IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 29, 2019, accepted June 10, 2019, date of publication June 13, 2019, date of current version June 28, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2922737

Statistics-Enhanced Direct Batch Growth Self-Organizing Mapping for Efficient Dos Attack Detection

XIAOFEI QU[1,2], LIN YANG[2], KAI GUO[2], LINRU MA[2], TAO FENG[2], SHUANGYIN REN[2], AND MENG SUN[1]
[1]Command and Control Engineering College, Army Engineering University of PLA, Nanjing 210007, China
[2]National Key Laboratory of Science and Technology on Information System Security, Institute of Systems Engineering, AMS, Beijing 100039, China

Corresponding authors: Xiaofei Qu (crane0106@163.com) and Kai Guo (guokai07203@hotmail.com)

ABSTRACT As an artificial neural network method, self-organizing mapping facilities efficient complete and visualize high-dimensional data topology representation, valid in a number of applications such as network intrusion detection. However, there remains a challenge to accurately depict the topology of network traffic data with unbalanced distribution, which deteriorates the performance of *e.g.* DoS attack detection. Hence, we propose a new model of the ''statistic-enhanced directed batch growth self-organizing mapping'', renew the definition of the growth threshold used to evaluate/control neuron expansion, and first introduce the inner distribution factor for fine-grained data distinguishing. The numerical experiments based on two datasets, KDD99, and CICIDS2017, demonstrate that the key performance in DoS attack detection including the detection rate, the false positive rate, and the training time are greatly enhanced thanks to the statistic concepts consulted in the proposed model.

INDEX TERMS DoS attack detection, statistic-enhanced directed batch growing self-organizing mapping, growth threshold, inner distribution factor.

I. INTRODUCTION

The denial of service (DoS) attack is arguably the most common network intrusion, which can greatly occupy the resources of legitimate requests and long-term paralyze the network. With the explosive increase of Internet bandwidth and the rapid development of various DoS hacking tools, the frequent DoS attack becomes emerged. According to Kappaski Labs, DoS attack keeps increasing in 2018 compared to that in 2017 [1], which reveals the fact that special attention is needed for efficient DoS attack detection. In the past decade, various methods of Dos attack detection have been developed, yet suffer from the challenge of data classification for large-scale datasets [2]–[9]. The conventional data classification approaches with the supervised learning mechanisms often require transcendental data characterization, which is not always valid in real-world cases. On the other hand, accurate data topology representation enables a stronger understanding of data characterization, which further facilitates high-performance intrusion detection. Hence, the strong need of realizable data classification, preferable with visualization scheme, is emerged.

Self-organizing mapping (SOM) is an visualized artificial neuron network method, which uses the unsupervised learning mechanisms to discrete high-dimensional data into low-dimensional data (often two-dimensional) for data topology representation [10]. More specifically, the input layer of SOM receives high-dimensional data, while the low-dimensional the output layer (also named competition layer) achieves the data mode clustering. For a certain data mode, the node (winning neuron) in output layer gets the maximal stimulus, while the neurons around the node are partially stimulated. Hence, the feature map of output layer well reflects the distribution of input data modes. Behaving as both a visual tool and a data mode classifier, SOM is naturally employed in a number of applications such as the pattern recognition [11], the fault diagnosis, the anomaly detection [12], and the DoS attack detection [13], [14].

The associate editor coordinating the review of this manuscript and approving it for publication was Jiafeng Xie.

However, previous studies have theoretically predicted that the simple form of SOM is insufficient for the topology representation of the network traffic data with unbalanced distributions [15], [16], that is, the SOM-based intrusion detection comes across low performance.

Since accurate and overall representation of data topology makes possible efficient DoS attack detection, several improved SOM models have been proposed in the past decade. The growing hierarchical self-organizing mapping (GHSOM) expanding neurons along both horizontal and vertical directions, benefits a more complete data topology representation [17], which further enables higher detection rate and lower false positive rate for DoS attack detection [18], [19]. However, the growth strategy in the GHSOM introduces vast unnecessary neurons, which greatly enlarges the computing redundancy and reduces the efficiency of the data clustering. To mitigate this, a growth threshold is introduced in both the growth self-organizing mapping (GSOM) [20] and the directed batch growth self-organizing mapping (DBGSOM), in which the neuron growth takes the cumulative error into account. Therefore, new neurons are inserted around each candidate boundary neuron with an optimized growth location and a proper weight. However, both the GSOM and the DBGSOM employ an input-data-independent constant growth threshold definition, thus remain insufficient in intrusion detection of network traffic data.

In this paper, we propose a new model, the statistic-enhanced directed batch growth self-organizing mapping (SE-DBGSOM), which well suffices DoS attack detection for different datasets. The novelties of this work are mainly three-folds: (1). For the first time to our best knowledge, we renew the definition of the growth threshold to make it input-data-dependent. Hence, the initial data mode clustering becomes more efficient. (2). For the first time to our best knowledge, we propose an inner distribution factor, which facilitates further fine-grained classification for the remaining deeply-bunched data after the neuron growth process. (3). We demonstrate the numerical experiments using datasets of KDD99 and CICIDS2017, where the detection rate, the false positive rate and the training time, are the best state-of-the-art compared to related works. The experiments explicitly validate the statistic enhancement for the proposed model, and enlighten a new direction of intrusion detection with various data types.

