국립서울과학기술대학교 SEOULTECH

A Survey on Data-driven Network Intrusion Detection

서울과학기술대학교 컴퓨터공학과 진호천

Depart. Computer Science, Seoul National University of Science and Technology

SeoulTech UCS Lab

CONTENTS

Methods for Feature-based Challenges: Noise

- ✓ Some traffic data in datasets may contain outliers. To combat noisy data or data with outliers, **feature normalization methods** have been applied to scale features and allow them to have similar effects in the model so noise would not weigh differently than the rest of the data.

- ✓ Feature Normalization: Feature normalization methods can be applied to scale features and allow them to have similar effects in the model so noise will not be weighed differently than the rest of the data.
  - ➢ Delahoz et al. [93] to overcome the noise challenge in network data, they normalized continuous variables to have mean 0 and variance 1, a standard normal distribution. They are encoded before normalization via $\frac{x-\bar{x}}{\sigma}$. The classification code is 1 if the feature is "activated", 0 otherwise.
  - ➢ Hsu et al. [73] developed an online intrusion detection system based on an autoencoder, SVM, and Random Forest ensemble where noise was dealt with feature normalization.

$$\bar{a} = \frac{log(a+1)}{(log(a+1))_{max}}, \quad \sigma \uparrow \quad (1)$$

$$\bar{a} = \frac{a}{a_{max}}. \quad \sigma \downarrow \quad (2)$$

Methods for Feature-based Challenges: Noise

- ✓ Density-based clustering: Density-based clustering is used to group data from the same class together and to identify outliers that are unusually distant from the observed cluster.
  - ➢ Due to the decentralized nature of DoS attacks in wireless sensor Networks (WSNs), Shamshirband et al. proposed an Imperial Competitive Algorithm (ICA) based on density algorithm and fuzzy logic. Dense areas in data space are clusters, and low-density areas (noise) surround them. Density-based clustering can detect shape clusters and deal with noise. Because network intrusion detection involves outlier detection, the density-based method can be extended to outlier detection.

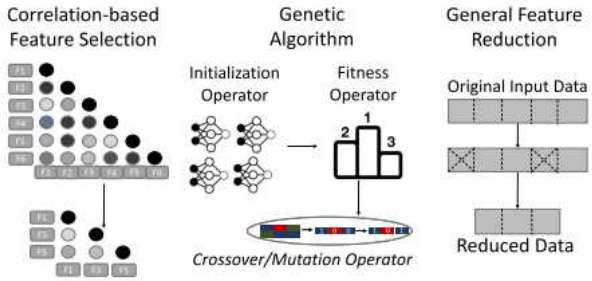
Methods for Feature-based Challenges: Redundant

- ✓ Some features in the collateral intrusion feature set may not contribute significantly to the predictive power of the model, so they may be removed based on their importance.

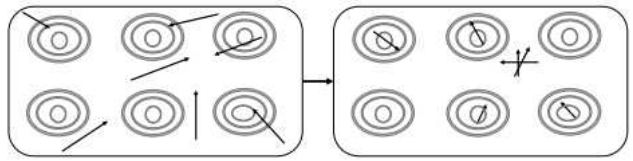
- ✓ Feature removal frameworks: Data redundancy is a common problem in network intrusion datasets, so researchers have developed frameworks that recommend the use of specific data deletion techniques
  - ➢ Ganapathy et al. [58] detailed introduces the characteristics of a gradual removal method and the modified method of mutual information, the method choice to maximize the output characteristics of the information (maximization of the correlation between the input and output), conditional random field (CRF) as a hierarchical method (each layer represents a type of attack), as well as the genetic feature selection, A set of trees is generated and the best feature set is extracted.
  - ➢ Bamakan et al. [18] proposed an effective intrusion detection framework by embedding feature selection into its target function combined with time-varying chaotic particle swarm optimization (TVCPSO) algorithm. They simplified the weighted objective function approach in the flow chart, where, with each iteration, the fitness of the particles is updated in the PSO and a chaotic search is performed to find the global optimal value.

Methods for Feature-based Challenges: Redundant

- ✓ Feature selection can rule out redundant features and select a subset of the features in the data without significantly degrading the performance of the model.
- ✓ In the realm of automatic feature extraction, rough set theory and autoencoders are two important automation methods. Rough set extracts features from network intrusion data and replaces original attribute values with discrete intervals to form an information system. Autoencoders are considered to be nonlinear generalizations of PCA, which use an adaptive, multi-layer network of "encoders" to reduce data dimensionality.



(a) Correlation-based feature selection [92]: The features are lined up on the horizontal and vertical axes of the correlation map. The method chooses features that are highly correlated with a class, but not correlated with each other. Genetic algorithm [58]: Ganapathy et al. identified a trending feature selection method using genetic algorithm that uses a fitness function and a decision tree where features are removed and model fitness so the optimal feature set is obtained.

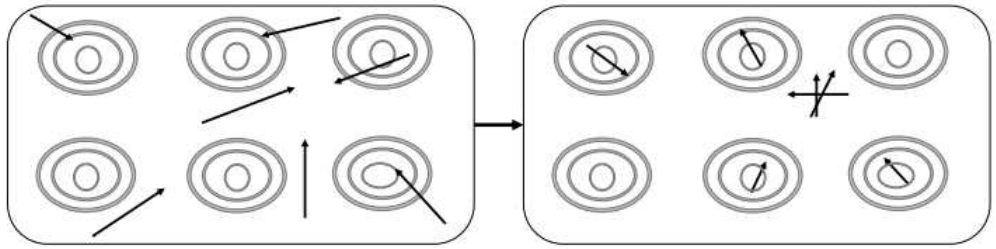


(b) Swarm optimization [34]: Chung and Wahid improved normal swarm optimization by conducted a local weighted search to avoid premature "optimal" solutions. Particles are shown as arrows in the figure and are updated by evolutionary operators. Depending on its fitness, its location is updated until the final feature set is optimal–the distribution of resulting particles after optimization shown below the first rounded rectangle.

