**INSIDER THREAT DETECTION**

**What are INSIDER THREATS?**

Insider Threats are a form of cybersecurity challenges that are caused by insider employees/working officials inside an organization.

**Why is it a challenge?**

Surveys tell that around 90% of the cybersecurity challenges are due to insider threats so they need to be tackled. Insiders posess unauthorized access and information that they may exploit.

On the technical front,

Existing approaches are inadequate in accuracy and precision. They suffer from a high false positive rate.

The reasons behind their failed accuracy are:

1. Heterogeneous nature of available data
2. The existing approaches tend to be overfitting on data.

**AVAILABLE DATA**:

Based on the available data, we have approaches varied for different types of data:

1. **Sequential Data**: This includes the **time-series data**.It uses both ML and DL techniques.
2. **Non-Sequential Data**: It comprises of the **image data** to identify insider threats. They are possible using Neural Networks.

What are the algorithms based on?

**ANOMALY DETECTION TASK.**

In this, we shall go through the various ANOMALY DETECTION TASKS.It can be done through ML and DL both.

Let’s begin with DL Techniques:

1. Sequential Data Approaches: LSTM , Hybrid LSTM and CNN models are used for that. For more information on LSTMs , click [here](VANILLA.docx).

(2)Non Sequential Data Approaches:

Different researchers proposed various techniques such as using RNN to extract features, generate images from features and use CNN to identify images as malicious/non-malicious. Some used Graph Convolution Network, Deep Feed Forward CNN. They were observed to have rendered the best results. Among the others , they include Geometric Transformations, OCSVM.

But with CNN, one problem was observed. Due to its pooling operations, it loses user behavioral data important for detection.

To rectify this problem, WCNN were used. WCNN are **Wavelet Convolution Networks.** They are better because they have infinitely strong and generalized convolution and pooling layer prior to losing important user information before detection. It uses both spectral and spatial analysis for the same.

**Spectral features** include the information related to the frequency domain from an image. They comprise of the information about different color channels and frequency components.

**Spatial Features** involve information like edges,corners and texture patterns of the image, spatial distribution or arrangement of pixels in the image. Convolution Neural Networks are good at capturing them through the use of convolution layers.

**WCNN** consists of **wavelets, convolution layers and pooling layers**. CNNs use spatial features while the spectral analysis uses invariant features. Thus, both spectral and spatial features are the art of a single model. They can filter and downsample image data in the frequency domain. Connect Convolutional and Pooling layers for MULTI RESOLUTION ANALYSIS.

Pooling Layer is added immediately after convolution layers and average pooling provides better texture features extraction.

**Data Collection:**

Data is collected from CMU CERTv4.2 Dataset (for WCNN). CERTv 5.2 dataset is used for Machine Learning Algorithms.

Data is stored in unified formats. The two main sources are user activities , organization structure and user profile information.

**Data Preprocessing:**

It involves aggregation of data from various log files into a single user file. The dataset contains multiple csv files. This if followed by feature extraction and feature selection. **Frequency based and statistical features** are retrieved from the user logs to improve threat detection.

We perform data processing in steps to obtain three levels of data granularity namely user-session, user-week and user-day.

Frequency features : Number of emails sent, number of long on after hour, number of USB connects to PC.

Statistical Features: Email size, file size, number of words in website visit.

After extracting the features, they are converted into a 1D feature vector.

Min-Max Normalization: It is carried out to smoothen the data since it ranges across varying magnitudes.

Then the features are generated for image based classification. The feature vector is generated image representation. Matlab’s mat2grey feature converts feature vector into greyscale image.

**Imbalance Handling**: An imbalanced Dataset is one having uneven distribution of quantity of samples within distinct classifications. One class has significantly more instances than another. Due to this, model tends to perform poorly on the minority class because it tends to be biased towards the majority class.

For handling imbalances in the dataset, we use algorithms such as SMOTEENN(Synthetic Minority Over Sampling Technique and Edited Nearest Neighbours). SMOTE is for oversampling the minority classes(it increases the number of samples) ad ENN is used to remove miss-classified instances from all the classes.

Other techniques for Imbalance handling are based on cost-sensitive learning algorithms. In Machine Learning, they include:

1. Class Weights: It adjust weights to optimize the output. Higher weights are assigned to minority class instances so while backpropagation, it is all rectified.
2. Ensemble Methods: Clustering based Subset ensemble learning technique. Clusters majority class samples into a fixed number of groups, subsequently reduce the number of majority class samles. Prediction from an ensemble of classifiers decision trees, Naive Bayes and SVM.
3. Algorithm Level Adjustments
4. Cost Matrices: Explicitly define the misclassification cost for different classes.

Modified SMOTE is integrated with SMOTE is employed to improve the classification accuracy by fusing outputs from several weak classifiers.

**Insider Threat Classification using WCNN**

WCNN is based on VGG-19 Network. The WCNN takes image representations as input, passes them from a sequence of computational units that produce representations with a relevant score for classifications.

