Am I Rare? an Intelligent Summarization Approach for Identifying Hidden Anomalies

Samira Ghodratnama[1](✉), Mehrdad Zakershahrak[2],
and Fariborz Sobhanmanesh[1]

[1] Macquarie University, Sydney, Australia
{samira.ghodratnama,fariborz.sobhanmanesh}@mq.edu.au
[2] Arizona State University, Arizona, USA
mehrdad@asu.edu

Abstract. Monitoring network traffic data to detect any hidden patterns of anomalies is a challenging and time-consuming task which requires high computing resources. To this end, an appropriate summarization technique is of great importance, where it can be a substitute for the original data. However, the summarized data is under the threat of removing anomalies. Therefore, it is vital to create a summary that can reflect the same pattern as the original data. Therefore, in this paper, we propose an INtelligent Summarization approach for IDENTifying hidden anomalies, called *INSIDENT*. The proposed approach guarantees to keep the original data distribution in summarized data. Our approach is a clustering-based algorithm that dynamically maps original feature space to a new feature space by locally weighting features in each cluster. Therefore, in new feature space, similar samples are closer, and consequently, outliers are more detectable. Besides, selecting representatives based on cluster size keeps the same distribution as the original data in summarized data. *INSIDENT* can be used both as the preprocess approach before performing anomaly detection algorithms and anomaly detection algorithm. The experimental results on benchmark datasets prove a summary of the data can be a substitute for original data in the anomaly detection task.

Keywords: Anomaly detection · Summarization · Network data · Clustering · Classification

1 Introduction

Monitoring the fast and large volume of Internet traffic data that is being generated is paramount since they may have instances of anomalous network traffic, which makes the system vulnerable. However, detecting anomalies when we face big data is computationally expensive and still an open challenge. To this end, summarization is a practical approach that produces a condensed version of the

H. Hacid et al. (Eds.): ICSOC 2020 Workshops, LNCS 12632, pp. 309–323, 2021.
https://doi.org/10.1007/978-3-030-76352-7_31

original data. Therefore, a summary of the network traffic data helps network managers quickly assess what is happening in the network. For instance, the summary should still give insight into most visited websites, frequently used applications, and incoming traffic patterns. In [23] authors defined three scenarios in which summarization can help in traffic data, including: Summarizing network traffic can give an overview of what is going on in the network to the administrator. Summarized network traffic can be used as input to anomaly detection algorithms to reduce the cost. A summary of intrusion detection alarms facilitates the administrator's duty. In all mentioned scenarios, a concise representation of the data helps both the administrator and the analysis algorithms.

Different data summarization techniques are designed for other applications such as transactional data or stream data [1], which can be applied to traffic data. However, they have some drawbacks to be used for anomaly detection purposes, including:

- Clustering is the most used approach for summarization, where centers are considered as the summarized data. The problem is that the centroids may not be a part of the original data.
- Detecting frequent itemsets is another approach which only captures frequent items in the summaries. Therefore, they ignore or leave out anomalies that may be infrequent. Consequently, anomaly detection techniques do not perform well on summaries as they do not contain any anomalies.
- Semantics-based techniques do not keep the same samples in the summarized data.
- Statistical based techniques such as sampling do not guarantee the representation of anomalies in summary since they use a sampling-based summarization technique.

Therefore, not all summarization approaches are proper for anomaly detection purposes. Consequently, there is a need for an efficient network traffic summarization technique so that the summary more closely resembles the original network traffic In this context, summarization aims to create a summary from original data that includes interesting patterns, especially anomalies, and normal data for further analysis.

This paper proposes an intelligent summarization approach suitable for anomaly detection on network traffic datasets, which guarantees the preservation of original data distribution. We investigate the adaptation of clustering and KNN algorithms to create a summary. The proposed algorithm is used in two scenarios: i) as the preprocess approach for performing anomaly detection, ii) to detect anomalies in supervised problems as it reveals the hidden structure of data. The proposed summarization technique can also be adapted to other domains where big data requires being minded for interesting and relevant information. The rest of this paper is organized as follows. Section 2 discusses the state-of-the-art methods. Section 3 presents the proposed method, and Sect. 4 explains the experimental results and justifies the obtained results. Finally, Sect. 5 concludes the paper and discusses future work.

