Drift-based approach for evolving data stream classification in Intrusion detection system

Sugandh Seth[a], Gurwinder Singh[a], Kuljit Kaur Chahal[a]

[a] *Guru Nanak Dev University,Amritsar, India*

Abstract

Machine learning, and deep learning are extensively used to augment the performance of Intrusion detection systems. While the existing work on intrusion detection system using data mining and machine learning is efficient, but it involves training static batch classifiers to detect intrusions irrespective of the regular data stream's time-varying characteristics. Aims: This paper proposes an adaptive approach for online intrusion detection using stream-oriented learning for adapting to concept drift in real world environment. Method: Adaptive Random Forest classifier with ADWIN change detector is used for detecting change in a data stream and adapting to drift detection in the streamed data resulting in agile adaptation against unknown intrusions and the proposed approach also overcomes the need to retrain the model with time. Results: The latest CIC-IDS 2018 dataset is used for evaluating the approach. With the proposed method, the final Accuracy obtained is 99.5 % and a recall rate of 99.8%.

Keywords 1

Intrusion Detection System, Concept Drift, Stream oriented learning, Adaptive Random Forest,

1. Introduction

Data mining and machine learning application's prominence is increasing with time. Recently much research is being proposed to utilize machine learning and deep learning techniques in many domains such as weather forecasting, Spam detection, detecting fraudulent financial transactions, Intrusion detection system etc. Traditionally, machine learning was primarily focused on using static data enough to represent underlying distribution. However, usually, real-world problems don't fit in models with such restrictions. Also, many real-world applications such as Intrusion Detection Systems have non-stationary data distributions that cause the problem of non-stationary learning or concept drift over the time. Many studies with good results are available to investigate IDS (Intrusion Detection Systems) using deep learning and machine learning approaches. However, most of the studies have deployed static data sources. These studies fail to take rapid technological developments, and the problem of concept drift into account [1], leading to poor performance of the system.

As an indicator, concept drift holds importance due to its ability to measure time-based data distribution variance. Besides, IDS can also be considered as a typical scenario of concept drift. Usually, for a single source providing a data stream to a network, the data under the scanner is in a stable state, distributed identically. And in case of an unknown intrusion, the current data distribution undergoes dynamic changes as compared to the historical data. This motivates for building and adaptive Intrusion detection method that involves incremental learning based on Concept Drift that quickly adapts to new intrusion types.

Moreover, the endless emergence of new attacks and security loopholes raises the need for an ideal classifier that quickly adapts to intrusion's emerging methods. In such a situation, the static batch

WCNC-2021: Workshop on Computer Networks & Communications, May 01, 2021, Chennai, India.
EMAIL: kuljitchahal.cse@gndu.ac.in (Kuljit Kaur Chahal)
ORCID: 0000-0003-3785-116X (Kuljit Kaur Chahal)

learning approach discussed above delivers poor performance. In other words, when a static classifier goes obsolete, its response to new intrusion types becomes slow, and a re-training [2] becomes mandatory that needs high-cost investments. Contrarily, incremental learning with the adaptive updating of the classifier to a regular data stream over time ensures the promising performance of an IDS.

Thus, current IDS models have numerous noteworthy drawbacks.

- The underlying model of current IDS usually detects only known network attacks whereas IDS are prone to novel malicious attacks.
- The IDS models are built on static data. Whereas data come in streams in an IDS, and the data distribution may vary over the time.
- IDS models become obsolete with time and needs to be retrained which is cost intensive.

To overcome the above research gaps, we propose a concept drift-based approach for evolving data stream classification in Intrusion detection system. With limited processing and memory time, the concept drift detection method ensures accurate and quick identification of changes in the underlying data point distribution, followed by the fastest possible adaptation in the model [3].

The major contributions of the paper are as follows:

- High performance intrusion detection system based on streamed data with concept drift.
- Incremental IDS model with adaptive updation of classification models, achieving high accuracy in real time.
- IDS with agile adaptation against unknown intrusions.

2. Related Work

Lot of research is done in the field of intrusion detection. Many researchers have proposed machine and deep learning techniques for detecting intrusions.