Fig. 3. Methods for dealing with redundant features: Besides the above two methods, Autoencoder [8] and Fuzzy Rough Set [153] have also been used for reducing redundant features.

✓ Chung and Wahid [34] went about improving the performance of it by conducting a local weighted search after SSO to produce more satisfactory solutions. They applied k-means clustering to continuous network data values and rough set theory to minimally sized subsets of the feature. The goodness in selected features is evaluated using the fitness function given input data D, |C| being the number of features, |R| being the length of a feature subset where R is a feature subset, and γR as the classification quality of feature set R:

$$\alpha \times \gamma_R(D) + \beta \times \frac{|C| - |R|}{|C|}. \tag{3}$$



- ✓ The lack of strong correlation between features in data may make the construction of a model more challenging. Correlation can be artificially made through increasing the dimensionality of the data by data fusion or the introduction of new features.
- ✓ Increase dimensionality. Given one-dimensional feature data, Li et al. [102] augmented the data to two dimensions and performed data segmentation where split data was later fused back together for network intrusion classification. They split feature data into four separate parts based on features that are correlated with one another. The one-dimensional feature space is converted to grayscale, then the data output from the four data components are merged and passed to the output layer of the multi-fusion CNN.

- ✓ The model is robust if the accuracy of the predictions is not affected by changes in the input data, such as changes in the distribution or outliers. For intrusion detection, changes in network traffic data can also come from "adversaries" who may "obfuscate" the attack payload to mimic its benign counterpart. In order to reduce the influence of noise or adversary on intrusion detection accuracy, different robustness methods have been proposed.
- ✓ Recent work focused on designing methods robust to distributional changes or the presence of outliers in network data. Papers either tackle method-specific limitations in robustness such as the sensitivity of SVMs to noise or general limitations that result in high **false positive rates** or undetected **false negative outliers**.

$$g(\mathbf{x}) = \begin{cases} +1 & \text{if } f(\mathbf{x}) \geq \epsilon_0 \\ -1 & \text{if } f(\mathbf{x}) \leq -\epsilon_0 \\ 0 & \text{otherwise} \end{cases}$$

normal network connection data: $T^2_{Fast-MCD,i} = (\mathbf{x}_i - \mathbf{T})^T \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{T}).$

control limit: $CL_{KDE} = \widehat{\tilde{F}}_h^{-1}(\tilde{t})(1-\alpha).$

Mininum Covariance Determinant: The accuracy of outlier detection is improved and more robust estimators are obtained T and S are the mean and covariance of the new permutation set.

Tolerance to Adversarial Attacks

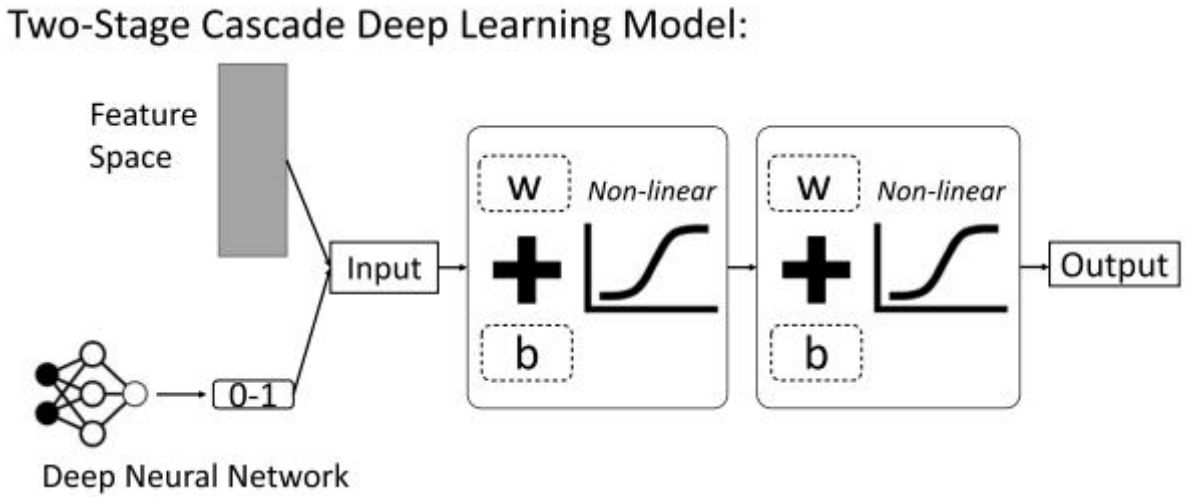
- ✓ An “adversary” can be a data generator or a network security expert that can mask network payloads to appear benign when they are in fact malicious.
- ✓ Marino and others have implemented an adversarial approach that attempts to improve the algorithm by enabling machine learning models to correctly classify erroneous samples by generating adversarial samples. Rather than tricking classifiers with linear models and multilayer perceptron models, the authors wanted to understand why their network data was misclassified.
- ✓ The constraint: $\mathbf{x}_{min} \preceq \widehat{\mathbf{x}} \preceq \mathbf{x}_{max}$

$$min_{\hat{x}} H(\hat{y}, p(y, \hat{x}, w)) \alpha I_{(\hat{x},\hat{y})} + (\hat{x} - x_0)^T Q(\hat{x} - x_0)$$

Q is a positive semidefinite matrix that can be adjusted using weights.