2 Related Work

Summarization has been widely explored in many domains and applications, using a variety of techniques [8,12,33]. When data size increases, the anomaly detection techniques perform poorly due to increasing false alarms and computational cost. Detecting anomalies from a summary could address these issues. However, existing summarization techniques cannot accurately represent the rare anomalies present in the dataset. In this section, we will present related work on traffic data summarization, along with anomaly detection techniques. It is worth mentioning although the general goal is to represent an input dataset in a condensed version, there is no definition of a good summary since each application requires a unique technique. For anomaly detection purposes, a good summary should be representative of all samples in the original dataset.

2.1 Network Analysis Tools

Different network analysis tools summarize network traffic data, such as Traffic Flow Analysis Tool, Flow-tools, Network Visualization Tools, and Network Monitoring Tools [2]. They produce a graphical report using different measurements, such as network bandwidth or latency. However, they only characterize and aggregate traffic instances based on a single attribute, such as the source/destination address or protocol. As a result, they are suitable to extract insights, not for further processing tasks such as anomaly detection. Besides, the objective of a summary is to provide an accurate report of the network's traffic patterns. Consequently, the summarization technique should identify traffic patterns based on arbitrary combinations of attributes efficiently.

2.2 Statistical Approaches

Statistical approaches aim to estimate the statistical distribution of data that could approximate the data set pattern. Sampling is a common technique in this category where a sample is a subset of the dataset. There are different kinds of sampling in practice, including i) simple random sampling, ii) stratified random sampling, iii) systematic sampling, iv) cluster random sampling, and v) multi-stage random sampling [15,17]. However, summarized data using sampling is under the threat of removing anomalies. To solve this problem, in a recent work [2], the author proposed a sampling-based summarization technique, called SUCh, which integrated the concept of sampling using the modified Chernoff bound to include anomalous instances in summary. *SUCh* is computationally effective than the existing techniques and also performs better in identifying rare anomalies. However, an essential aspect of the summarization is representing all different types of traffic behavior. Although *SUCh* ensures the presence of anomalies, it ignores other types of traffic as they focus only on anomalous data.

2.3 Machine Leaning Approaches

Supervised and unsupervised learning techniques are two widely used knowledge discovery techniques. Two common machine learning algorithms used in summarizing network traffic data are *frequent itemsets* and *clustering.* Frequent itemsets are a set of items that appears more frequently than the rest of the samples. Different algorithms are used to detect frequent itemsets [14]. However, they are proper for detecting frequent items, not rare anomalies. Two main clustering-based algorithms for network traffic data summarization include centroid-based and feature-wise intersectin clustering algorithms. In a centroid-based summarization, after clustering samples, centroids are used to form the summary. Different variations of the k-means algorithm are widely used due to its simplicity, which can handle high-dimensional data [20,37]. In a feature-wise intersection-based summarization, the summary is created from each cluster using the feature-wise intersection of the data instances after clustering [14,23]. Consequently, summaries from all the clusters are combined to produce the final summary. This approach is best fitted for datasets with identical attribute values and, therefore, not suitable for detecting rare anomalies.

2.4 Semantic-Based Approaches

Semantic-based approaches are not suitable for anomaly detection since they do not produce a summary, which is part of the original data. Examples are linguistic summaries, which are based on the fuzzy. These approaches produce natural language expressions that describe important facts about the given data to enhance the human understanding of the network traffic summaries [31]. Attribute Oriented Induction (AOI) is another semantic-based approach aims to describe data in a concise and general manner [21]. AOI is a generalization process that abstracts a large dataset from a low conceptual level to a relatively higher conceptual level. Other semantic-based approaches include Fascicles [24], which relies on an extended form of association rules and perform lossy semantic compression. SPARTAN is another semantic-based summarization technique [10], which generalizes the fascicles approach.

2.5 Anomaly Detection Techniques

Anomaly detection is an important data analysis task that detects anomalous or abnormal data from a given dataset. Anomalies are patterns in data that do not follow the well-defined characteristic of typical patterns. Anomalies are important because they indicate significant but rare events that may have a detrimental impact on the system. Therefore, they require prompt critical actions to be taken in a wide range of application domains. An anomaly can be categorized in the following ways [3].

