EAI Endorsed Transactions
on Scalable Information Systems

Research Article ICST.ORG

An Efficient Technique for Network Traffic Summarization using Multiview Clustering and Statistical Sampling

Mohiuddin Ahmed[1,*], Abdun Naser Mahmood[1] and Michael J. Maher[1]

[1]School of Engineering and Information Technology, UNSW Canberra, Australia

Abstract

There is significant interest in the data mining and network management communities to efficiently analyse huge amounts of network traffic, given the amount of network traffic generated even in small networks. Summarization is a primary data mining task for generating a concise yet informative summary of the given data and it is a research challenge to create summary from network traffic data. Existing clustering based summarization techniques lack the ability to create a suitable summary for further data mining tasks such as anomaly detection and require the summary size as an external input. Additionally, for complex and high dimensional network traffic datasets, there is often no single clustering solution that explains the structure of the given data. In this paper, we investigate the use of multiview clustering to create a meaningful summary using original data instances from network traffic data in an efficient manner. We develop a mathematically sound approach to select the summary size using a sampling technique. We compare our proposed approach with regular clustering based summarization incorporating the summary size calculation method and random approach. We validate our proposed approach using the benchmark network traffic dataset and state-of-the-art summary evaluation metrics.

Keywords: Scalable Data Mining, Network Traffic Summarization, Multiview Clustering.

Received on 15 March 2015, accepted on 24 March 2015, published on 02 July 2015

doi:10.4108/sis.2.5.e4

1. Introduction

Summarization is considered a key knowledge discovery approach that produces a concise, yet informative version of the original dataset[1]. Clustering, which groups together similar data instances, is often used for summarization[2–6]. Among the large pool of clustering algorithms[7], *k-means*[8] clustering has been widely used since it is easy to implement and understand. The resulting cluster centroids are considered the summary of the original data. However, *k-means* introduces several problems in terms of summarizing a dataset. First, the *k-means* algorithm generates a centroid calculating the mean of the data instances within a cluster, which in general is not an actual member of the dataset. A summary produced using these centroids might be misleading. Another important problem for summarization using unsupervised techniques on unlabelled data is that the number of clusters is generally unknown. Importantly, traditional clustering techniques focus on producing only a single solution, even though multiple alternate clustering may exist. It is thus difficult for the user to validate whether the given solution is in fact appropriate, particularly if the dataset is large and high dimensional (such as network traffic), or if the user has limited knowledge about the clustering algorithm being used. In this case, it is highly desirable to provide alternative clustering solution, which is able to extract more information about the underlying pattern from different dimensions of the dataset.

Figure 1 shows the run time complexity of basic *k-means*[8] clustering algorithm on different sizes of data. It is clearly visible that, as data size increases the run time complexity also increases. As a result, knowledge discovery from large datasets becomes very inefficient. Consequently, summarization is a necessary step before performing data mining (such as anomaly detection from network traffic), which can expedite the process of knowledge discovery. Existing summarization techniques based on clustering do not produce a summary that can be used for anomaly detection[9, 10] because of not using original data

[*] Corresponding author. Email: mohiuddin.ahmed@student.unsw.edu.au

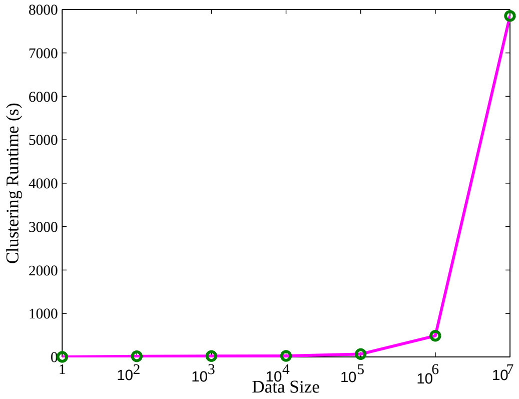


Figure 1. Run Time Complexity

instances. So, summarization using original data instances is important for further data mining and knowledge discovery process.

Rest of the paper contains the related works in Section 2, analysis of network traffic as complex data in Section 3. The theoretical background on multiview clustering and sampling techniques are discussed in Section 4. We discuss our proposed approach in Section 5 and experimental results in Section 6. Section 7 concludes the paper.