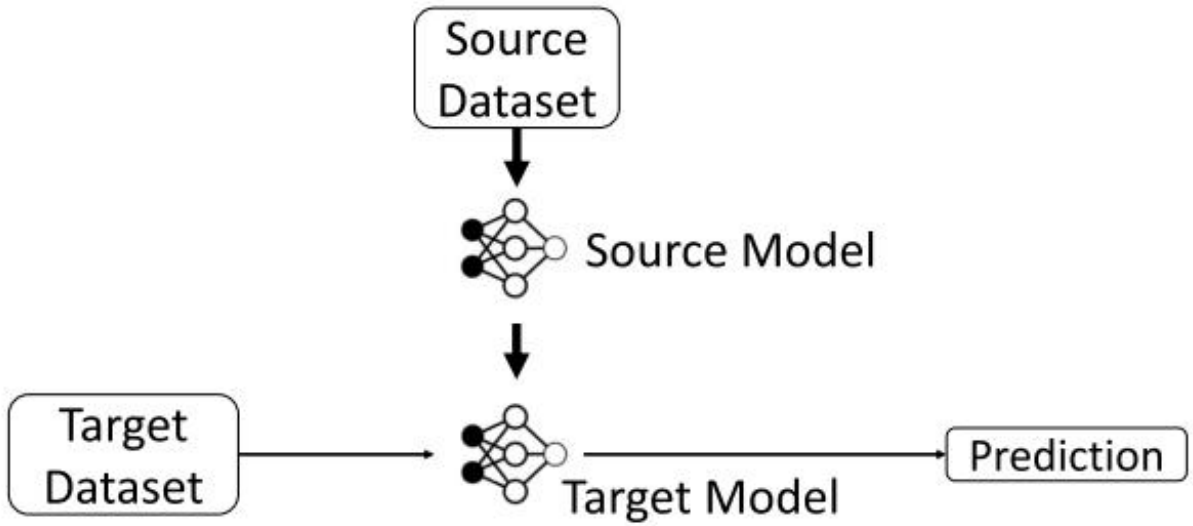
Methods for Label-based Challenges: Too Few Labels

- ✓ Data may have a lack of labels, particularly when network traffic is ambiguous or unlabeled. This poses another challenge between the stages of data preprocessing and model creation.
- ✓ Figure 4 illustrates transfer learning, adversarial sample generation, and deep learning paradigms used to resolve the issue of unlabeled data.

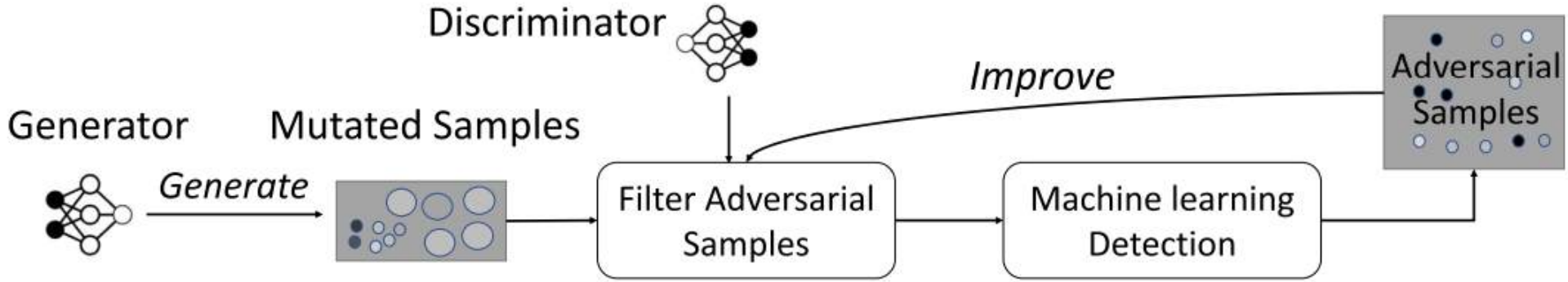


Methods for Label-based Challenges: Too Few Labels

✓ Transfer learning can compensate for the lack of labeled data via transfer of knowledge from other labeled data sources.

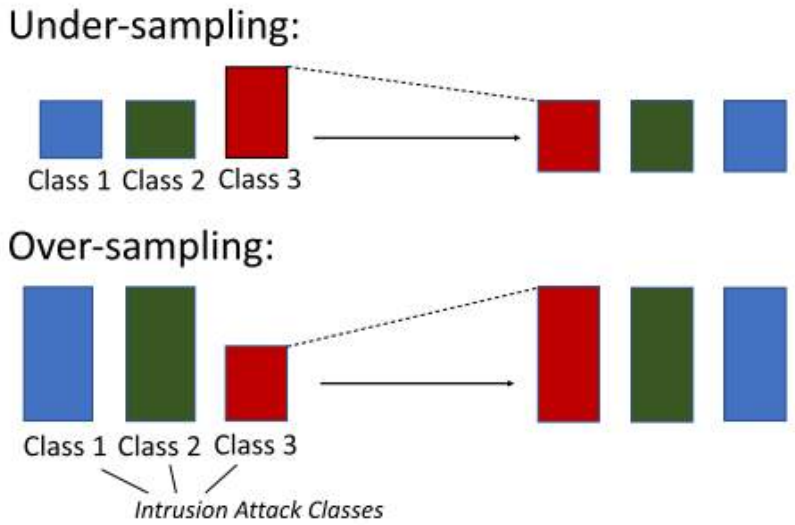


- ✓ Cheng [32] used a generative adversarial network (GAN) where a generator aims to refine the generation of fake data while a discriminator determines which network traffic flows are legitimate or anomalous.