II. RELATED WORKS

For the first time, [21] proposed the GHSOM in network intrusion detection, demonstrating that the unsupervised learning mechanism sufficed the cases with complex and unbalance-distributed data. Reference [19] demonstrated DoS attack detection using the GHSOM, where the detection rate reached 97.59%. Yang *et al.* first introduced the tension-mapping ratio to control the neuron growth, where the detection rate was 96.71% [22]. Note that the strategy of inserting neuron-row or -column, corresponding to large computing redundancy, is not always necessary. Hence, with the dynamic incremental strategy, GSOM makes possible a more elastic neuron growth, where the training time is greatly shortened [20], [23]. Nevertheless, conventional GSOM models often fill all free spaces around the candidate neuron, which makes the representation of data topology low-quality and time-consuming. Although reference [24] modified the GSOM algorithm using batch learning strategy and shortened the training time, previous drawbacks remained unsolved. On the other hand, Vasighi and Amini developed the DBGSOM in 2017 [25], where a batch learning strategy taking the cumulative error into account was employed. Only a single neuron was inserted at a suitable position around the candidate boundary neuron with a proper initial weight. Hence, the neuron growth could be controlled along the right direction in the topology mapping. Compared to the GHSOM and the GSOM, the DBGSOM overcomes huge computing burden, improves the representation quality of data topology and enlarges the training efficiency in large-data-scale cases.

Although the DBGSOM introduces a good mechanism of data topology representation, the growth threshold that determines whether new neurons are inserted, is independent of the input data. Specifically for network intrusion detection, such a constant growth threshold becomes insufficient for various data types, which results in low performance. Moreover, the data clustering in the neuron growth process may leave some deeply-bunched attack and normal data, which cannot be classified through the cumulative error between neuron and data. Hence, it is of great necessity to introduce another parameter, quantifying the inner distribution properties of data, to enable the further fine-grained data classification. Since both the renewed data-dependent growth threshold and the newly-given inner distribution factor consult the statistic mechanism, the proposed model used for *e.g.* DoS attack detection in this paper, behaves as the new edition of the DBGSOM with statistic enhancement.

III. THEORETICAL APPROACH OF SE-DBGSOM

A. BASIC MODEL SET-UP

Similar to that shown in [25], our model holds the merits of the DBGSOM, where a visualize feature map can be set up without needing pre-specification of network size in the initialization step. In the neuron growth process (training process), the cumulative error (*i.e.* the sum of Euclidean distance) between all input data vectors and a specific neuron (serial number of l) follows

$$CE_l = \sum_{j=1}^{m} ||x_j - w_l|| \tag{1}$$

where subscripts j denotes the serial number of input data (total amount of m). The data vector x_j and the initial neuron weight vector w_l share the same dimension. With the minimal cumulative error with respect to all neurons, the weight vector

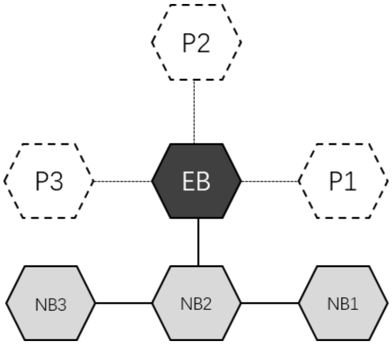


FIGURE 1. Three available positions for the 3p neuron insertion strategy.

of the winning neuron becomes

$$w_l^{new} = \frac{\sum_{j=1}^{m} h_{c_j,l} x_j}{\sum_{j=1}^{m} h_{c_j,l}} \tag{2}$$

where $h_{c_j,l}$ denotes the Gaussian neighborhood function

$$h_{c_j,l} = \exp\left(-\frac{\left\| w_l - w_{c_j} \right\|^2}{2\sigma^2(t)}\right) \tag{3}$$

On the other hand, the growth threshold GT is utilized to decide the new neuron growth, where $CE_l > GT$ corresponds to the insertion action, with a position handled in three strategies, 1p, 2p and 3p [25]. As an example, Figure. 1 shows the three available positions (P1, P2 and P3) for the most common 3p neuron insertion strategy, where the gray and the black hexagons denote the neighbor boundary (NB) and the expandable boundary (EB), respectively. The new neuron is inserted to Pi when the cumulative error CE_{NBi} with respect to NBi reaches the maximum among all neighbor boundaries. In addition, the weight vector of the new neuron is initialized by three strategies, 1W, 2W and 3W [25]. The most representative 3W strategy of weight vector initialization follows

$$w_{new} = \begin{cases} [(2w_{eb} - w_{nb2}) + w_{nbi}]/2 & \text{insert } p_i \; i \neq 2 \\ 2w_{eb} - w_{nb2} & \text{insert } p_2 \end{cases} \tag{4}$$

where w_{eb} and w_{nbi} denote the weight vectors of the expandable boundary and the neighbor boundary, respectively.

