Machine Learning in Network Intrusion Detection

Liru Long / Xilei Wang / Xiaoxi Zhu
CS229 Machine Learning

Data Pre-processing

Data Set
- Data set is downloaded from KDD Cup '99 - Data Mining and Knowledge Discovery competition organized by ACM Special Interest Group on Knowledge Discovery and Data Mining
- Raw training-test data originates from TCP dump data of real network traffic, processed to generate seven million connection records.
- The subset used in this project includes a training set of 494021 samples and a test set of 110283 samples.
- Each sample contains 41 features grouped into three categories: basic features, content features, and time-based features.
- Each sample is labeled with integer 0-4 - 0 indicates normal traffic and 1-4 corresponds to a specific class of network attack as follows: 1—Probing, 2—DoS, 3—U2R, 4—R2L.

Feature Pre-processing

Feature values in the original data set vary significantly (continuous vs discrete, order of magnitude, etc.). As many Machine Learning algorithms are sensitive to variations in feature values, sample features are normalized using following approaches:
1. Test features (e.g., protocol, service type) are mapped onto discrete integer values between 0 to 10.
2. Large numerical values are normalized to the range [0, 1].
3. Extremely large numerical values (e.g., source bytes) are converted to range [1, 10] using logarithm base 10.

Naïve Bayes Classifier

Multi-class Classification
- A multinomial multi-class Naïve Bayes Classifier is trained with training data set labeled with 0-4.
- Test data is classified into 0-4 with trained NB classifier.
- Incorrect labeling across different classes of attacks (e.g. label class-1 attack as class-2 attack) is ignored.

Binary Classification
- A binary Naïve Bayes classifier is trained with training data labeled as true vs false, (where label 0 mapped to true, label 1 mapped to false).
- Test data is then classified as normal (true) vs attack (false).

Decision Tree & Random Forest

Decision Tree
- By using the largest information gain, Decision Tree would build a tree based on the most distinguishable feature.
- Since decision tree suffered from overfitting, different max depths are tried. It turns out that depth of 9 gives the most satisfactory result.
- Decision Tree could successfully identify the normal network traffic with 98.4% recall, but only with 72.9% precision.

Random Forest
- Random Forest builds multiple decision trees using subsets of the training samples to improve classification or regression results.
- Since max depth of 9 in decision tree method, decision trees in random forest all use max depth of 9 to build.
- Number of decision trees are varied when build the random forest; it turns out that with number of 16, we get the most satisfactory result.
- Random forest gets every class improved, but for the attack and normal classification, it does not improve too much.
- The random forest tends to classify the data with unseen pattern to normal class.

K-means Clustering

K-means Clustering
- Unsupervised learning - suitable to deal with data with unknown characteristics.
- Time Complexity: O(NKID)

Approach
- Normalize the data so that values of each feature range in [0, 1].
- Select the k of cluster k and generate random initial centroids for each cluster.
- The centroids are trained with training set.
- Label each cluster by identifying the majority class of samples in this cluster.

Protocol-based Classification

Protocol-based Classification
- Training data & test data are further divided into three subsets based on transport layer protocol - TCP, UDP and ICMP.
- Three multi-class NB classifiers are trained for each subset and test data is classified using corresponding classifier.

Feature Reduction

Motivation
- The data occupies extremely high-dimensional feature space.
- Redundant features are identified through inspection based on prior networking knowledge (e.g. SYN error rate vs REJ error rate).
- Observation from single classifier results proves the assumption such that:
  - Features are correlated as overfitting has been detected - best performing DT classifier uses tree depth of 8 (against 41 features)
  - Certain features are more closely correlated with single class traffic - protocol-based NB classifier generates improved performance.

Approach
- A pair-wise feature correlation matrix is generated.
- Pairs of features with high correlation (p-value & mag) are selected.
- Identify groups of highly correlated features (each pair of features within the group are highly correlated).
- Apply PCA feature reduction to each group of highly correlated features.

Results
- The following group of features are identified as highly correlated:
  - features 13 & 16
  - features 2 & 23 & 24 & 36
  - features 27 & 28 & 40 & 41
- Both training data & test data are reduced to 30-dimensional after applying PCA.

Reflection & Future Work

Results
- Consistent performance across different Machine Learning algorithms.
- Decision Tree and Random forests outperforms the other algorithms with higher recall.
- Precision in identifying normal traffic is ~79%.
- Recall of normal traffic classification varies across different approaches.
- PCA shows marginal improvement in classification results.

Proposed extension - cascaded DT-GMM classifier - does not lead to expected significant performance improvement.

Reflections
- Dimension of data set features is in the medium to high range, which increases the complexity of learning.
- As features are non-uniform - test vs numerical, range variation etc., results may be susceptible to data preprocessing.
- Features are correlated - an effective feature reduction technique is critical to successful training.
- Training data set and test data set do not have same sample distribution across multiple classes.
- Sample size of class 3 (U2R attack) is too small.
- Class-4 (R2L attack) samples in training data set and test data set do not have similar feature signature.

Cascade Classifier

Cascaded DT-GMM
- The misclassification from RLD class to Normal class brings down the precision of normal class precision in all classifier models. The reason of this is lack of training samples.
- The Decision Tree Model gives the best result when classifying, using the Decision Tree Model as the first layer model.
- GMM model-models the distribution of the data and does not suffer from initialization problem of K-Means.
- Unsupervised learning is preferred when dealing with unseen data.
- Building GMM to identify abnormal network traffic, after going through the decision tree model.
- The cascade improves the application’s ability to identify the attack but impairs its ability to classify normal traffic.

Comparison of Classifiers

Future work
- A more effective data pre-processing technique may be developed, preserving unique feature characteristics.
- Improved feature reduction technique - generate feature correlation for individual class to help identify signature patterns.
- Finer classification of traffic types, especially for attack classes.
- Explore unsupervised learning algorithm for effective abnormal detection as the network attack keeps evolving.

References