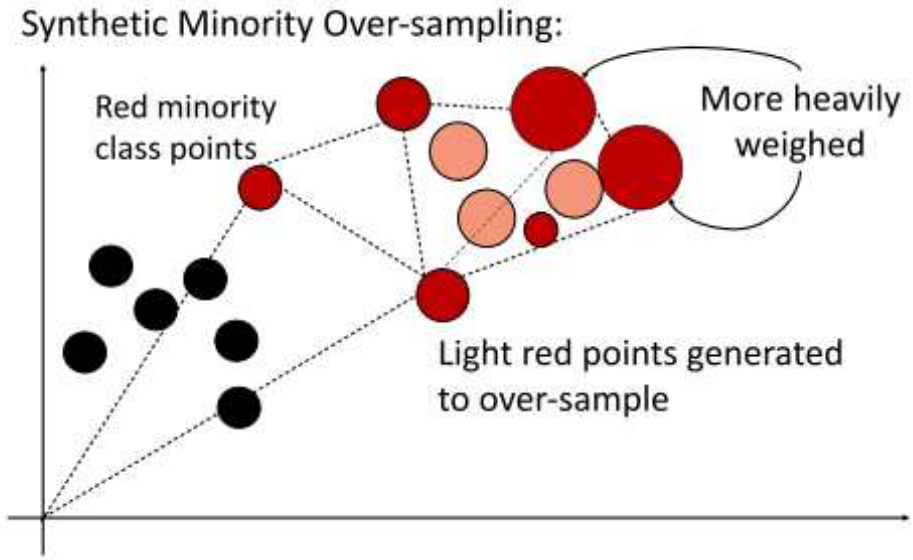


✓ The data may also be imbalanced where network intrusion attacks are disproportionately smaller than that of normal network activity. Figure 5 highlights the random oversampling and under-sampling techniques used to handle minority and majority classes in network intrusion datasets and main machine learning models implemented to handle unbalanced classes.

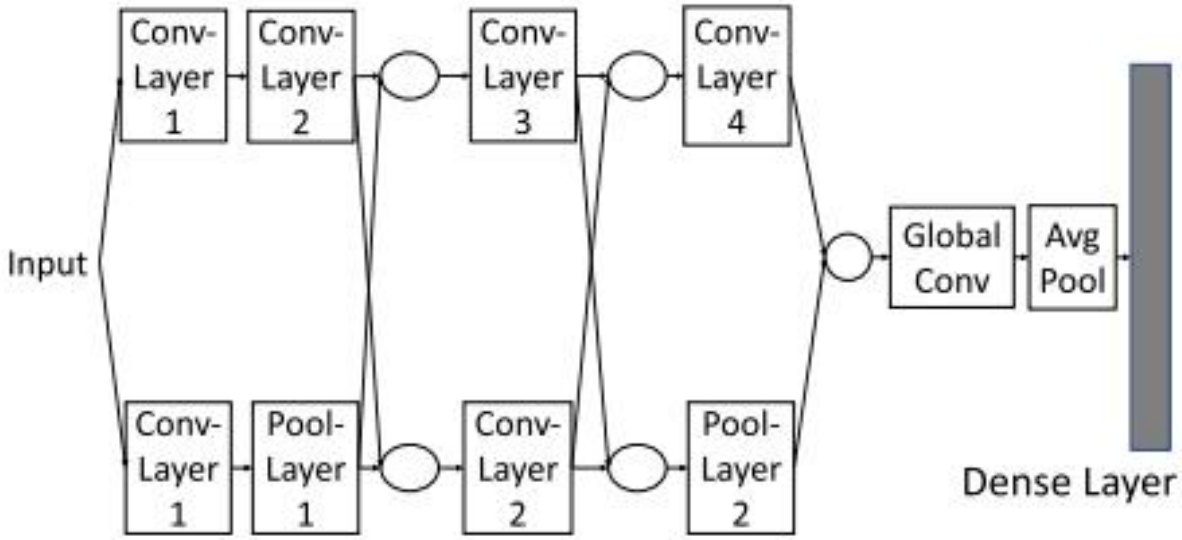
- ❖ Oversampling is meant to increase samples from the minority class and balance the distribution of data among attacks and normal activity in a network.
- ❖ Under-sampling moves samples from the majority class to allow minority and majority classes to become similar in size, disallowing misclassifications of underrepresented network attacks



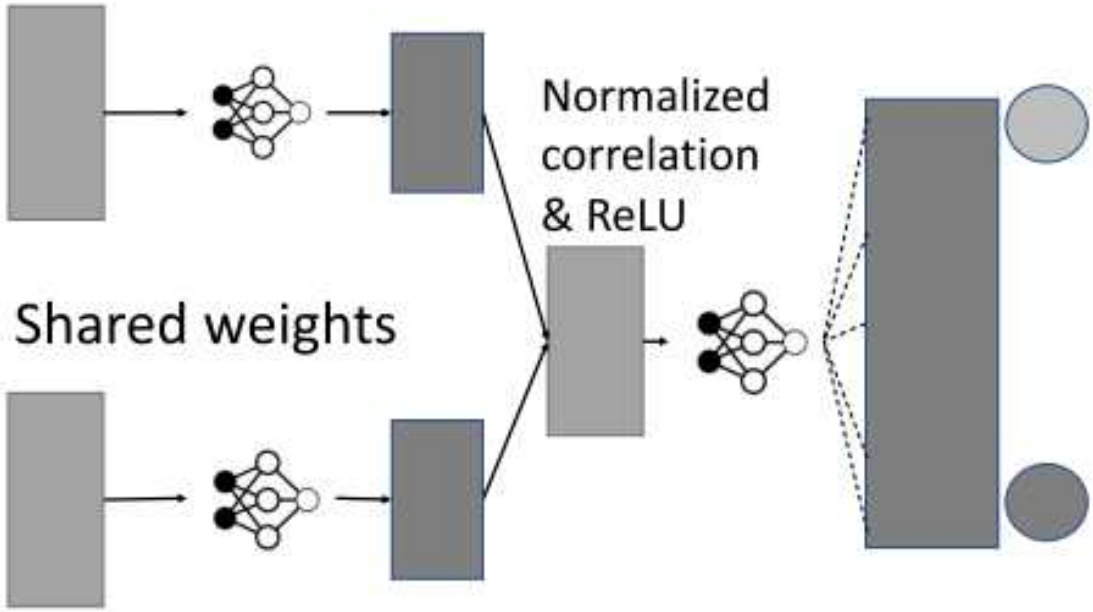
✓ Zhang et al. [193] resolved this issue by combining weighted oversampling with a boosting method. The weighted oversampling technique updates weights associated with minority classes and the misclassified majority class observations are forced on the classifier to learn.