2. Related Works

In this Section, we briefly review the existing clustering based summarization approaches. Although, there are different approaches of data summarization, the clustering based summarization approaches fall within the scope of this paper.

Ha-Thuc et al[3] proposed a quality-threshold data summarization method modifying the *k-means* algorithm. The number of clusters is determined using the characteristics of dataset and a threshold. The algorithm partitions a dataset until the distortion or sum of squared error(2)(SSE) is less than a given threshold. It starts by finding the cluster centroids as *k-means* but next steps are executed only if the SSE is above the given threshold and the existing cluster is split. A new cluster centroid is introduced which is closer to the larger cluster centroid. This process is repeated until all the clusters' SSE exceeds the given threshold. They did not explain the method to choose the threshold and how the characteristics of datasets are analysed.

$$SSE = \sum_{i=1}^{k} \sum_{C_i} d(c_i, x)^2;\ x \in C_i \ and \ c_i : centroid \ of \ C_i$$

Patrick et al[4] proposed a distributed clustering framework, where the dataset is partitioned between several sites and output is mixture of gaussian models. Each distributed dataset is summarized using *k-means* algorithm and sent to a central site for global clustering.

Prodip et al[5] proposed an approach for clustering large datasets by randomly dividing the original data into disjoint subsets. The *k-means* algorithm is applied to summarize the dataset as well as to form ensemble using the centroids.

Wagstaff et al[2] presented a semi-supervised summarization approach for hyperspectral images. Hyperspectral images produce very large image in which each pixel is recorded at hundreds or thousands of different wavelengths. The ability to automatically generate summaries of these dataset enables important applications such as quickly browsing through a large image repository. However, this technique uses pre-specified knowledge to seed the initial centre for clustering which is not directly applicable in different domains.

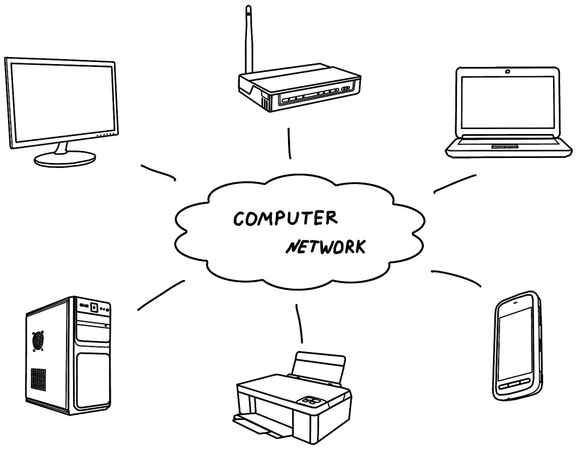


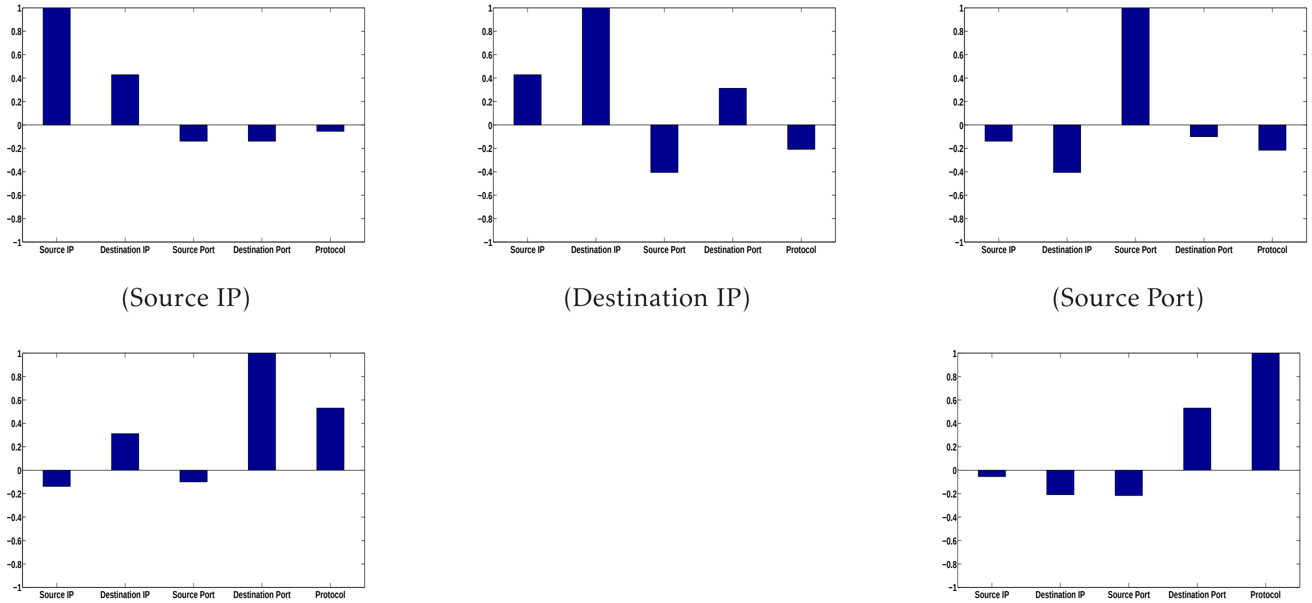
Figure 2. A common network architecture, adapted from Internet