– Point anomaly: When a data instance deviates from the normal pattern of the dataset, it can be considered a point anomaly.

Table 1. Example of network traffic samples.

Source IP	Source port	Destination IP	Destination port	Protocol
192.168.5.10	1234	192.168.1.1	80	TCP
192.168.5.12	4565	192.168.1.2	20	TCP
192.168.5.10	20	192.168.28.80	119	HTTP
192.168.5.10	70	192.168.1.1	50	TCP
211.204.12.10	31	192.168.28.80	119	HTTP
192.168.5.1	3214	192.168.1.2	86	TCP

- Contextual anomaly: When a data instance behaves anomalously in a particular context, it is called a contextual or conditional anomaly.
- Collective anomaly: When a collection of similar data instances behave anomalously compared to the entire dataset, the group of data instances is called a collective anomaly.

Different supervised, unsupervised, and semi-supervised approaches have been proposed for this purpose. These techniques, including classification based network anomaly detection such as support vector machine [11], Bayesian network [27], neural network [30], and rule-based approaches [38]. Statistical anomaly detection techniques, including mixture model [16], signal processing technique [36], and principal component analysis (PCA) [34]. Other category includes information theory-based and clustering-based [1]. The proposed summarization approach is a general approach used in two scenarios: i) as the preprocessing approach where results are used as the input for anomaly detection algorithm, and ii) as an anomaly detection technique in a supervised setting discussed in the next section.

3 The Proposed Approach (INSIDENT)

This section discusses our proposed methodology. At first, we define the problem and then discuss our algorithm.

3.1 Problem Definition

In this paper, x_i is a sample vector and $X = [x_1, x_2, ..., x_N]$ is traffic data consists of N sample where $x_i \in R^d$ which d denotes the number of features. K is the number of clusters, and cluster centroids are denoted by c. $x_=$ is the closest similar sample to x, and $x_{\neq}$ is the closest different sample. An example of network traffic data with few attributes is reported in Table 1. The goal is to find a cluster of similar samples and find representatives for each cluster as the summary S where they keep the same distribution but less in size.

3.2 Methodology

Previous approaches used different clustering or sampling algorithms to summarize data. However, there is no guarantee that the summarized data has the same distribution as the original data, and therefore as the substitute for the original data. In this paper, we investigate the adaptation of clustering and the KNN algorithm to understand the data's underlying structure. In our previous work, this structure was used in the context of multi-document summarization [19] and image retrieval [18], demonstrating promising results. For this reason, the error rate of the nearest neighbor classifier in each cluster is minimized by locally weighting features in each cluster. INSIDENT transforms the feature space into a new feature space by weighting features separately in each cluster, where outliers are recognized easier in the new feature space. To this end, the weighted Euclidean distance is used. In our problem, these weights are arranged in a $d \times K$ weight matrix $W = \{w_{ij}, 1 \leq i \leq d, 1 \leq j \leq K\}$ where d is the number of features, and is K the number of clusters. To be more specific, for each cluster we have a vector of weights corresponding to each feature which are representative of the importance of each feature in each cluster. Our objective function is designed to minimize the error of 1NN in each cluster by regulating weights of each feature, and consequently cluster centers. To estimate the error of 1NN the following approximation function defined in [29] is used:

$$J(\mathbf{W}) = \frac{1}{N} \sum_{s \in XS} S_\beta(\frac{d_w(x, x_=)}{d_w(x, x_{\neq})}) \tag{1}$$

where the sample $x_=$ is the nearest similar sample, and the sample $x_{\neq}$ is the closest different sample to the input sample x. Respectively d_w is the weighted Euclidean distance, and S_β is the sigmoid function, defined as:

$$S_\beta(z) = (\frac{1}{1 + e^{\beta(1-z)}}) \tag{2}$$

The objective function of K-means, which aims to minimize the errors of each cluster, is defined as:

$$J(\mathbf{W}, \mathbf{C}) = \sum_{k=1}^{K} \sum_{i=1}^{|N_K|} d^2_{W_K}(x_i, c_K) \tag{3}$$

Thus, the overall objective function is defined as:

$$J(\mathbf{W}, \mathbf{C}) = (\sum_{k=1}^{K} \sum_{i=1}^{|N_K|} d^2_W(s_i, c_K) + \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{|N_K|} S_\beta(\frac{d_w(x, x =)}{d_w(x, x_{\neq})})) \tag{4}$$

where the first term is the objective function of K-means, and the second term is the summation of the classification errors over the K clusters.