WCNN Architecture:

**Based on VGG-19 Network.**

Uses convolution layers with 3x3 kernels and extended stride to minimize the size of the feature maps instead of pooling layer.The global average Pooling Layer is put before fully linked layers. It serves as an Energy layer as it pools up the averages of the activated output of each feature map. It also helps in avoiding overfitting. They reduce model complexity by downsampling the feature map to a single value equal to the average for each channel.

In a nutshell, the training model consists of generalized convolution and pooling layers. There are four decomposition levels(as wavelets for Multi-resolution analysis)

**Evaluation and performance metrics:**

WCNN gave 97.19% accuracy.

Performance Metrics:

1. **Accuracy=** (True positives+ True negatives)/(True positives+ True -ves + False+ves + False-ves): Number of correctly classified instances/ Number of total instances.
2. **Precision**: Instances originally malicious/ Total instances labelled as malicious.
3. **Recall**: Ability to recognize all the malicious instances.
4. **F1 Score:** Harmonic mean of precision and recall.
5. **AUC-ROC Curve**: Visualisation Metric for ensuring the effectiveness of the proposed approach(WCNN).Higher value implies better classification results.

**ANALYZING DATA GRANULARITY and DETECTING INSIDER THREATS USING MACHINE LEARNING :**

MODEL TRAINING:

1. **Logistic Regression:**

Uses a logistic function to model a binary dependent variable based on independent variables. It returns the probability of an input vector belonging to a class. The aim of logistic regression with l2 regularization is to find the coefficient vector that minimizes the sum of squared errors between logistic function and labels.

1. **Neural Networks:**

Here, it will be a multi layer perceptron with upto three hidden layers. Back propagation with Adam formulation of stochastic gradient descent is assumed here. This also implies that each weight has its own learning rate.Weight vectors of the neurons are updated via backpropagation algorithm over a number of epochs using chain rule.

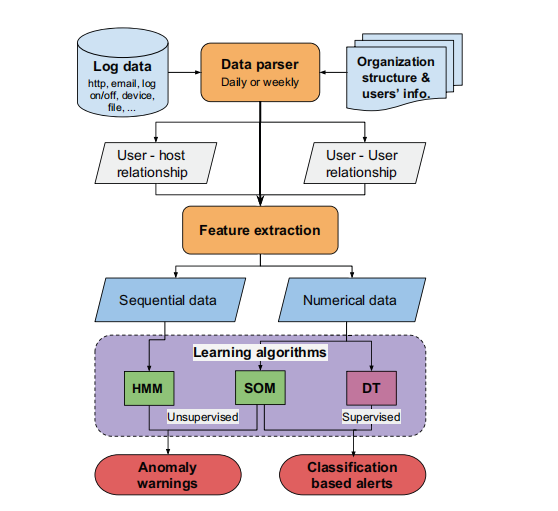
1. **Random Forest:**

Decision tree learns tree-like nonlinear classification model,where each leave of the tree represents a ‘decision’.

Process for building an ensemble of decision tree classifiers that collectively vote to provide a single class label for each exemplar. They combine the output of multiple decision trees to predict the final output. Number of decision trees and their depth are the hyperparameters here. Random Forest trains individual decision trees through a random subset of input features and training data. These are called as **bootstrap aggregating** methods. Individual trees are trained using **CART Algorithm** which seeks to gain maximum information at the end of each split of the tree.

1. **XGBoost:**

eXtreme Gradient Boosting used for classification in this case.It is an ensemble learning method wherein boosting involves training a series of decision trees and rectifying the misclassified instances along each iteration. He objective function contains loss function and regularization term.



OPTIMIZATION OF THESE MACHINE LEARNING ALGORITHMS TO FETCH BETTER RESULTS:

1. **Logistic Regression**: lbfgs solver is used to speed up training by parallelization. It finds the optimal parameters that minimise the logistic regression loss function.
2. **Random forest**: We use random search with Cross Validation to adjust the following hyperparameters:
3. the number of individual decision trees
4. The number of features available for training individual decision trees
5. The number of maximum leaf nodes in a tree.
6. **Artificial Neural Networks:**We use ADAM optimization algorithm for training upto 250 epochs and then aain applied Random Search for hyperparameter tuning.

**RESULTS ON TEST DATA:**

Logistic Regression usually achieves high Recall-malicious ,low precision. This results in more time investment of the analyst in false alarms.

RF and ANN show better precision . The show an overall better balance in maintaining between false positive rates and high malicious insider detection rates.

**RF shows better results** than ANN and LR. It has higher precision and a lower false positive rate .

DATASET COMPARISON:

On **CERT r5.1**, the models perform **well with low false alarm rates** and **75% malicious insider detection rate**, where only the scenario 2 insider is missed.