B. STATISTIC-BASED GROWTH THRESHOLD

The growth threshold behaves as the minimal cumulative error for new neuron insertion, which heavily determines the trade-off between complete data topology representation and few unnecessary neurons. **It is worth noting that the constant growth threshold in previous studies does not suffice all types of datasets, even using an empirical value from large-scale experiments.** Hence, we renew the definition in a statistic manner as

$$GT = \lambda \sqrt{\sum_{i=1}^{D} std_i^2} \tag{5}$$

where subscript i denotes the serial number of items for the D-dimensional data vectors. The proposed growth threshold depends on the input data vectors, thus makes possible a more accurate and overall reflection of various datasets. More significantly, such a definition reflects the standard derivation of the i-th items for all input data vectors (total amount of m)

$$std_i = \sqrt{\frac{\sum_{j=1}^{m} (X_{ij} - \frac{1}{m}\sum_{j=1}^{m} X_{ij})^2}{m-1}} \tag{6}$$

where X_{ij} denotes the i-th items for the data vector with a serial number of j. Note that Eq. (6) takes the similar formula to Eq. (1), that is, the cumulative error and the growth threshold are comparable. Hence, for a specific input dataset, the growth threshold remains unchanged, yet holds the statistic properties of all data vectors. Additionally, a regulation coefficient λ is introduced in Eq. (5), as a degree of freedom for various attack detection cases.

C. INNER DISTRIBUTION FACTOR

A single network traffic data is commonly a D-dimensional column vector, where the items in normal data vectors greatly differ from that in attack data vectors. **To quantify the statistic properties of these items, we newly define the inner distribution factor, which makes possible another approach of data topology representation.** For a specific type of data, the mean value of all items is given by

$$\overline{MX} = \frac{1}{mD} \sum_{j=1}^{m} \sum_{i=1}^{D} X_{ij} \tag{7}$$

meanwhile the mean standard error is given by

$$M\delta = \frac{1}{mD} \sum_{j=1}^{m} \sqrt{\sum_{i=1}^{D} (X_{ij} - \frac{1}{D}\sum_{i=1}^{D} X_{ij})^2} \tag{8}$$

By taking the concept of mean summarization [26] into account, the inner distribution factor follows

$$\text{IDF} = (\overline{MX} - M\delta, \overline{MX} + M\delta) \tag{9}$$

Since IDF for normal data differs from that for attack data (especially DoS attack data), Eq. (9) provides a scheme of statistic-based data classification (detection). More specifically for the proposed SE-DBGSOM model, IDF facilitates fine-grained classification for deeply-bunched data vectors, remaining after the neuron growth clustering.

D. SCHEMATIC AND ALGORITHM

A schematic of the proposed SE-DBGSOM model is shown in Figure. 2, where the first step is the initialization with respect to both input data vectors and neuron weight vectors. Based on the input-data-dependent growth threshold, the neuron growth (training) process takes place, where most of the normal/attack data vectors are clustered in different neurons. There may remain a few deeply-bunched data vectors, normal

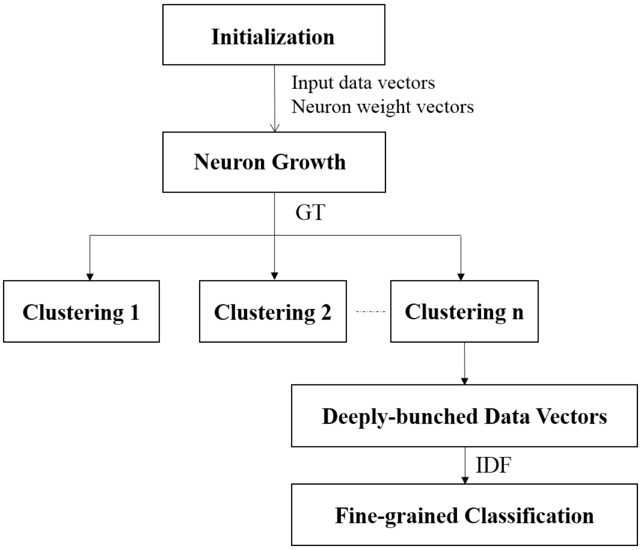


FIGURE 2. A schematic of the proposed SE-DBGSOM model.

and attack types included, that cannot be classified through further neuron growth, thus IDF is utilized. Thanks to the statistic enhancement, the SE-DBGSOM enables efficient intrusion detection with higher detection rate, lower false positive rate and shorter training time. The algorithm of the SE-DBGSOM training and the fine-grained classification are given in Algorithm.1 and Algorithm.2, respectively.

Algorithm 1 SE-DBGSOM Training

```
Initialization: Set the training epochs at 100; Calculate the growing threshold using the training dataset; Optimize the regulation coefficient using the training dataset.
for $i = 1$ to 100 do
    The growth threshold follows Eq. (5) while other parameters value from Ref. [25]
end for
Return weight vectors and labels for winning neurons
```

IV. NUMERICAL EXPERIMENTS

A. DATASET

We based on the computing environment of @MATLAB 2017 a, operating system of @Windows 7 Professional and computer of @Intel Core i7-7700, 3.6GHz CPU, 8.0GB RAM, characterize the proposed SE-DBGSOM model using entire KDD99 and CICIDS2017 datasets. The benchmark KDD99 is arguably the most widely used dataset for intrusion detection experiments, which is created by MIT Lincoln Lab for IDS evaluation competitions held in 1998 and 1999. A detailed description of the 1998 DARPA off-line intrusion detection competition is shown in [27]. The KDD99 dataset [28], [29] investigated in this work includes four types of attacks: the DoS, the user to root (U2R, unauthorized access to local superuser by a local unprivileged user), the probe (surveillance and probing), and the remote to low frequency (R2L, unauthorized access from a remote machine to a local machine). Being written in CSV format, KDD99 data is a 42-dimensional vector, of which the last term labels the data type (normal or attack).