Two parameters are optimized in this objective function. The first is the weights matrix. The feature-dependent weights associated with the sample are trained to make it closer to x, while making the sample $x_{\neq}$ further from x. Then, the cluster centroid update is based on the learned weighted distance. Since this function is differentiable, we can analytically use gradient descent for estimating the matrix W, guaranteeing convergence. The iterative optimization of a learning parameter like w is given below.

$$W^{t+1} = W^{t} - \alpha(\frac{J(\mathbf{W}, \mathbf{C})}{\delta(W)}) \tag{5}$$

To simplify the formula, the function $R(x)$ is defined [29] as:

$$R_w(x_i) = (\frac{d_w(x_i, x_{i,=})}{d_w(x_i, x_{i,\neq})}) \tag{6}$$

The partial derivative of $J(W, C)$ with respect to W is calculated by:

$$\frac{\delta J(\mathbf{W}, \mathbf{K})}{\delta W_K} \cong \sum_{i=1}^{|N_K|} 2W_K \odot (x_i - C_K)^2 + \frac{1}{N} \sum_{i=1}^{|N_K|} S_{\beta}^{'}(R(x_i)) \frac{\delta R(x_i)}{\delta W_k} \tag{7}$$

where $\odot$ is the inner product and $\frac{\delta R(x_i)}{\delta W_K}$ is :

$$\frac{\delta R(s_i)}{\delta W_K} = \frac{1}{d_{W_K}^2}(x_i, x_{i,\neq})(\frac{1}{R(x_i)} W_K \odot (x_i - x_{i,=})^2 - R(x_i) W_K \odot (x_i - x_{i,\neq})^2) \tag{8}$$

The derivative of $S_{\beta}(z)$ is defined as:

$$\begin{aligned} S_{\beta}(z)' &= \frac{\delta S_{\beta}(z)}{\delta z} \\ &= \frac{\beta e^{\beta(1-z)}}{(1 + e^{\beta(1-z)})^2} \end{aligned} \tag{9}$$

The partial derivative of $J(\mathbf{W}, \mathbf{C})$ with respect to C is calculated as:

$$\frac{J(\mathbf{W}, \mathbf{C})}{\delta C_k} \cong \sum_{i=1}^{|N_k|} -2W_k^2 \odot (x_i - C_k) \tag{10}$$

Since we need to optimize the weight of features for each cluster's samples, along with the center of clusters, we first update W in each cluster, and then we update C (center of clusters). The INSIDENT algorithm is depicted in Algorithm 1 for more clarification. Since the algorithm performs in an iterative process using gradient descent, the simplest clustering (k-means) and (KNN) algorithms are used for efficiency. However, K-means is one of the most reliable and most widely used clustering algorithms. Besides, the K-nearest neighbor (NN) has been successfully used in many pattern-recognition applications [9]. Similar samples are close to each other in new feature space, making a point, and contextual type anomalies easily detectable. In the case of collective anomalies, we select the number of each cluster's representative based on its size to keep the distribution the same as the original data.

Algorithm 1. INSIDENT

```
Input: Traffic Data X, learning rate $\gamma$ and $\alpha$.
Output: Summary ($S$).

procedure INSIDENT.
  while $iter < MaxIterations$ do
    Clusters (C) ← K-means(X)
    for each clusters c in C do
      for each sample x in c do
        $x_{=} \leftarrow findSimilarCloseSample()$
        $x_{\neq} \leftarrow findDifferentCloseSample()$
        $W^{iter+1} = W^{iter} - \gamma \frac{\delta J(W)}{W}$
      end for
    end for
    Update Clusters
  end while
  return Summary(S)
```

4 Experiments and Evaluation

In this section, the dataset, the evaluation method, and the performance of INSIDENT are explained and compared with existing state-of-the-art approaches.