✓ Zhang et al. [196] implemented a parallel cross-convolutional neural network (CNN), which fused the traffic characteristics learned by two independent convolutional neural networks, making the network pay more attention to a few attack classes. After under-sampling the two neural networks, the number of channels in the output feature map is doubled, and then a pooling layer is applied to reduce the dimensionality of the data by combining the outputs clustered in one layer into one neuron in the next layer.



✓ Data may have a lack of labels, particularly when network traffic is ambiguous or unlabeled. Marino and others have implemented an adversarial approach that attempts to improve the algorithm by enabling machine learning models to correctly classify erroneous samples by generating adversarial samples. Rather than tricking classifiers with linear models and multilayer perceptron models, the authors wanted to understand why their network data was misclassified.

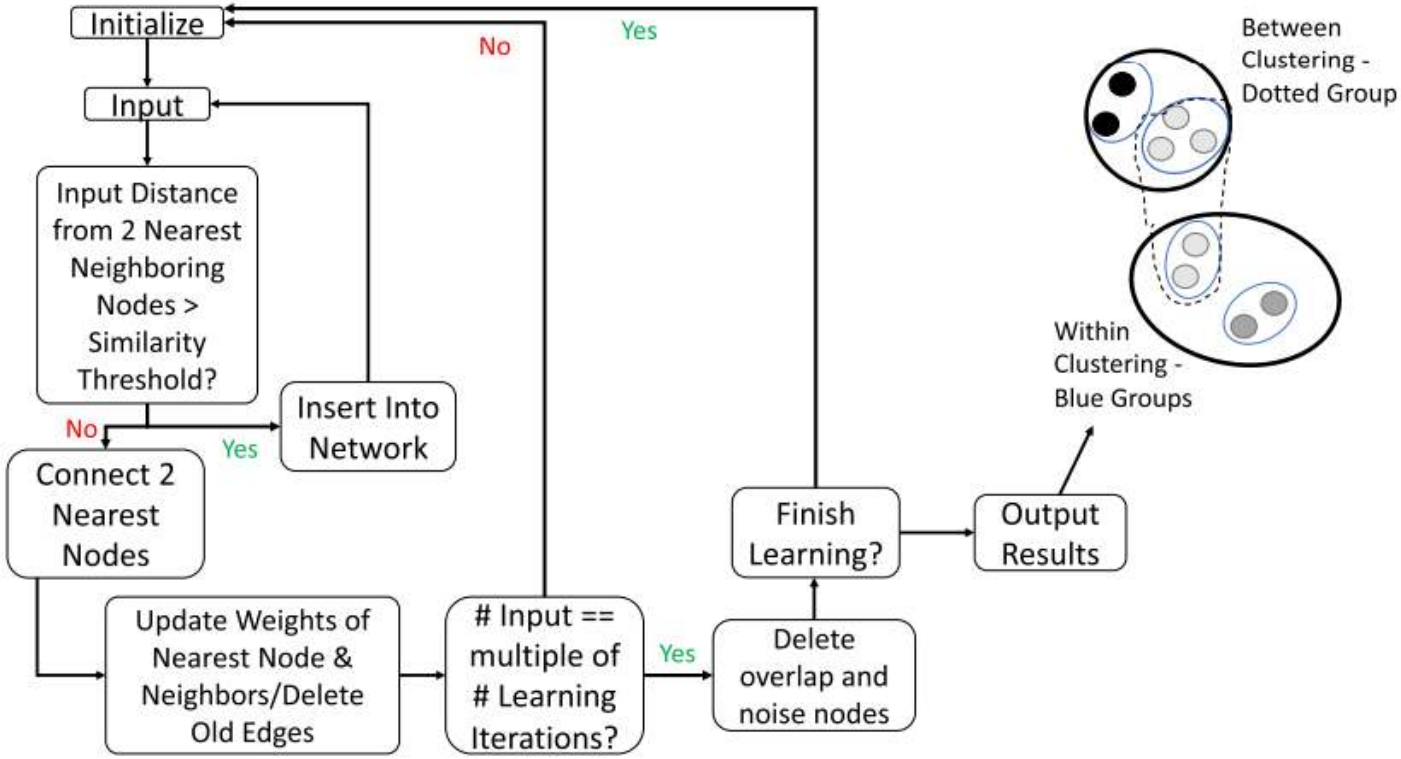


Methods for Label-based Challenges: Explainability

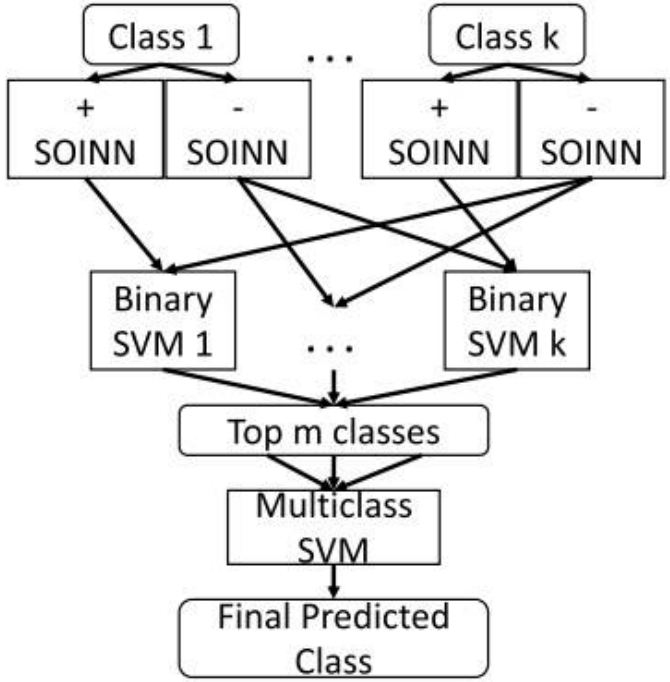
- ✓ Model interpretability is the ability to understand how a machine learning model behaves and why it proposes a particular solution or prediction for a given task. The interpretability of network intrusion detection models is very important because of the side effects of misprediction. If a model predicts a false negative, then the entire network may be at risk. Without an explanation for why a model makes such poor and costly decisions, individuals will be less receptive to new intrusion detection systems driven by machine learning models. Therefore, studying intrusion detection models that are easier to interpret can increase the chances of integrating models into corporate or public networks.
- ✓ To improve detection accuracy while maintaining model explainability, Amarasinghe and colleagues [12] present a post hoc framework of explanations for their deep neural network (DNN) predictions. They provide the relevance of input features for the prediction, the prediction confidence, and a textual summary of why the prediction was made such as “DoS attack WITH high confidence BECAUSE connections to the same host in the last 2 seconds is high.” The textual summary and feature relevancy are similar to rough set’s decision rules that are based off relevant features.