Algorithm 2 Fine-Grained Classification

```
Initialization: Input testing and training dataset; Calculate IDF for normal and DoS attack data vectors, respectively.
for $j = 1$ to length(testing dataset) do
    Calculate mean value of all items in each data vector in testing dataset.
end for
if the mean value is in-between IDF (DoS attack) then
    This data is DoS attack
end if
Return normal data in testing dataset.
```

Another investigated dataset CICIDS2017 comes from the project between the Communications Security Establishment and the Canadian Institute for Cybersecurity [30]. This dataset corresponds to the user profile that records network events and behaviors, to produce a diverse and comprehensive baseline dataset from intrusion detection. The original files (PACP and logs) can be used to summarize new features of the network traffic data, while CICFlowMeter is used for network traffic data analysis with respect to the tagged flows based on time timestamps, the source & destination IP, the source & destination ports, the protocols and the attacks. CICIDS2017 dataset is updated every 5 days, mainly including six types of attacks: the DoS, the Brute-force, the Heartbleed, the Botnet, the Web and the infiltration of the network inside. The exampled CICIDS2017 dataset comes from Friday traffic tracking, which is written in an 80-dimensional CSV format.

For the KDD99, the features extracted to model a malicious behavior of DoS attack mainly reflect in four items of the data vector. The 5-th, 6-th, 25-th and 27-th items in the DoS attack data vectors, corresponding to the src_bytes (the number of data bytes from source to destination), the dst_bytes (the number of data bytes from destination to source), the serror_rate (the percentage of connections that have "SYN" errors) and rerror_rate (the percentage of connections that have "REJ" errors), respectively, are far higher than those in the normal data vectors. For the CICIDS2017, the events used to distinguish DoS attack, mainly reflect in two types of items in the data vectors. The 2-th, 4-th, and 7-th items in the DoS attack data vectors, corresponding to the tot_fw_pk (the total packets in the forward direction), the tot_l_fw_pkt (the total size of the packet in forwarding direction) and the fw_pkt_l_avg (the average size of the packet in forwarding direction), respectively, are far higher than those in the normal data vectors. On the other hand, the 18-th, 23-th, and 28-th items in the DoS attack data vectors, corresponding to the fl_iat_min (the minimal time between two flows), the fw_iat_min (the minimal time between two packets sent in the forward direction) and the bw_iat_min (the minimal time

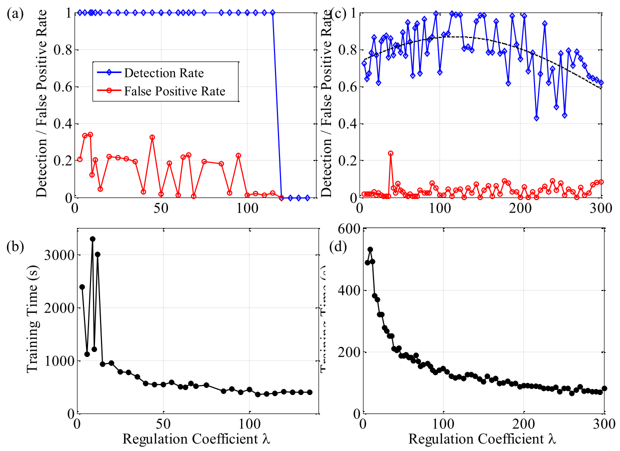


FIGURE 3. Neuron Expansion training using KDD99 : (a). DR (red) and FR (blue) versus λ (b). Training time (black) versus λ. and CICIDS2017 : (c). DR (red) and FR (blue) versus λ (d). Training time (black) versus λ.

between two packets sent in the backward direction), respectively, are far lower than those in the normal data vectors.

The numerical experiments are carried out using three manually-setting datasets. Set1 comes from the KDD99 with 21003 data vectors (including 9505 normal, 2008 DoS attack, 9173 U2R attack and 317 R2L attack) for training, and 8000 randomly-selected data vectors for testing. Being a large-scale dataset, Set2 also comes from KDD 99 with 95557 data vectors (including 76815 normal, 13467 DoS attack, 42 U2R attack, 1126 R2L attack and 4107 Prob attack) for training, and 10000 randomly-selected data vectors for testing. To validate that the SE-DBGSOM-based DoS attack remains efficient for different datasets, we also use Set3 from CICIDS2017 with 22705 data vectors (including 11469 normal and 11236 DoS attack) for training, and 8000 randomly-selected data vectors for testing.

B. THE REGULATION COEFFICIENT OPTIMIZATION

The key degree of freedom for the proposed SE-DBGSOM model is the regulation coefficient λ, which heavily determines the trade-off between high detection rate and low false positive rate. The detection rate is defined as $\mathrm{DR} = 1 - N_{fn}/N_{ta}$, where N_{fn} and N_{ta} denote the number of false negatives and the total number of attack connections, respectively. The false positive rate is defined as $\mathrm{FR} = N_{fp}/N_{tn}$, where N_{fp} and N_{tn} denote the number of false positives, and the total number of normal connections, respectively. DR quantifies the ratio of the attack data that can be correctly detected, while FR quantifies the probability that the normal data are misdetected as the attack data [31].