4.1 Data Set

Experiments on six benchmark datasets are performed. The details of this dataset and the distribution of normal and anomalous samples in each dataset are reported in Table 2. KDD1999 contains collective anomalies were the other five datasets contain only rare anomalies. These rare anomalous datasets are from SCADA network, including real SCADA (WTP), simulated anomalies (Sim1 and Sim2), and injected anomalies (MI and MO).

4.2 Evaluation Metrics

To evaluate network traffic summary, we explain two widely used summary evaluation metrics including *conciseness*, and *information loss* [5].

- Conciseness: The size of the summary influences the quality of the summary. At the same time, it is important to create a summary that can reflect the underlying data patterns. Conciseness is defined as the ratio of input dataset size (N) and the summarized dataset size (S) defined as:

$$Conciseness = \frac{N}{S} \tag{11}$$

- Information Loss: A general metric used to describe the amount of information lost from the original dataset due to the summarization. Loss is defined as the ratio of the number of samples not present by samples present in summary defines as:

$$InformationLoss = \frac{L}{T} \tag{12}$$

where T is the number of unique samples represented by the summary, and L defines the number of samples not present in the summary.

Table 2. Dataset sescription.

Dataset	Sample number	Normal Percentage	Anomalies percentage
KDD1999	494020	19.69	80.310
WTP	527	97.34	2.66
MI	4690	97.86	2.14
MO	4690	98.76	1.24
Sim1	10501	99.02	0.98
Sim2	10501	99.04	0.96

Besides, to evaluate the performance of the anomaly detection algorithms used in supervised approaches, three measures, including accuracy, recall, and F1 discussed below, are used. Before we define these measure, four values included in the confusion needs to be discussed [3].

- True Positive (TP): Number of anomalies correctly identified as anomalous.
- False Positive (FP): Number of normal data incorrectly identified anomaly.
- True Negative (TN): Number of normal data correctly identified as normal.
- False Negative (FN): Number of anomalies incorrectly identified as normal.

Based on the above definitions, we define the evaluation metrics.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{15}$$

4.3 Result Analysis

In this section, we discuss the performance evaluation of the existing summarization methods compared to INSIDENT, along with the anomaly detection result.

Table 3. Real SCADA dataset (WTP) result.

Model	WTP-Recall	WTP-Accuracy	WTP-F1
KNN	85.71	97.39	85.71
LOF	78.57	97.38	78.57
COF	57.14	97.35	57.14
LOCI	85.71	97.39	85.71
LoOP	42.85	97.33	42.85
INFLO	57.14	97.35	57.14
CBLOF	92.85	97.40	92.85
LDCOF	85.71	97.39	85.71
CMGOS	57.14	97.35	57.14
HBOS	28.57	97.32	28.57
LIBSVM	85.71	97.39	85.71
INSIDENT	94.87	97.91	94.87

Anomaly Detection Evaluation. This section contains the performance analysis of anomaly detection techniques. The baseline algorithms include Nearest Neighbor-based algorithms (K-NN [32], LOF [13], COF [35], LOCI [28], LoOP [26], INFLO [25]), clustering-based approach(CBLOF [22], LDCOF [6], CMGOS [6]), and statistical appraoches (HBOS and LIBSVM [7]). These approaches are compared with INSIDENT on different variations of the SCADA dataset, including WTP, MI, MO, Sim1, and Sim2, where their values are reported by [3]. Results are reported respectively in Table 3, Table 4, and Table 5.

From Table 3, it can be seen that for the real SCADA dataset(WTP), INSIDENT has higher values. Then the clustering-based anomaly detection technique, CBLOF, performs best, and third, the nearest-neighbor-based approach attains the best performance. It is an expected result showing the combination of clustering and KNN can perform better. Statistical based approach HBOS dis not perform well. Table 4 displays the results on simulated datasets (Sim1 and Sim2). LIBSVM has better recall than others, and INCIDENT performs as the second best. Clustering-based approaches are not well suited for the simulated datasets. For the datasets with injected anomalies (MI, MO), INCIDENT, along with clustering-based approaches, are the best considering the evaluation measures. Nearest neighbor-based approaches are the next best. It is interesting to observe that the Recall and F1 values are identical for all the anomaly detection techniques. The reason is that since the top N anomalies detected by the techniques match the actual N number of anomalies in the dataset, the Recall, and F1 scores are always the same.