✓ Due to the changing landscape of new data being generated daily, adaptive models have been ever more important to dynamic data, especially as data has been growing exponentially for the past decade and now the digital world contains roughly 2.7 zetabytes.

Figure 6 summarizes the significant and novel dynamic network intrusion models developed recently.

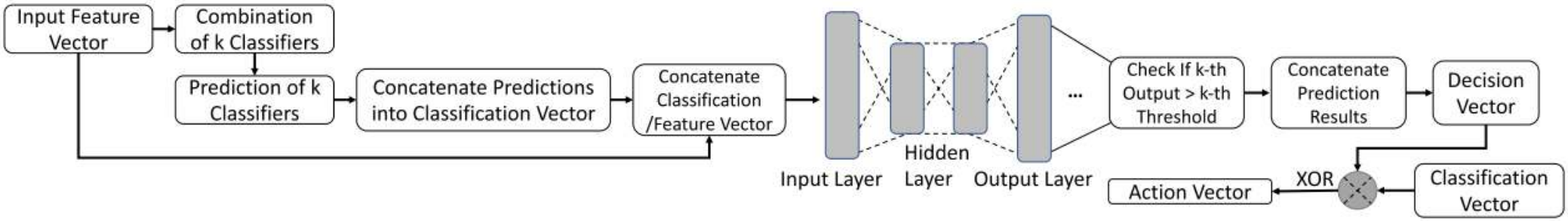


✓ Constantinides et al. self-organizing incremental neural network [36]: The detection system is initialized with a dataset containing k attack classes. Each attack class category is modeled with two self-organizing incremental neural networks. The input vector per SVM is constructed from that SVM's positive n-SOINN and other negative n-SOINNs from the other classes.



Methods for Label-based Challenges: Dynamic Data

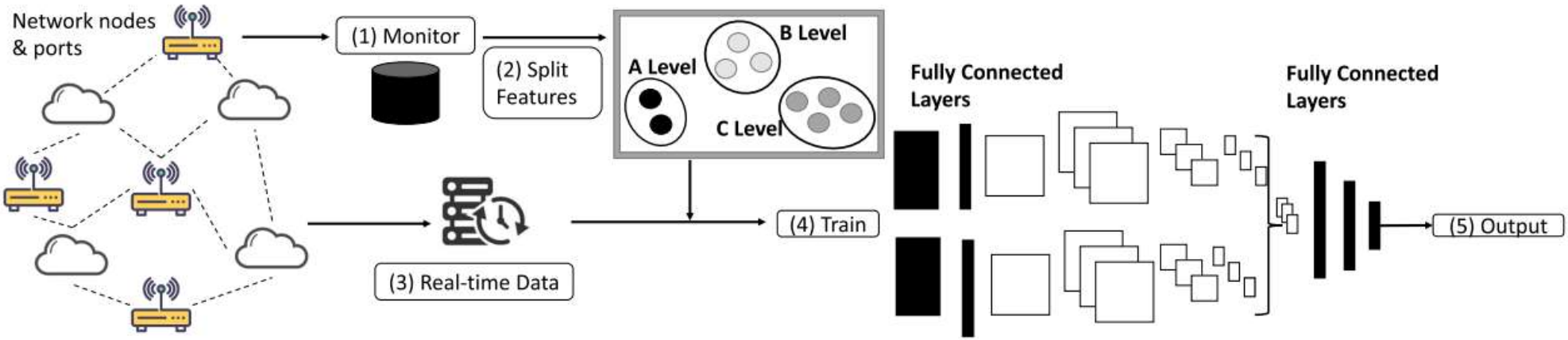
- ✓ There has recently been research on cloud environments and applying reinforcement learning to changing data in the cloud by Sethi et al. [154], who applied reinforcement learning to the cloud where a host network communicates with an agent network through VPN. Log generation from the virtual machine was provided to an agent that applied a deep Q-network and compared the model's result with the actual result from the administrator network, calculating the reward and iterating until the reward was maximized.