Ferrag et al. in [4] evaluated seven deep learning techniques namely: Deep neural network(DNN), recurrent neural network(RNN), restricted Boltzmann machine(RBM), deep belied networks(DBN), recurrent neural network, convolutional neural network(CNN), deep Boltzmann machine(DBM) and Deep Autoencoders(DA) on the latest CIC-IDS 2018 dataset. Though this paper has comprehensively evaluated the deep learning techniques, but all the evaluation is done on the static data.

(Karatas et al. in [5] evaluated six machine learning-based IDS using K Nearest Neighbour, Random Forest, Gradient Boosting, Adaboost, Decision Tree, and Linear Discriminant Analysis algorithm on the CIC-IDS 2018 dataset. This proposed approach focuses on balancing the skewed dataset using oversampling with SMOTE. Though this approach gives good results, but it is also on static data.

In [6] Roshan et al. proposed an adaptive Intrusion detection system based on extreme learning an clustering. The proposed model was evaluated using the NSL-KDD dataset and have claimed to achieve an accuracy rate of 89% for novel attacks.

Feng et al. in [7]developed a plug and play to capture the packets. Deep learning techniques were used for detecting DOS attacks, CNN was used to detect XSS and LSTM was used for detecting SQL Injection.

In [8] Yuan et al.proposed concept drift-based ensemble incremental approach for intrusion detection system. The HDDM drift detection method based on Hoeffding's bounds was used to detect the anomaly and the ensemble based incremental learning using weighted voting was used for classification. All the experiments were done on the NSL-KDD dataset. They claimed to have achieved an accuracy of 94.91 % with the proposed approach. In [3] Breve & Zhao,proposed semi supervised classification with concept drift for intrusion detection. The study is based on passive drift detection

without explicitly using any algorithm and is inspired by the competitive and cooperative behavior of some animals to protect their territory. It is based on natural way of learning new data and forgetting the older ones. The proposed algorithm evaluated on the KDD Cup 1999 dataset.

Park et al. in [9] proposed online eigenvector transformation for reflecting concept drift detection in Intrusion Detection System. In Online PCA eigenvectors were computed by converting the existing eigenvector as per the latest data without generating a new eigenvector. They compared the performance of network intrusion detection using both online and offline PCA. Both the methods gave good precision rate but the recall rate for online PCA outperformed the offline method.

The majority of the proposed work in literature is on old datasets or is based on static datasets. To overcome the above research gaps, In this paper Adaptive Random forest classifier with drift detection to classify attacks in stream data using the latest CIC-IDS 2018 dataset.

The rest of the paper is structured as follows. Section 2 reviews the current literature Intrusion detection systems. Section 3 discusses the research methodology. Section 4 discusses the results obtained. Section 5 concludes the paper with a summary.

3. Research Methodology

In the network security domain, malicious intrusions have increased, making the IDS (Intrusion Detection System) design vital for securer systems. Recently, machine learning methods are increasingly used for network abnormality detection. However, currently available works do not explore the variation in data over the time, restricting their ability to detect new intrusion types. Therefore, for unpredicted changes in the status data's statistical properties over time, we propose an IDS with a concept drift-based incremental learning using adaptive random forest classifier with adwin drift detection.

3.1. Data Preprocessing

3.1.1. Data Collection

The proposed study is done on the latest CIC IDS 2018 dataset. The CIC IDS 2018 is a massive dataset that incorporates 14 modern-day attacks. The CIC IDS dataset was published by Communications Security Establishment (CSE) & the Canadian Institute for Cybersecurity (CIC). The dataset comprises of 80 features with 16 million rows.

3.1.2. Data Transformation

Table 1: Count of benign and various attack sessions in the CIC-IDS 2018 dataset

Label	Count
Benign	13484708
Bot	286191
Brute Force -Web	611
Brute Force -XSS	230
DDOS attacks-HOIC	686012
DDOS attacks-LOIC-UDP	1730
DDOS attacks-LOIC-HTTP	576191
DoS attacks- Golden Eye	41508
DoS attacks-Hulk	461912
Dos attacks-SlowHTTPTest	139890

Dos attacks-Slowloris	10990
FTP-BruteForce	193360
Infiltration	161934
Sql Injection	87

The CIC-IDS dataset comprises of 13 modern-day attacks as listed in the table 1. The dataset is relabeled to Attack and Benign sessions. All the 13 attack types listed in the table are relabeled to attack class. Thus, the problem of multi classification of attacks is reduced to binary classification. Thus, the dataset mix after preprocessing is listed in table 2.