Network Traffic Summarization Evaluation. For summarization evaluation, the KDD dataset is used. Summarization size, which defines conciseness, is

Table 4. Simulated SCADA datasets result(Sim1 and Sim2).

Model	Sim1Recall	Sim1Accuracy	Sim1F1	Sim2Recall	Sim2Accuracy	Sim2F1
KNN	64.7	99.03	64.7	63	99.05	63
LOF	0	99.01	0	0	99.03	0
COF	0	99.01	0	2	99.03	2
LOCI	0	99.01	0	0	99.03	0
LoOP	0.98	99.01	0.98	0	99.03	0
INFLO	0	99.01	0	0	99.03	0
CBLOF	0	99.01	0	0	99.03	0
LDCOF	0	99.01	0	0	99.03	0
CMGOS	18.62	99.02	18.62	97	99.05	97
HBOS	30.39	99.02	30.39	27	99.04	6
LIBSVM	74.50	99.03	74.50	68	99.05	68
INSIDENT	72.13	99.07	72.13	78.21	99.05	78.21

Table 5. Simulated SCADA datasets with Injected Anomalies result (MI and MO).

Model	MI-Recall	MI-Accuracy	MI-F1	MO-Recall	MO-Accuracy	MO-F1
KNN	96	97.09	96	91.37	98.77	91.37
LOF	38.33	97.43	38.33	55.17	98.76	55.17
COF	9	97.82	9	25.86	98.75	25.86
LOCI	91	97.9	91	84.48	98.77	84.48
LoOP	10	97.83	10	27.58	98.75	27.58
INFLO	12	97.83	12	43.1 0	98.76	43.10
CBLOF	24	97.84	24	63.79	98.76	63.79
LDCOF	100	97.91	100	63.79	98.76	63.79
CMGOS	100	97.91	100	50	98.76	50
HBOS	98	97.91	98	65.51	98.76	65.51
LIBSVM	86	97.9	86	91.37	98.77	91.37
INSIDENT	100	98.76	100	94.21	99.04	94.21

considered as a constraint in summarization algorithms. When the summary is small, it has maximum information loss. On the other hand, when conciseness is small, the summary contains the whole dataset has no information loss. Therefore, information loss and conciseness are orthogonal parameters. Our experiments used five different summary sizes, and then information loss was measured for each summary size. In practice, the network manager/analyst decides the summary size based on the network. The results are compared with NTS and FIB approaches [4]. Since our algorithm is based on k-means, we test three times with different initial points for each summary size. Results are depicted in Fig. 1. Besides, the percentage of anomalies compared with SUCh [2] is reported in Table 6 proving that INSIDENT well-preserved the percentage of anomalies in generated summaries.

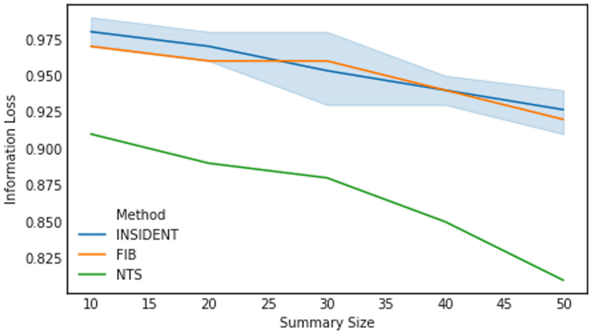

Fig. 1. The result of comparing information loss based on different summary size.

Table 6. Comparing the distribution of anomalies in summaries and original data.

Dataset	Original data	SUCh Alg	INSIDENT
WTP	2.66	N/A	2.33
MI	2.14	2.61	2.76
MO	1.24	1.46	1.52
Sim1	0.98	1.04	1.11
Sim2	0.96	0.94	1.01

5 Conclusion and Future Work

Monitoring network traffic data to detect any hidden patterns of anomalies is a challenging and time-consuming task which requires high computing resources. Therefore, in this paper, we proposed an INtelligent Summarization approach for IDENTifying hidden anomalies, called *INSIDENT*. In data summarization, it is always a dilemma to claim the best summary. The proposed approach claim is to guarantee to keep the original data distribution in summarized data. The INSIDENT's backbone is the clustering and KNN algorithm that dynamically maps original feature space to a new feature space by locally weighting features in each cluster. The experimental results proved that the proposed approach helps keep the distribution the same as the original data, consequently making anomaly detection easier. In future work, we aim to focus on real-time network traffic summarization.