We utilize datasets from both the KDD99 (Set1) and the CICIDS2017 (Set3) to optimize the regulation coefficient λ, with respect to the DoS attack detection. Through the neuron growth process using KDD99, Figure. 3 (a) shows that when λ takes a value in-between 100 and 120, a near-unity detection rate and a near-zero false positive rate (less than 0.03) can be simultaneously achieved. Moreover, Figure. 3(b) shows that using a regulation coefficient of less than 15, the training time increases dramatically. It reveals the fact that a small value of λ results in a strict condition of clustering, thus the training time of neuron growth naturally trends longer. In addition, when λ is higher than 120, the resulting growth threshold fails to separate the normal and attack data vectors, where the detection rate drops to zero.

TABLE 1. Comparison between DBGSOM, GHSOM and SE-DBGSOM (without IDF).

Methods	Minimal FR	Maximal DR	Dataset
DBGSOM	4.14%	99.68%	Set1
SE-DBGSOM	**2.37%**	**99.70%**	
GHSOM	5.92%	99.72%	Set2
SE-DBGSOM	**1.48%**	**100%**	
DBGSOM	1.40%	71.69%	Set3
SE-DBGSOM	**0.65%**	**99.85%**	

For the neuron growth process using CICIDS2017, Figure. 3(c) shows that most of the detection rates for all regulation coefficient values are higher than 0.6, yet the near-unity results take place randomly, which reveals the fact that the DoS attack (DDoS attack included) data vectors in CICIDS2017 are not so characterization-clear compared to those in KDD99. Moreover, Figure. 3(d) shows that a large value of λ facilitates a small training time. In addition, through a trendline fitting for the detection rate versus λ, the optimized value is 115, which coincidentally approaches to that for KDD99, and is used in the following.

C. EXPERIMENTAL CHARACTERIZATION FOR SE-DBGSOM

To validate that the renewed definition of the statistic-based growth threshold facilitates higher detection rate and shorter training time, we experimentally compare the SE-DBGSOM (without IDF) and the conventional DBGSOM [25]. A series of repeating experiments (based on Set3) are carried out to present the evolution of the neuron growth process. The schematic of the numerical experiments includes three steps: The test data vectors are incident in the grown neuron network with trained weights; similar to that in the training process, the normal and DoS attack data vectors are clustered in existing neurons; the IDF is used to fine-grained classify the deeply-bunched data. As a result, Figure. 4(a) shows that the SE-DBGSOM facilitates a higher detection rate of DoS attack (mean value of 93.3% for 20-times experiments) compared to the DBGSOM (mean value of 67.9%). Figure. 4(b) shows that the neuron growth process for DBGSOM commonly takes a longer training time (mean value of 630 seconds) than that for SE-DBGSOM (mean value of 145 seconds). Moreover, Table. 1 shows that the SE-DBGSOM facilitates a higher detection rate and a lower false positive rate for Set3 compared to the DBGSOM. Although both the SE-DBGSOM and the DBGSOM achieve near-unity detection rate for Set1, an obvious enhancement is reflected at the false positive rate.

Thanks to the visualization mechanism, it is valid to learn the statistic enhancement from the output map of the neuron growth process. Figure. 5 shows a part of the output neuron map for both DBGSOM (a) and SE-DBGSOM (b), where

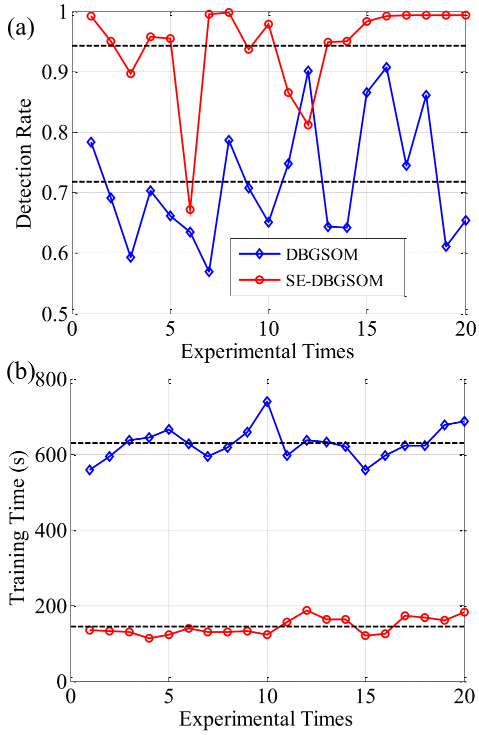


FIGURE 4. DBGSOM and SE-DBGSOM (without IDF) characterization (a). Detection rate. (b) Training time.