3. Network Traffic as Complex Data

Network traffic can be considered as complex data where the straightforward data mining applications may not be effective. Data comes from more than one process. Each entry in the dataset is usually not only the outcome of a single characteristic; but also the combination different process. For example, in

Table 1. A sample of Network traffic

Source IP	Destination IP	Source Port	Destination Port	Protocol
192.168.5.10	192.168.12.1	20	80	TCP
192.168.5.12	192.168.11.1	21	80	TCP
192.168.12.28	192.168.1.11	22	21	TCP
192.168.5.22	192.168.12.20	23	443	ICMP
192.168.12.32	192.168.1.2	25	80	ICMP
192.168.5.26	192.168.1.1	53	21	UDP
88.34.224.2	192.168.1.2	110	443	UDP
88.36.226.2	192.168.1.1	119	25	TCP
88.34.226.12	192.168.1.2	143	21	TCP
192.168.5.10	192.168.1.1	443	80	TCP



(Source IP)

(Destination IP)

(Source Port)

(Destination Port)

(Protocol)

Figure 3. Correlation among the network traffic attributes

benchmark KDD Cup 99 dataset[11] has four different attribute types as follows:

- **Basic:** These features are corresponds to fields in the network packet headers and session timeouts. These are useful for detecting attacks which target protocol and service vulnerabilities.

- **Time:** These are important for identifying high volume fast rate DoS attacks based on the number of connections requests to the same destination host or service in a very short time frame.

- **Host:** These features store the number of connections to the same host, port or service by a destination host in the last 100 connections. They are useful for identifying Probe attacks.

- **Content:** These are based on domain knowledge. Important for detecting stealthy attacks (U2R, R2L) by observing the payload section of the packets.

The relationship among these attributes is not always significant. Also the network traffic as complex data has the following characteristics-

- A computer network or data network is a telecommunications network that allows computers to exchange data (Figure 2). In computer networks, networked computing devices pass data to each other along data connections. The connections (network links) between nodes are established using either cable media or wireless media. The best-known computer network is the Internet.

Network computer devices that originate, route and terminate the data are called network nodes. Nodes can include hosts such as personal computers, phones, servers as well as networking hardware. Two such devices are said to be networked together when one device is able to exchange information with the other device, whether or not they have a direct connection to each other. Computer networks support applications such as access to the World Wide Web, shared use of application and storage servers, printers, and fax machines, and use of email and instant messaging applications. Computer networks differ in the physical media used to transmit their signals, the communications protocols to organize network traffic, the network's size, topology and organizational intent. So, the network traffic data is suppose to be complex data based on these facts.

- Data has multiple causes. The relationship among the attributes and between each attribute are subtle and some attributes are predictive only for some records. For example, in Table 1, we display a sample of network traffic instances and Figure 3 shows the correlation (1) among the different network attributes. In equation(1), $\overline{x}, \overline{y}$ are the means of the variables X,Y. More over, network traffic dataset contains mixed attributes (such as numerical, categorical) and thus the relationship among the attributes are quite insignificant.