Acknowledgement. We acknowledge the AI-enabled Processes (AIP) Research Centre (https://aip-research-center.github.io/) for funding this research. We also acknowledge Macquarie University for funding this project through IMQRES scholarship.

References

1. Ahmed, M.: Data summarization: a survey. Knowl. Inf. Syst. **58**(2), 249–273 (2019)
2. Ahmed, M.: Intelligent big data summarization for rare anomaly detection. IEEE Access **7**, 68669–68677 (2019)
3. Ahmed, M., Anwar, A., Mahmood, A.N., Shah, Z., Maher, M.J.: An investigation of performance analysis of anomaly detection techniques for big data in scada systems. EAI Endorsed Trans. Indust. Netw. Intell. Syst. **2**(3), e5 (2015)
4. Ahmed, M., Mahmood, A.N., Maher, M.J.: A novel approach for network traffic summarization. In: Jung, J.J., Badica, C., Kiss, A. (eds.) INFOSCALE 2014. LNICST, vol. 139, pp. 51–60. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16868-5_5
5. Ahmed, M., Mahmood, A.N., Maher, M.J.: An efficient technique for network traffic summarization using multiview clustering and statistical sampling. EAI Endorsed Trans. Scalable Inf. Syst. **2**(5), (2015)
6. Amer, M., Goldstein, M.: Nearest-neighbor and clustering based anomaly detection algorithms for rapidminer. In: Proceedings of the 3rd RapidMiner Community Meeting and Conference (RCOMM 2012), pp. 1–12 (2012)
7. Amer, M., Goldstein, M., Abdennadher, S.: Enhancing one-class support vector machines for unsupervised anomaly detection. In: Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description, pp. 8–15 (2013)
8. Amouzgar, F., Beheshti, A., Ghodratnama, S., Benatallah, B., Yang, J., Sheng, Q.Z.: iSheets: a spreadsheet-based machine learning development platform for data-driven process analytics. In: Liu, X., et al. (eds.) ICSOC 2018. LNCS, vol. 11434, pp. 453–457. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-17642-6_43
9. Anava, O., Levy, K.: k*-nearest neighbors: From global to local. In: Advances in Neural Information Processing Systems, pp. 4916–4924 (2016)
10. Babu, S., Garofalakis, M., Rastogi, R.: Spartan: a model-based semantic compression system for massive data tables. ACM SIGMOD Rec. **30**(2), 283–294 (2001)
11. Balabine, I., Velednitsky, A.: Method and system for confident anomaly detection in computer network traffic, uS Patent 9,843,488, 12 December 2017
12. Beheshti, A., et al.: iProcess: enabling IoT platforms in data-driven knowledge-intensive processes. In: Weske, M., Montali, M., Weber, I., vom Brocke, J. (eds.) BPM 2018. LNBIP, vol. 329, pp. 108–126. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-98651-7_7
13. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: Lof: identifying density-based local outliers. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, pp. 93–104 (2000)
14. Chandola, V., Kumar, V.: Summarization-compressing data into an informative representation. Knowl. Inf. Syst. **12**(3), 355–378 (2007)
15. Cochran, W.G., William, G.: Sampling Techniques. Wiley, New York (1977)
16. Eskin, E.: Anomaly detection over noisy data using learned probability distributions (2000)
17. Ghodratnama, S., SadrAldini, M.: An innovative sampling method for massive data reduction in data mining. In: The 3rd Iran Data Mining Conference, Tehran (2009)
18. Ghodratnama, S., Abrishami Moghaddam, H.: Content-based image retrieval using feature weighting and C-means clustering in a multi-label classification framework. Pattern Anal. Appl. **24**(1), 1–10 (2020). https://doi.org/10.1007/s10044-020-00887-4