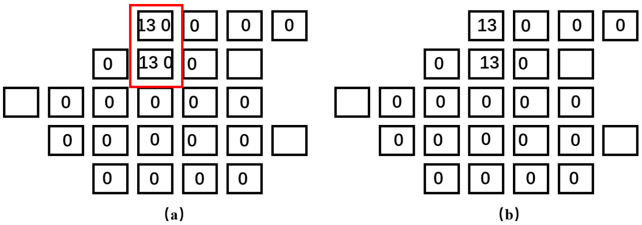


FIGURE 5. Output neuron map for (a). DBGSOM and (b). SE-DBGSOM.

each square represents a winning neuron. Through the neuron growth process, the data vectors sharing the same type are clustered in same neurons. However, false clustering takes place where the normal data vectors are marked attack, vice versa, or different types of data vectors are deeply-bunched. In Figure. 5, the normal data marked 0 and the DoS attack data marked 13 are clustered in the same neuron for DBGSOM, but is classified for SE-DBGSOM due to the input-data-dependent growth threshold definition. Additionally, since the output neurons for SE-DBGSOM (amount of 38) are fewer than that for DBGSOM (amount of 81), the training time is naturally shortened.

Numerical experiments using Set2 (from KDD99 but with more data vectors) are also carried out, to compare the SE-DBGSOM and the GHSOM. Figure. 6 shows a part of the output neuron map for GHSOM, where the normal data (marked 0) are not only bunched with the DoS attack data (marked 4 and 5), but also bunched with the U2R attack data (marked 2) and the Prob attack (marked 10 and 15). Such a deeply-bunched issue is mitigated for SE-DBGSOM, thus the false positive rate drops down to 2.37% (see Table. 1).

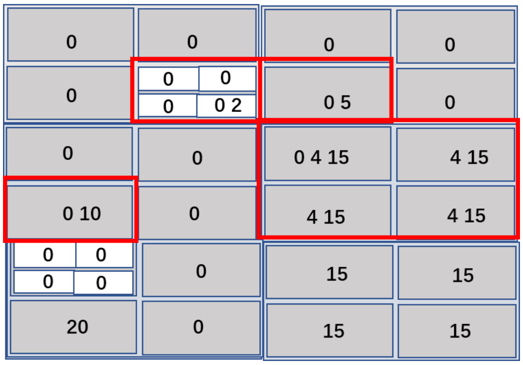


FIGURE 6. Output neuron map for GHSOM.

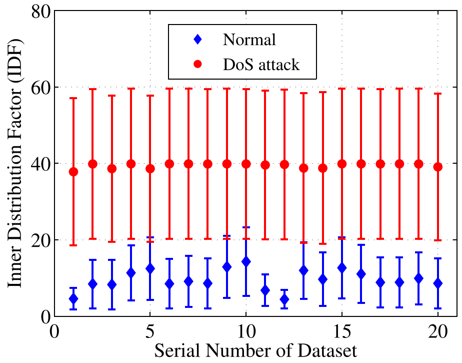


FIGURE 7. The IDFs of normal and DoS attack datasets.

D. FINE-GRAINED DATA CLASSIFICATION USING IDF

Although the statistic-based growth threshold facilitates higher detection rate and shorter training time, the remaining deeply-bunched data vectors contribute to a high false positive rate. Hence, the IDF is utilized for the fine-grained data classification mechanism after the neuron growth process, such that the false positive rate can be further reduced. We randomly extract 40 datasets from the testing part in Set2, either normal or DoS attack data included only, calculate the corresponding IDFs via Eq. 9, and compare them in an errorbar manner. IDF for normal and DoS attack datasets value in two independent zones except a few overlaps. By taking all measured results into account, the proper IDF for identifying DoS attack data, is at (19.92, 59.04). When the mean value of all items for an unknown data vector (in the testing dataset) falls into this range is marked DoS attack, or else marked normal.

Although the proposed IDF definition seems simple, it somehow reflects the inner characterization of items for different data types. Through the fine-grained classification, the remaining deeply-bunched data vectors can be further separated. To quantify this, experiments based on Set2 and Set3 are carried out, mainly concerning the false positive rate. Table. 2 shows that all of the concerned models facilitates

TABLE 2. Comparison between DBGSOM, GHSOM, REversible Sketch, and SE-DBGSOM (with IDF).

Methods	Minimal FR	Maximal DR	Dataset
GHSOM	3.25%	99.99%	Set2
DBGSOM	0.65%	100%	
SE-DBGSOM (with IDF)	**0.59%**	**100%**	
GHSOM	7.8%	86.5%	Set3
DBGSOM	1.4%	71.79%	
SE-DBGSOM (with IDF)	**0%**	**100%**	

TABLE 3. Comparison between SE-DBGSOM (with IDF) and other methods based on KDD99.

Methods	Minimal FR	Maximal DR
K-means [2]	0.76%	98%
NB [3]	8.4%	94.3%
KM+NB [3]	0.5%	99.6%
k-NN [4]	40.24%	95.87%
SVM [4]	4.19%	82.85%
TANN [4]	3.08%	90.94%
PCA-SVM [6]	0.89%	96.19%
LDA-SVM [6]	2.03%	91.60%
PCA-LDA-SVM [6]	1.96%	96.77%
A machine-learning-based [5]	2.57%	97.70%
DCNN [7]	1.31%	93.20%
Sequential classifiers [8]	27.42%	99.70%
A centroid-based technique [9]	29.66%	99.24%
SE-DBGSOM (with IDF)	**0.59%**	**100%**

a near-unity detection rate for Set2, yet the SE-DBGSOM with IDF has the lowest false positive rate of 0.59% (calculated from 40 parallel experiments using independent randomly-selected datasets), which evaluates the potentially best behavior of correcting the false positive. On the other hand, only the SE-DBGSOM with IDF possibly achieves a (coincidental) unity detection rate and a zero false positive rate for Set3 simultaneously. Since the IDF-based data classification benefits an ultra-low false positive rate, it becomes another evidence of the statistic enhancement.