Table 2. CIC-IDS 2018 dataset after pre-processing

Label	Count
Benign	12615791
Attack	2586295

3.2. Training the Model

Any modification in the underlying process of data generation is referred to as concept drift. In the classification context, concept drift points to variation in the target variable's statistical properties. The target variable is the one the model is trying to make a time-based prediction for, and the term concept is used for the quantity the researcher aims to predict. As previously mentioned, the distribution that creates the data stream's items can change with time. To address this problem of concept drift in Intrusion Detection System the proposed model uses Adaptive Random Forest with Adaptive Windowing method for concept drift detection (ADWIN).

Adaptive Random Forest

Random Forest is a popular learning algorithm in regression and non-stream classification (batch) tasks. This approach creates multiple trees, avoiding overfitting of the branches through bootstrap aggregation for decorrelation [10] and random feature selection when nodes split. For creating each tree's bootstraps, the original Random Forest passes over the input data multiple times. It also passes over a part of the original features for each of the tree's internal nodes.

Performing multiple passes becomes infeasible when using data stream learning with input data. Thus, Random Forests need to adapt to the streaming data based on:

- A suitable process for online bootstrap aggregation
- Limitation of each leaf split decision into a feature subset

To achieve the second requirement, the algorithm for the base tree is modified [11]. For effective modification, the set of features to be taken for further splits is limited to a random subset of m size. Here, m<M with M is the total feature count. For non-streaming bagging, random samples are drawn by replacing the training set to create bootstrap samples (size Z) that are then used to train each of the n base models. An original training instance can be found K times in every bootstrap sample, with P (K=k) following a binomial distribution. Binomial distribution for instances with higher Z values adheres to a Poisson distribution (λ=1). Adaptive Random Forests, uses Poisson (λ=6) instead of (λ=1) as in leveraging bagging [12]. This uses resampling, causing a practical impact of increasing the chances of higher weight assignment to instances while the base models are trained.

However, when working with Adaptive algorithms, the primary aim is to deal with data streams that evolve over time. Thus, including other strategies it is also essential to cope with the problem of concept drifts. Adaptive Random Forest deploy a permissive threshold in ARF for warning detection instead of a tree reset process on drift detection. Alongside, background trees are created and trained along with

the ensemble without influencing its predictions. However, on the detection of a drift a background tree is used to replace the originating tree for the warning signal.

Thus, three major features of Adaptive Random Forest are as follows:

1. Increasing the variance with resampling
2. Increasing the variance with random selection of feature subsets for node splits based on hoeffding tree
3. Drift detection per base tree

Adwin Drift Detection

Adaptive Random Forest algorithm uses ADWIN (Adaptive Windowing) for drift detection. ADWIN [13] is a sliding window algorithm to detect drifts in a data stream. This algorithm is based on keeping statistics of a variable sized window to detect concept drift. The window size is computed by cutting the statistics widows at various points and comparing the average of various statistic measure over different windows. A drift is detected if the difference between the average statistics is above a certain threshold value. The proposed model uses adaptive random forest with adwin drift detection to adapt the model to concept drift in streams over the time. The results of the proposed model is discussed in section 4.

4. Results and Discussions

In this study, the performance of the adaptive Random Forest algorithm is evaluated. on the latest CIC-IDS 2018 dataset. The proposed system was implemented using scikit-multiflow library in python. All the experiments were performed on AWS cloud platform using the configuration in Table 4.