19. Ghodratnama, S., Beheshti, A., Zakershahrak, M., Sobhanmanesh, F.: Extractive document summarization based on dynamic feature space mapping. IEEE Access **8**, 139084–139095 (2020)
20. Ghodratnama, S., Boostani, R.: An efficient strategy to handle complex datasets having multimodal distribution. In: Sanayei, A., E. Rössler, O., Zelinka, I. (eds.) ISCS 2014: Interdisciplinary Symposium on Complex Systems. ECC, vol. 14, pp. 153–163. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-10759-2_17
21. Han, J., Fu, Y.: 16 exploration of the power of attribute-oriented induction in data mining. In: Advances in Know Ledge Discover and Data Mining, pp. 399–421. AAAI/'&I1T Press, Cambridge (1996)
22. He, Z., Xu, X., Deng, S.: Discovering cluster-based local outliers. Pattern Recogn. Lett. **24**(9–10), 1641–1650 (2003)
23. Hoplaros, D., Tari, Z., Khalil, I.: Data summarization for network traffic monitoring. J. Netw. Comput. Appl. **37**, 194–205 (2014)
24. Jagadish, H., Madar, J., Ng, R.T.: Semantic compression and pattern extraction with fascicles. VLDB **99**, 186–97 (1999)
25. Jin, W., Tung, A.K.H., Han, J., Wang, W.: Ranking outliers using symmetric neighborhood relationship. In: Ng, W.-K., Kitsuregawa, M., Li, J., Chang, K. (eds.) PAKDD 2006. LNCS (LNAI), vol. 3918, pp. 577–593. Springer, Heidelberg (2006). https://doi.org/10.1007/11731139_68
26. Kriegel, H.P., Kröger, P., Schubert, E., Zimek, A.: Loop: local outlier probabilities. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, pp. 1649–1652 (2009)
27. Kruegel, C., Mutz, D., Robertson, W., Valeur, F.: Bayesian event classification for intrusion detection. In: 19th Annual Computer Security Applications Conference. Proceedings. pp. 14–23. IEEE (2003)
28. Papadimitriou, S., Kitagawa, H., Gibbons, P.B., Faloutsos, C.: Loci: fast outlier detection using the local correlation integral. In: Proceedings 19th International Conference on Data Engineering (Cat. No. 03CH37405), pp. 315–326. IEEE (2003)
29. Paredes, R., Vidal, E.: Learning weighted metrics to minimize nearest-neighbor classification error. IEEE Trans. Pattern Anal. Mach. Intell. **28**(7), 1100–1110 (2006)
30. Poojitha, G., Kumar, K.N., Reddy, P.J.: Intrusion detection using artificial neural network. In: 2010 Second International conference on Computing, Communication and Networking Technologies, pp. 1–7. IEEE (2010)
31. Pouzols, F.M., Lopez, D.R., Barros, A.B.: Summarization and analysis of network traffic flow records. In: Mining and Control of Network Traffic by Computational Intelligence, pp. 147–189. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-18084-2_4
32. Ramaswamy, S., Rastogi, R., Shim, K.: Efficient algorithms for mining outliers from large data sets. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, pp. 427–438 (2000)
33. Schiliro, F., et al.: iCOP: IoT-enabled policing processes. In: Liu, X., et al. (eds.) ICSOC 2018. LNCS, vol. 11434, pp. 447–452. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-17642-6_42
34. Shyu, M.L., Chen, S.C., Sarinnapakorn, K., Chang, L.: A novel anomaly detection scheme based on principal component classifier. Miami Univ Coral Gables FL Dept of Electrical and Computer Engineering, Technical report (2003)
35. Tang, J., Chen, Z., Fu, A.W., Cheung, D.W.: Capabilities of outlier detection schemes in large datasets, framework and methodologies. Knowl. Inf. Syst. **11**(1), 45–84 (2007)

36. Thottan, M., Ji, C.: Anomaly detection in IP networks. IEEE Trans. Signal Process. **51**(8), 2191–2204 (2003)
37. Wendel, P., Ghanem, M., Guo, Y.: Scalable clustering on the data grid. In: 5th IEEE International Symposium Cluster Computing and the Grid (ccGrid) (2005)
38. Yang, Y., McLaughlin, K., Littler, T., Sezer, S., Wang, H.: Rule-based intrusion detection system for scada networks (2013)