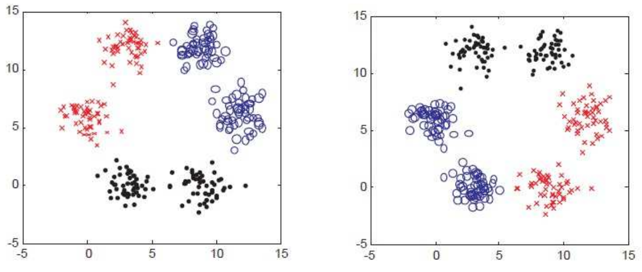
$$Correl(X,Y) = \frac{\sum(x-\overline{x})(y-\overline{y})}{\sqrt{\sum(x-\overline{x})^2 \sum(y-\overline{y})^2}} \qquad (1)$$

4. Theoretical Background

In this section, we provide a brief discussion on multiview clustering followed by sampling techniques and calculation of sample size.

4.1. Multiview Clustering

Exploratory data analysis aims to identify and generate multiple views of the structure within a dataset. Conventional clustering techniques[7], however, are designed to only provide a single grouping or clustering of a dataset. Data clustering is challenging, because there is no universal definition of it. Labelled data is generally not available that may help in the understanding of the underlying structure of the data, moreover, there is no unique similarity measure for differentiating clusters. Consequently, it is evident that there is no single clustering solution that explains the structure of a given dataset, especially if it is large (such as network traffic) and represented in a high dimensional space. This challenge has given rise to the recently emerging area of multiview clustering analysis[13], where the goal is to explore different partitions, in order to describe different grouping aspects for a given dataset.



(a) 1st clustering (b) 2nd clustering

Figure 4. Two alternative clusterings of the same dataset, each with 3 clusters. Point shapes show cluster membership, adapted from[12].

For example, consider the data given in Figure 4 and assume the number of clusters to be uncovered is 3. It is clear that both of the clustering solutions found in two Figures 4a and 4b are equally valid and logical, since they fit the data well and have the same clustering quality. It would be difficult to justify keeping only the first clustering, while omitting the second. We can also identify similar examples in real life applications. For example, in network traffic analysis, one can cluster traffic instances by their basic attributes; or content attributes, both clustering solutions are equally important and each could be used to provide a different interpretation of the data. In this paper, we study the application of multiview clustering on summarization of large and high dimensional data.

The multiview clustering problem can be formulated using the information theoretic concepts. For example, if we are given a dataset X with N points, such as $X = (x_1, x_2, ..., x_N)$, the task is to find a set of alternative clustering solutions, $C = (c_1, c_2...)$, where the clustering quality in terms of an objective function will be high and simultaneously the clustering solutions will be highly dissimilar to one another i.e. mutual information $I(c_1; c_2)$ is close to zero and $c_1 \neq c_2$.

4.2. Sampling Methods

The rationale behind integrating sampling methods into summarization is based on the need to construct a summary from original data instances. Sampling is a popular choice for reduction of input data in data mining and machine learning techniques. It has been applied in various aspects of network management, such as traffic measurement and reporting, traffic characterization and intrusion detection[14]. The principal advantages of sampling over the complete enumeration are the reduced cost and greater speed[18]. There are different kinds of sampling in practice. We briefly discuss three major categories of sampling and choose the appropriate sampling for our proposed summarization algorithm.

- **Simple Random Sampling:** Given the dataset size N and sample size n, simple random sampling chooses sample at random from the $\binom{N}{n}$ distinct possible samples, where no data instance is included more than once.
- **Stratified Random Sampling:** Here the dataset of size N is divided into non-overlapping subsets. These subsets are called *strata*. The sampling scheme selects a random element from each *strata* and produces a stratified sample. Basically, a simple random sampling is applied on each *strata* to have a stratified random sample.
- **Systematic Sampling:** This method of sampling is quite different from the other two. Here a data instance is sampled from the dataset, beginning from a specified starting point to the end, at equal intervals. To select a sample of n units, the first k unit is taken at random and every k^{th} unit afterwards. If the first random unit number is 2 and the value of k=5, then for sample of size 3, the sample units will be of number 2,7,12 respectively. Here the k is calculated as N/n, and for fractions it is rounded up.