Table 3. AWS configuration for conducting the experiments

Hardware	Properties
VCPU	32
PLATFORM	AMAZON LINUX(Version 31.0)
MEMORY	256 GB
INTERNAL STORAGE	24* 1980 GB
NETWORK PERFORMANCE	25 GIGABIT

The proposed model is evaluated on various performance metrics - Accuracy, Precision, Kappa, Recall, F- Measure. The performance metrics [14] can be evaluated using the following equations(1-5)

$$Accuracy = \frac{TruePositive + TrueNegative}{TruePositive + FalseNegative + TrueNegative + FalsePositive} \quad (1)$$

Accuracy is the percentage of samples that were correctly classified.

$$Recall = \frac{TruePositive}{TruePositive+FalseNegativ} \quad (2)$$

Recall is the ratio of samples correctly classified as attack among all attack samples.

Kappa = (total accuracy – random accuracy) / (1- random accuracy) (3)

The coefficient of kappa calculates an agreement between the values of classification and truth value.

A kappa value of 1 stands for perfect agreement, while a value of 0 does not stand for agreement.

$$Percision = \frac{TruePositive}{TruePositive + FalsePositive} \quad (4)$$

Precision is the ratio of correctly classified samples as attacks with total samples graded as attacks.

$$F1\ Score = 2 * \frac{Percision * Recall}{Percision + Recall} \quad (5)$$

The F1 Score is the harmonic mean of recall and Accuracy.

Besides Accuracy other performance evaluation parameters such as Precision, F-Measure, Kappa and Recall metrics are also important to evaluate the performance of the system. Analysis of result using the above performance metrics with other adaptive algorithms is listed in Table.

The CIC-IDS 2018 dataset is converted into a data stream and holdout evaluation method is used for evaluating the performance of the model. The holdout evaluation method updates the statistics of incoming samples without evaluating the performance or predicting the labels. The performance of the model is evaluated after every n samples. The evaluation is done on the unseen test samples and the test sets are dynamically generated from the data stream.

4.1. Analysis of Results of the proposed method in Comparison to other Adaptive Learning Algorithms on Stream Data

Table 4: Results of the proposed method in comparison to other adaptive learning algorithms on stream data

Model	Drift Detector	Accuracy	Kappa	Precision	Recall	F1 Score
Adaptive Random Forest	ADWIN	0.995	0.9983	0.9992	0.9980	0.9986
Adaptive KNN	ADWIN	0.9743	0.9088	0.9257	0.9228	0.9242

The comparison of the proposed adaptive Random Forest with other adaptive learning methods is done in Table 4. It is evident from the above results that the adaptive random forest model outperforms the other models with high accuracy rate of 99.5% and recall rate of 99.8%.

4.2 Analysis of Results of the proposed method in Comparison to other Learning Algorithms on Stream Data

Table 5. Results of the proposed method in Comparison to other Learning Algorithms on Stream Data

Model	Accuracy	Kappa	Precision	Recall	F1 Score
Adaptive Random Forest with ADWIN	0.995	0.9983	0.9992	0.9980	0.9986
Naïve Bayes	0.9821	0.9368	0.9454	0.9497	0.9476

KNN	0.9158	0.7469	0.6739	0.9774	0.7978

The performance of the proposed model is further compared with other stream-based learning models in Table 5. The performance of the Adaptive Random forest is undoubtedly better that other stream-based learning approaches.

4.3 Analysis of Results of the proposed method in Comparison to other Machine and Deep learning algorithm on static data.

Table 6 lists the comparison of average accuracy of proposed adaptive random forest approach on stream with deep learning approaches on the static data. It is evident from the above results that the adaptive random forest outperforms other deep learning approaches [4] Ferrag et al. (2020).

Table 6: Results of the proposed method in Comparison to other Deep learning algorithm on static data.

Model	**ARF(Proposed)**	**DNN**	**RNN**	**CNN**
Average Accuracy	99.5	95.95	97.64	83.90
Model	RBM	DBN	DBM	DA
Average Accuracy	95.49	96.99	96.79	97.44

5. Conclusion

Most of the applications based on streaming data need a fast response, requiring re(training) of an algorithm with the latest data available. Most of the existing Artificial Intelligence based IDS models are trained on static data. Whereas data come in streams in an IDS, and the data distribution may vary over the time since attack patterns tend to evolve over the time resulting in a concept drift. Moreover, for the IDS to work efficiently it needs to adapt itself to be able to detect new attack classes over the time. In such scenarios the static data models performs poorly since static batch learning models get outdated and needs to be updated with time. To overcome the above challenges, this paper uses adaptive random forest classifier with ADWIN drift detector for building concept drift-based model for evolving data stream classification in Intrusion detection system. The proposed algorithm outperforms other approaches in literature with high accuracy, precision and recall rate of 99.5 %, 99.9% and 99.8% respectively.