Methods for Label-based Challenges: Big Data

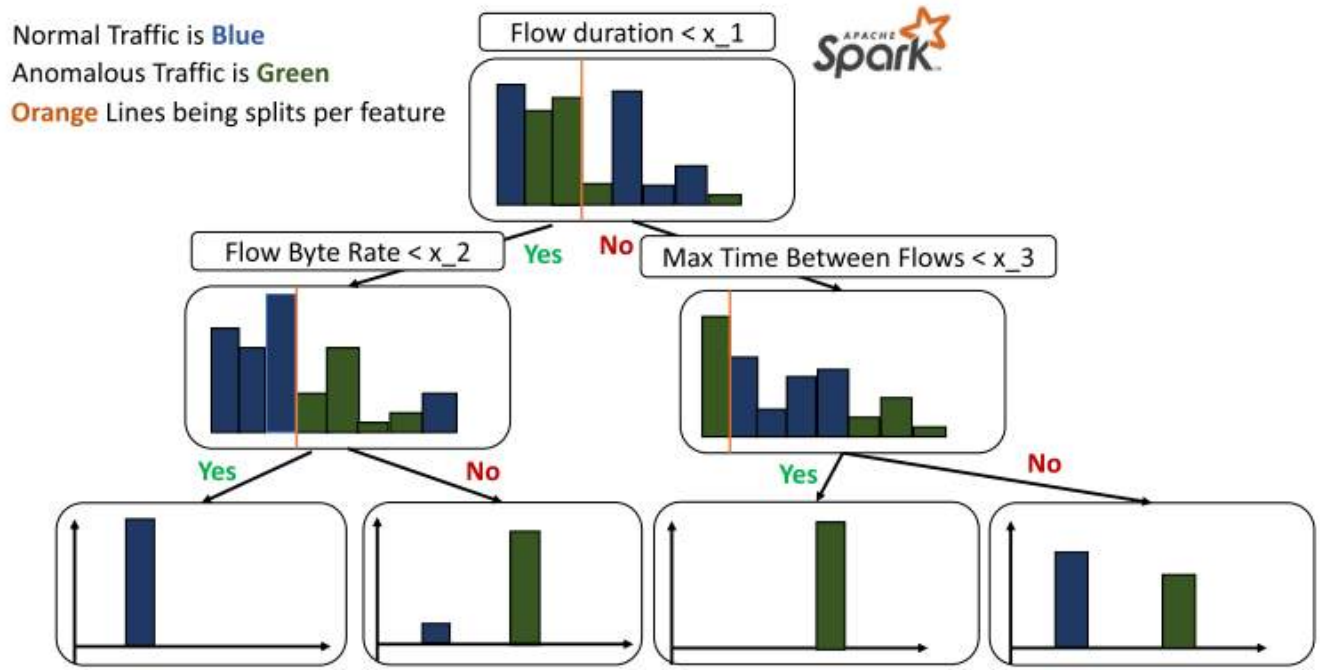
- ✓ For big data, processing such large amounts of data is overwhelming, so optimization methods were devised to speed up preprocessing, such as reduction methods that remove redundant features and reduce the size of the data.

- ✓ Figure 7 depicts the paradigms from three pivotal methods handling large amounts of data using incremental learning, parallel processes, and Apache Spark for Cloud Computing.



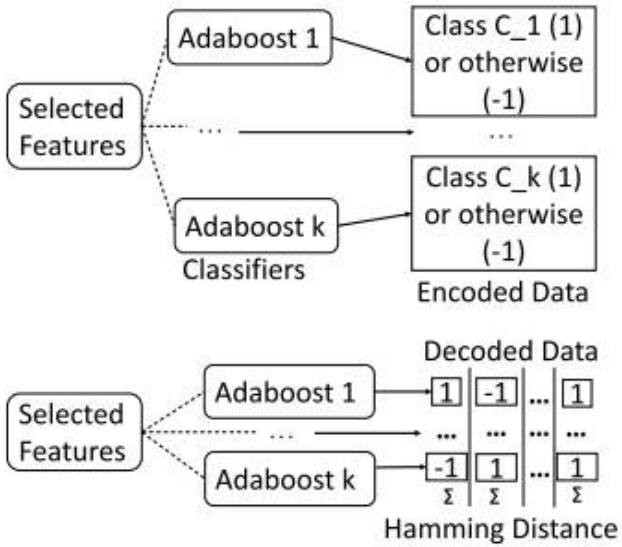
Methods for Label-based Challenges: Big Data

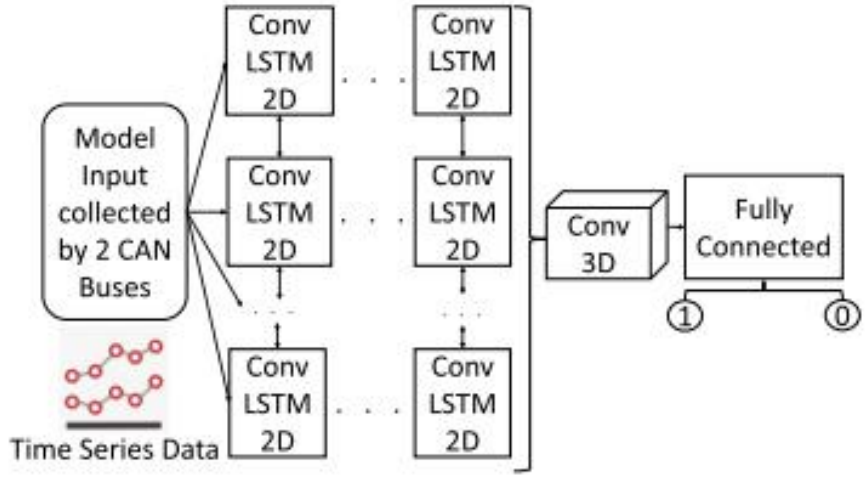
- ✓ Morfino and Rampone Apache Spark with decision tree [121]: Morfino et al. used Apache Spark's Machine Learning Library, MLlib, which stores filter/map operations in a directed acyclic graph and uses “Catalyst" to optimize an efficient execution plan. The decision tree splits a distribution by features until splits divide features into more homogeneous groups of normal and anomalous traffic.



