

Types of Attacks

- DoS(Denial of service) Attack :- makes a computing resource too busy to serve legitimate networking requests and hence denying users access to a machine[2]
- R2L (Remote to user) Attack :- sending packets to a machine to which it doesn't have access in order to expose it's vulnerabilities [1]
- U2R (User to Root) Attack :- hacker tries to get the access rights from a normal host in order, for instance, to gain the root access to the system.
- Probing :- scanning a machine to exploit it later.

Features of dataset

- **Training attack types:** back dos, buffer_overflow u2r, ftp_write r2l, guess_passwd r2l, imap r2l, ipsweep probe, land dos, loadmodule u2r, multihop r2l, neptune dos, nmap probe, perl u2r, phf r2l, pod dos, portsweep, probe, rootkit u2r, satan probe, smurf dos, spy r2l, teardrop dos, warezclient r2l, warezmaster r2l.
- There are 23 types of attacks that can be broadly classified into 4 primary types - DOS,
- **41 features:** protocol_type, service, flag, src_bytes, dst_bytes, land, … etc.
- The test data is not from the same probability distribution as the training data, and it includes specific attack types not in the training data.
- Most novel attacks are variants of known attacks and the "signature" of known attacks can be sufficient to catch novel variants.

Models to be used

Naive Bayes

- It is termed as ‘Naive’ because it assumes independence between every pair of feature in the data.
- In case of multiclass classification problems, Guassian NB gives good results.
- Precision = 98%

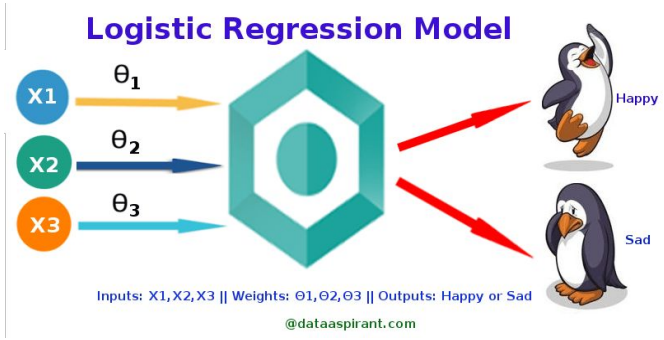
Multilayer Perceptron

- MLP has shown that they are capable of approximating an XOR operator as well as many other non-linear functions.
- MLPs with one hidden layer are capable of approximating any continuous function.
- In case of multiclass classification problems, softmax function is used on output layer.

Models to be used

Logistic Regression

- Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary).
- In case of multiclassification problems, a one vs all approach can be used to obtain the optimal results.



Decision Trees

- Different in the way it generates it's decision boundaries. Decision Trees bisect the space into smaller and smaller regions, whereas Logistic Regression fits a single line to divide the space exactly into two.

Ensemble Models

Random Forests

- **Bagging**: Here, after selecting multiple samples of data vua bootstrapping, we train multiple hypthesis(models). Once each model has developed a hypothesis. The models use *voting for classification* or *averaging for regression*. This is where the “Aggregating” in “Bootstrap Aggregating” comes into play. Each hypothesis has the same weight as all the others.
- **Boosting**: Boosting refers to a group of algorithms that utilize weighted averages to make weak learners into stronger learners. Unlike bagging that had each model run independently and then aggregate the outputs at the end without preference to any model.
- Since decision trees are susceptible to problems of overfitting and high variance, random forests is a way of using the above techniques on decision trees to create an ensemble model that can be used to predict outcomes with a much higher accuracy.

Timeline

Searching

KDD dataset was found to be the most suited towards our domain of research with 41 features.

Model selection

We study about the different classifiers and choose the ones that best suit domain of our research.

Ensembling

Once tested on individual models, we select a few of them and create an ensemble model which gives the highest accuracy.

Dataset Cleaning

Clean the data by converting categorical data into numerical., fill missing values,

Implementing

We try out our cleaned data on several models such as decision trees / naive bayes classifier etc.

In the end

Submit a final report eliciting the results obtained.

References

1. https://arxiv.org/pdf/1801.02330.pdf Evaluation of Machine Learning Algorithms for Intrusion Detection System

2. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7883100 Research on Intrusion Detection Model Using Ensemble learning Methods

3. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8442693 An Ensemble Approach for Intrusion Detection System Using Machine Learning Algorithms

4. ACM Proceedings, June 21 - 23, 2000 Springer-Verlag London, UK MCS '00 Proceedings of the First International Workshop on Multiple Classifier Systems

References

5. ACM Computing Surveys 50 A Survey on Ensemble Learning for Data Stream Classification

6. National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 2 Ensemble learning Methods

7. Oregon State University, Oregon, USA Ensemble Methods in Machine Learning

4. Ensemble Learning History Ensemble learning methodologies