On **CERT r6.2,** due to a different organization structure, **accuracy is lower**. On **user-week data**, most of the **insider cases are missed**, while on user-session data, although malicious insider detection rates are acceptable, **the precision is low**. On the other hand, accuracy could be increased on CERT r6.2 by using **RF or ANN with user-day data.**

Results by Data Granularity Levels:

On Data Granularity , user session data appears to give the best results. This is because it allows a system with higher malicious insider detection rate and faster response time.

RF can be employed where manpower for investigating the alarms is limited, as it gives high precision.

On the other hand, ANN gives higher malicious insider detection-rate.

**USING UNSUPERVISED ENSEMBLES:**

**This uses CERTr4.2 Dataset.**

**It has various data representations like percentile,concatenation mean/median differences.**

1. **AUTOENCODERS:**

Type of Artificial Neural Network used for unsupervised learning and dimensionality reduction. Encoder takes the input data and compresses its representation. This compressed representation is referred to as ‘bottleneck’ or ‘encoding’ layer. The bottleneck layer has lower dimensionality than the input layer. The decoder reconstructs the original data from it. The autoencoder is trained to minimize loss function between input data and reconstructed output.

1. **ISOLATION FOREST:**

An ensemble of isolation trees. They are an ensemble based learning algorithm for anomaly detection. The idea is to isolate the anomalies in a dataset by building a model that separates normal data points from rare, potential anomalous ones.

The isolation trees are trained using recursive partitioning of data. Each tree in IF works on a subset of training data and feature set. Binary splits are generated in each node of a tree by random feature selection and splitting.

The anomaly score for a data point is averaged across all trees.

1. **Light Weight on-line Detector of Anomalies:**

It is an ensemble method combining weak histogram based anomaly detectors into a strong detector. Each histogram anomaly detector in LODA works on a subset of input features .Individual histograms are calculated for each of the vectors.

To produce anomaly score for a data sample, LODA uses the average of logarithm of probabilities estimated on individual projection vectors.

1. **Local Outlier Factor(LOF):**

It is an unsupervised anomaly detection algorithm to identify anomalies/outliers in a dataset.It focuses on the local density of data points to detect anomalies,makig it effective in identifying outliers in regions of varying data density. It also uses k-nearest neighbours for the same. Calculates the local density of a data point by comparing its distance with nearest neighbors.

The LOF Score for a data point is a measure of the difference in its local density between the data point and its neighbors. Data points with significantly lower local density than their neighbours are considered to be potential anomalies.

**Performance Metrics:**

Here, the performance is measured based on ROC-AUC Curve. ROC(Receiving Characteristic Curve) depicts the relationship rate between detection rate and false positive rate under different decision thresholds .

RESULTS:

**Autoencoders exhibit the best performance** in detecting anomalies representing insider threats, especially at very low FPRs. LOF performs well on lower data counts.

BEST DATA REPRESENTATION: **PERCENTILE REPRESENTATION. COFIRMED BY FRIEDMAN TEST**.

**OTHER ALGORITHMS:**

(1)**SELF-ORGANIZING MAP:**

An unsupervised, competitive-learning based neural network based on how the human neural system works. The SOM produces a non

linear, ordered, low dimensional projection of data from multi-dimensional input space. The SOM consists of nodes that can act as decoders or detectors of their respective input space domains post training.

1. **HIDDEN MARKOV MODEL**:Statistical model in which states are hidden. A HMM is trained to model each user’s action sequence over a given time, in this work: weekly. Then for each of the user’s new action sequence, the user’s HMM is used to calculate the log probability of the sequence. The sequence is flagged as an anomaly for further analysis if the log probability value is larger than a threshold. If the

action sequence is not flagged, or flag is cleared by an analyst, it is used in combination with the previous action sequence to train the user’s HMM again.

1. **DECISION TREES:** Discussed above.

RESULTS:

In **Hidden Markov Model**, due to time limitation, only weekly user actions are used for training. HMM gives the best balance between training time and the detection performance.

**SOM** visually shows that the insider threat data exhibits different characteristics than Normal data.

**CONCLUSION:**

We will use **CERTr4.2 dataset.** The data consists of user log information. The data preprocessing is done. Firstly the imbalanced data is handled using **SMOTE** , other cost-sensitive learning algorithms. This enables better accuracy. Then feature extraction takes place. For **image based approaches**, where **WCNN is used**, 1D feature vector is converted into image feature vector. This is done using mat2grey feature present in MATLAB. For sequential data, **LSTMs are used along with CNNs**. Talking about ML Algorithms, among **supervised ones**, **Random Forests** give the best results. Others include Logistic Regression. Among **unsupervised ensembles**, **autoencoders** give the best results. Others include Isolated Forest, Light weighted on-line anomaly detection, Local outlier Factor. The anomaly scores are calculated from each algorithm and the **highest Anomaly score** represents the best classifier.The data representation is tested using Friedman tests and determines percentile as the best data representation for unsupervised learning ensembles. The best **data granularity is user-session**.

SOMs are used for visualisation purposes. They are unsupervised Neural Networks.