Methods for Label-based Challenges: Small Data

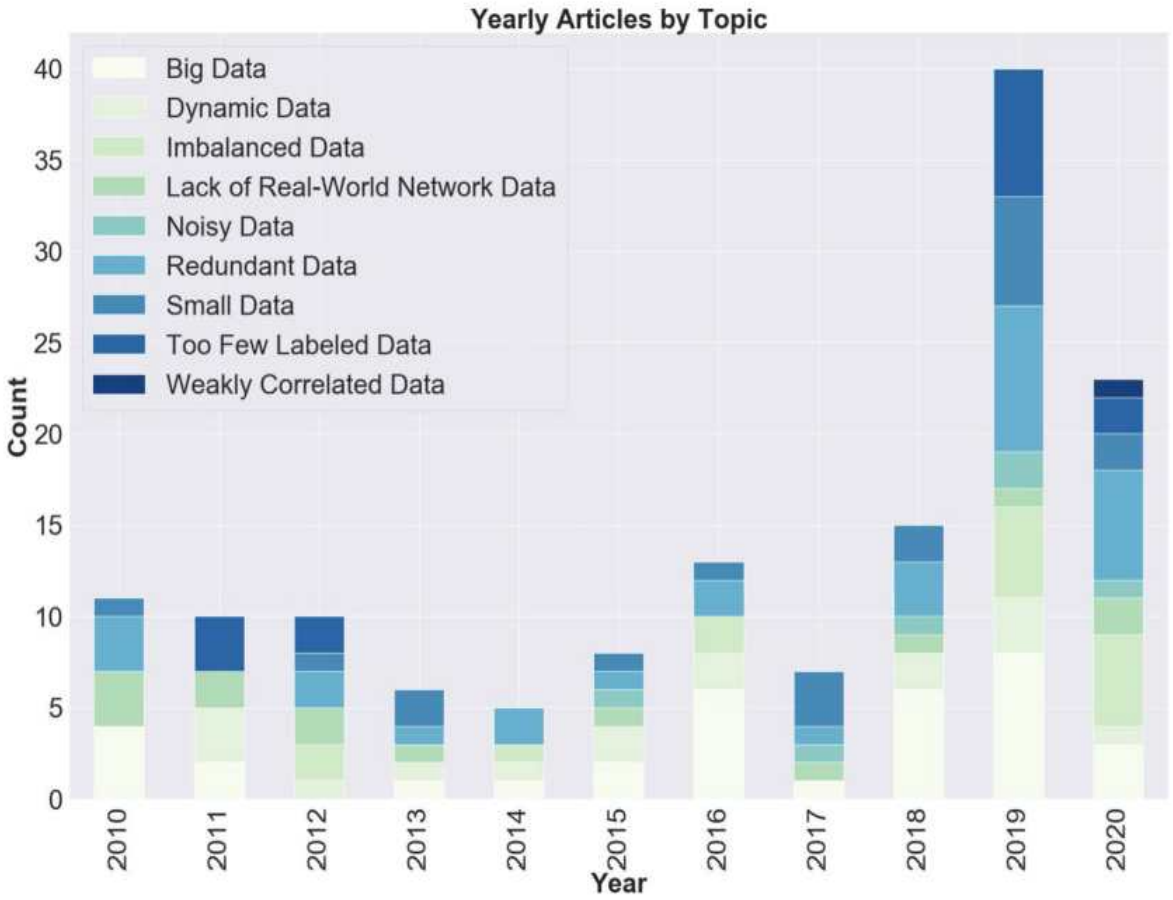
- ✓ The concomitant challenge with a growing network intrusion data repository is the continued lack of data on more current, diverse network attack types. The dataset is littered with attack categories that lack uniform representation. Some datasets are dominated by specific attacks, while other types of attacks are less representative. Some datasets are dominated by non-attack classes (benign classes), and all attack types are minority classes. To solve the problem of small data size, especially the lack of attack types, meta-learning and transfer learning techniques have been explored.





Research Trends

- There was already a pre-existing interest in big data research since 2010.
- 2019 saw the largest number of articles on big data where researchers continued to study parallel processing techniques and incremental learning methods to handle processing large amounts of data.
- Small data issues were first researched in the early 2010s, particularly with meta-learning.
- With noisy data challenges, authors have done more extensive research into methods that weigh noisy observations over others in network intrusion datasets since 2017.
- However, one area of research that has not seen much attention is real-world network data.



Discussion on Future Directions

- ✓ **Real World Data Collection**: There was initially a step towards real-world network intrusion data by emulating a realistic network environment. However, simulated data may not be as valuable to fit and test a model on as data collected on a real-world network due to possibly incorrect network attack models and behaviors in sandbox network environments. Network intrusion research requires further data collection of realistic attacks in real-world networks.
- ✓ **Labeling Real-world Traffic**: Although web traffic may be manually flagged by cybersecurity experts, real-world web traffic can easily grow into the millions. Future research focuses on the design of more adaptive detection models and the development of efficient traffic data annotation paradigm and technology
- ✓ **Consumer Network Intrusion**: Because consumer networks such as home networks do not have the same security resources as enterprise networks, there is a lack of datasets for collecting data on home networks.
- ✓ **Extending Anomaly Detection to Cloud Environments**: In addition to using cloud computing to accelerate model convergence and reduce anomaly detection time, exploring network intrusion in cloud environment has not been studied in depth. Another feature of today's cloud environment is that data is constantly changing. Due to the large amount of data stored on the cloud, developing machine learning for dynamic data on the cloud should be a future research step.
- ✓ **Machine Learning Scalability and Performance Improvements**: Parallelism in big data machine learning models can help researchers improve anomaly based intrusion detection methods. In the future, researchers should continue to investigate ways to make incremental machine learning models more scalable with huge data growth.

Conclusion

- ✓ This paper introduces a general taxonomy of data-driven network intrusion detection methods, and tests the common public data sets used in this taxonomy.

- ✓ Given the research trends over time, areas that require future research are web big data, streaming and changing data, and real-world web data collection and availability.

- ✓ Many solutions have been implemented for the other challenges specified in the taxonomy, but there is still a lack of real network data, especially consumer network data, which may limit the accuracy of model performance in simulated network environments using real network traffic data.

SEOULTECH 국립서울과학기술대학교

THANKS

서울과학기술대학교 컴퓨터공학과 진호천