Finally, we compare our experimental results to those using conventional DoS attack detection methods based on KDD99. Table. 3 shows that by trading off high detection rate and low false positive rate, the proposed SE-DBGSOM model facilities the best state-of-the-art among all related works. Since IDF in the proposed model only concerns the DoS attack data vectors, it naturally comes at the cost of failing to improve the detection performance of other attack types. This issue can, in future works, be solved by employing various statistic-based evaluation parameters, such that the data topology can be more exhaustively represented.

V. CONCLUSION

We propose a new model of directed batch growth self-organizing mapping concerning the emerged needs of DoS attack detection, which facilitates accurate topological representation for network traffic data. Special attention is given to the definition of the input-data-dependent growth threshold and the data-vector-items-based inner distribution factor, which consult simple statistic principle but is proved efficient. Based on the optimized regulation coefficient, numerical experiments using two different datasets, KDD99 and CICIDS2017, are carried, where the detection rate, false positive rate and training time, reaches the state-of-the-art among related works. Being the first proof-of-concept of the statistic enhancement in the proposed model, this work holds great potential in intrusion detection and other applications which need accurate data topology representation.

REFERENCES

[1] *Kappaski Labs*. Accessed: Apr. 26, 2018. [Online]. Available: https://securelist.com/ddos-report-in-q1-2018/85373/

[2] K. Nalavade and B. B. Mehsram, "Evaluation of k-means clustering for effective intrusion detection and prevention in massive network traffic data," *Int. J. Comput. Appl.*, vol. 96, no. 7, pp. 9–14, 2014.

[3] Z. Muda, W. Yassin, M. N. Sulaiman, and N. I. Udzir, "Intrusion detection based on K-Means clustering and Naïve Bayes classification," in *Proc. 7th Int. Conf. Inf. Technol. Asia*, 2011, pp. 1–6.

[4] C.-F. Tsai and C.-Y. Lin, "A triangle area based nearest neighbors approach to intrusion detection," *Pattern Recognit.*, vol. 43, no. 1, pp. 222–229, 2010.

[5] I. Z. Muttaqien and T. Ahmad, "Increasing performance of IDS by selecting and transforming features," in *Proc. IEEE Int. Conf. Commun.*, Dec. 2017, pp. 85–90.

[6] A. A. Aburomman and M. B. I. Reaz, "Ensemble of binary svm classifiers based on pca and lda feature extraction for intrusion detection," in *Proc. Adv. Inf. Manage., Commun., Electron. Autom. Control Conf.*, 2017.

[7] X. Wang, C. Zhang, and K. Zheng, "Intrusion detection algorithm based on density, cluster centers, and nearest neighbors," *Chin. Commun.*, vol. 13, no. 7, pp. 24–31, 2016.

[8] D. C. Corrales, J. C. Corrales, A. Sanchis, and A. Ledezma, "Sequential classifiers for network intrusion detection based on data selection process," in *Proc. IEEE Int. Conf. Syst.*, Oct. 2016, pp. 001827–001832.

[9] M. S. Gondal, A. J. Malik, and F. A. Khan, "Network intrusion detection using diversity-based centroid mechanism," in *Proc. 12th Int. Conf. Inf. Technol.-New Generat.*, 2015, pp. 224–228.

[10] *Self-Organizingmap*. Accessed: Jun. 9, 2018. [Online]. Available: https://commons.wikimedia.org/wiki/Category:Self-organizing_map

[11] M. Liukkonen and Y. Hiltunen, "Recognition of systematic spatial patterns in silicon wafers based on SOM and K-means," *IFAC-PapersOnLine*, vol. 51, no. 2, pp. 439–444, 2018.

[12] P. Mishra, V. Varadharajan, U. Tupakula, and E. S. Pilli, "A detailed investigation and analysis of using machine learning techniques for intrusion detection," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 686–728, 1st Quart., 2019.

[13] T. M. Nam, P. H. Phong, T. D. Khoa, T. T. Huong, and V. D. Loi, "Self-organizing map-based approaches in DDoS flooding detection using SDN," in *Proc. Int. Conf. Inf. Netw.*, Jan. 2018, pp. 249–254.

[14] D. Li, G.-Q. Ni, Z.-S. Pan, and G.-Y. Hu, "DDoS intrusion detection using generalized grey self-organizing maps," in *Proc. IEEE Int. Conf. Grey Syst. Intell. Services*, Nov. 2007, pp. 1548–1551.

[15] T. Villmann, R. Der, M. Herrmann, and T. M. Martinetz, "Topology preservation in self-organizing feature maps: Exact definition and measurement," *IEEE Trans. Neural Netw.*, vol. 8, no. 2, pp. 256–266, Mar. 1997.

[16] H. Ritter, "Neural computation and self-organizing maps: An introduction," in *Neural Computation and Self-Organizing Maps; An Introduction*. Reading, MA, USA: Addison-Wesley, 1992.