6. References

[1] Xu, R., Cheng, Y., Liu, Z., Xie, Y., & Yang, Y. (2020). Improved Long Short-Term Memory based anomaly detection with concept drift adaptive method for supporting IoT services. Future Generation Computer Systems, 112, 228–242. https://doi.org/10.1016/j.future.2020.05.035

[2] Folino, G., Pisani, F. S., & Pontieri, L. (2020). A GP-based ensemble classification framework for time-changing streams of intrusion detection data. Soft Computing, 24(23), 17541–17560. https://doi.org/10.1007/s00500-020-05200-3

[3] Breve, F., & Zhao, L. (2013). Semi-supervised Learning with Concept Drift Using Particle Dynamics Applied to Network Intrusion Detection Data. 2013 BRICS Congress on Computational Intelligence and 11th Brazilian Congress on Computational Intelligence. https://doi.org/10.1109/BRICS-CCI-CBIC.2013.63

[4] Ferrag, M. A., Maglaras, L., Moschoyiannis, S., & Janicke, H. (2020). Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study. Journal of Information Security and Applications, 50, 102419. https://doi.org/10.1016/j.jisa.2019.102419
[5] Karatas, G., Demir, O., & Sahingoz, O. K. (2020). Increasing the Performance of Machine Learning-Based IDSs on an Imbalanced and Up-to-Date Dataset. IEEE Access, 8, 32150–32162. https://doi.org/10.1109/access.2020.2973219
[6] Roshan, S., Miche, Y., Akusok, A., & Lendasse, A. (2018). Adaptive and online network intrusion detection system using clustering and Extreme Learning Machines. Journal of the Franklin Institute, 355(4), 1752–1779. https://doi.org/10.1016/j.jfranklin.2017.06.006
[7] Feng, F., Liu, X., Yong, B., Zhou, R., & Zhou, Q. (2019). Anomaly detection in ad-hoc networks based on deep learning model: A plug and play device. Ad Hoc Networks, 84, 82–89. https://doi.org/10.1016/j.adhoc.2018.09.014
[8] Yuan, X., Wang, R., Zhuang, Y., Zhu, K., & Hao, J. (2018). A Concept Drift Based Ensemble Incremental Learning Approach for Intrusion Detection. 2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), 350-357.
[9] Park, S., Seo, S., Jeong, C., & Kim, J. (2019). Online eigenvector transformation reflecting concept drift for improving network intrusion detection. Expert Systems, 37(5), 1. https://doi.org/10.1111/exsy.12477
[10] Breiman L. Random Forests. Machine Learning. 2001; 45:5-32. Available from: https://doi.org/10.1023/a:1010933404324
[11] Gomes, H. M., Bifet, A., Read, J., Barddal, J. P., Enembreck, F., Pfharinger, B., Holmes, G., & Abdessalem, T. (2017). Adaptive random forests for evolving data stream classification. Machine Learning, 106(9–10), 1469–1495. https://doi.org/10.1007/s10994-017-5642-8
[12] Bifet, A., Holmes, G., Kirkby, R., & Pfahringer, B. (2010). Moa: Massive online analysis. The Journal of Machine Learning Research, 11, 1601–1604.
[13] Bifet, A., & Gavaldà, R. (2007). Learning from Time-Changing Data with Adaptive Windowing. Proceedings of the Seventh SIAM International Conference on Data Mining, April 26-28, 2007, Minneapolis, Minnesota, USA. https://doi.org/10.1137/1.9781611972771.42
[14] Shone, N., Ngoc, T. N., Phai, V. D., & Shi, Q. (2018). A Deep Learning Approach to Network Intrusion Detection. IEEE Transactions on Emerging Topics in Computational Intelligence, 2(1), 41–50. https://doi.org/10.1109/tetci.2017.2772792