[17] A. Rauber, D. Merkl, and M. Dittenbach, "The growing hierarchical self-organizing map: Exploratory analysis of high-dimensional data," *IEEE Trans. Neural Netw.*, vol. 13, no. 6, pp. 1331–1341, Nov. 2002.

[18] S.-Y. Huang and Y. Huang, "Network forensic analysis using growing hierarchical SOM," in *Proc. IEEE Int. Conf. Data Mining Workshops*, Dec. 2014, pp. 536–543.

[19] D. Ippoliti and X. Zhou, "A-GHSOM: An adaptive growing hierarchical self organizing map for network anomaly detection," *J. Parallel Distrib. Comput.*, vol. 72, no. 12, pp. 1576–1590, 2012.

[20] D. Alahakoon, S. K. Halgamuge, and B. Srinivasan, "Dynamic self-organizing maps with controlled growth for knowledge discovery," *IEEE Trans. Neural Netw.*, vol. 11, no. 3, pp. 601–614, May 2000.

[21] E. J. Palomo, E. Domínguez, R. M. Luque, and J. Muñoz, "Network security using growing hierarchical self-organizing maps," in *Adaptive and Natural Computing Algorithms*. Berlin, Germany: Springer, 2009.

[22] Y. Yang, D. Jiang, and X. Min, "Using improved GHSOM for intrusion detection," *J. Inf. Assurance Secur.*, vol. 5, pp. 232–239, May 2010.

[23] A. L. Hsu, I. Saeed, and S. K. Halgamuge, "Dynamic self-organising maps: Theory, methods and applications," in *Foundations of Computational, Intelligence*. Berlin, Germany: Springer, 2009.

[24] Y. Yu and D. Alahakoon, "Batch implementation of growing self-organizing map," in *Proc. Int. Conf. Comput. Inteligence Modelling Control Automat. Int. Conf. Intell. Agents Web Technol. Int. Commerce*, 2006, p. 162.

[25] M. Vasighi and H. Amini, "A directed batch growing approach to enhance the topology preservation of self-organizing map," *Appl. Soft Comput.*, vol. 55, pp. 424–435, Jun. 2017.

[26] M. Zaki, *Introduction to Data Mining*. Beijing, China: China Machine Press, 2010.

[27] R. Lippmann, J. W. Haines, D. J. Fried, J. Korba, and K. Das, "The 1999 DARPA off-line intrusion detection evaluation," *Comput. Netw.*, vol. 34, no. 4, pp. 579–595, 2000.

[28] *KDD CUP 99 Dataset*. Accessed: Oct. 28, 1999. [Online]. Available: http://kdd.ics.uci.edu/databases/kddcup99/task.html

[29] *KDD99 Dataset*. Accessed: Oct. 28, 1999. [Online]. Available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[30] *CICIDS Dataset*. Accessed: Aug. 19. 2018. [Online]. Available: https://www.unb.ca/cic/datasets/flowmeter.html

[31] H. G. Kayacik, A. N. Zincir-Heywood, and M. I. Heywood, "A hierarchical SOM-based intrusion detection system," *Eng. Appl. Artif. Intell.*, vol. 20, no. 4, pp. 439–451, 2007.

XIAOFEI QU received the master's degree from the Beijing University of Posts and Telecommunications, in 2011. She is currently pursuing the Ph.D. degree with the College of Command and Control Engineering, Army Engineering University of PLA. Her research interests include cybersecurity, artificial intelligence, computer immunity, and game theory.

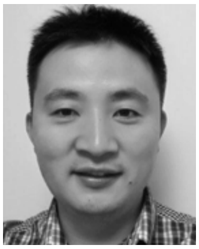
LIN YANG received the Ph.D. degree from the National University of Defense Technology, in 1998. He is currently a Researcher fellow with the National Key Laboratory of Science and Technology on Information System Security, Institute of Systems Engineering, AMS, Beijing. His current research interests include cybersecurity and component-based design.

KAI GUO received the Ph.D. degree from the College of Advanced Interdisciplinary Studies, National University of Defence Technology, in 2018. He is currently an Engineer with the National Key Laboratory of Science and Technology on Information System Security, Institute of Systems Engineering, AMS, Beijing. His current research interests include silicon nanophotonics, quantum communication, and optical neuron networks.

LINRU MA received the Ph.D. degree from the National University of Defense Technology, in 2007. She is currently a Senior Engineer with the National Key Laboratory of Science and Technology on Information System Security, Institute of Systems Engineering, AMS, Beijing. Her current research interests include intrusion detection systems and network security.

TAO FENG received the Ph.D. degree from Tsinghua University, in 2016. He is currently a Senior Engineer with the National Key Laboratory of Science and Technology on Information System Security, Institute of Systems Engineering, AMS, Beijing. His current research interests include network architecture, SDN, network management, and network security.

SHUANGYIN REN received the Ph.D. degree from the National University of Defense Technology, in 2017. He is currently an Engineer with the National Key Laboratory of Science and Technology on Information System Security, Institute of Systems Engineering, AMS, Beijing. His current research interest includes the security of unmanned aerial systems.



MENG SUN received the Ph.D. degree from the Department of Electrical Engineering, Katholieke University Leuven, in 2012. He is currently an Associate Professor with the Laboratory of Intelligent Information Processing, Army Engineering University of PLA, China. His research interests include speech processing, unsupervised/semi-supervised machine learning, and sequential pattern recognition.