For the network traffic summarization purpose, systematic sampling is advantageous over the other two because it involves choosing the data instances to be sampled at equal intervals. However, it can suffer from periodicity of the data but we address the issue by using clustering. It is an efficient sampling scheme for our purposed technique because, we think of choosing the samples from the clusters produced from the original dataset. Since, the clustering process groups together the similar data instances, the systematic sampling scheme will encompass the total cluster and be able to represent the cluster well. Additionally, this technique results better when the sample size is known and we plan to calculate the sample size of the produced cluster using statistical formula.

Sample size determination is a very important issue because a large sample size is a wastage of time and resource; on the other hand a smaller sample may lead to faulty results[18]. Since, we are interested in identifying the summary size automatically without user input, calculation of sample size from the produced clusters is a necessary step. In this scenario, sample mean and the original dataset mean is different and this difference is considered as an error. The margin of error E is the maximum difference between the sample mean and the actual dataset mean. According to statistical view point[19], this error E, can be defined using the following equation(2).

$$E = z_{\alpha/2} * \frac{\sigma}{\sqrt{n}} \tag{2}$$

- $z_{\alpha/2}$ is the critical value, α is the significance level. A positive z value corresponds to the area of $\alpha/2$ in the right tail of standard normal distribution;
- σ is the dataset standard deviation;
- n is the sample size;
- E is the margin of error, difference between the sample mean and original dataset's mean.

After rearranging the above formula, the sample size (summary size) can be calculated(3)

$$n = \left[\frac{z_{\alpha/2} * \sigma}{E}\right]^2 \tag{3}$$

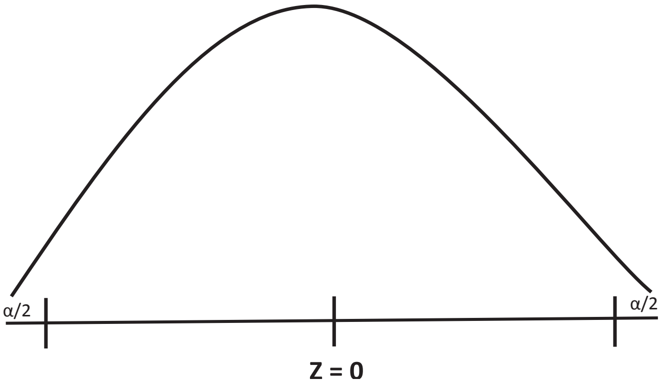


Figure 5. Critical value of the Standard Normal Distribution

5. Proposed Multiview Clustering based Network Traffic Summary

In this Section, we describe our proposed method for network traffic summarization. At first we present the regular clustering based technique for summarization incorporating the summary size calculation. Then we explain our algorithm based on multiview clustering.

In the **Algorithm 1**, regular clustering based network traffic summarization is presented, where we introduced the summary size calculation technique. The algorithm uses partitional clustering algorithm, *k-means* clustering. From each of the clusters produced, we calculate the sample/summary size using the statistical theories discussed in previous section. Once the summary size of the cluster is calculated, we take

ALGORITHM 1: ***RCNTS(Regular Clustering based Network Traffic Summarization)***

```
Input  : $D$, Dataset;
         $k$, number of clusters
Output: $S$, The summary of $D$
Begin
$\{C_1, C_2, ....C_n\} \leftarrow$ k-means $(D,k)$
for each cluster $C_i$, $i = 1{:}n$ do
    Calculate the summary size (3)
    $S_i$ = Systematic Sample of $C_i$
end
$S = \bigcup_1^n \ |S_i|$
End
```

a systematic sample (discussed in previous section 4.2) from the cluster. Finally, we merge all the systematic samples from all the clusters produced to create the final summary. This approach does not require any external input and overcomes the problems with the existing summarization techniques where the sample size and the representation of original data in the summary are the main constraints. However, for high dimensional and complex network traffic, regular clustering may not discover the underlying pattern effectively. Next, we present our proposed methodology based on multiview clustering.

ALGORITHM 2: ***Multiview Clustering of Network Traffic***

```
Input  : $D$, Dataset
Output: $C_n$, Clusters from Multiview Clustering
Begin
Classify the dataset according to different attribute types
For Network Traffic, $D_i = \{D_{Host}, D_{Basic}, D_{Time}, D_{Content}\}$
for $i = 1{:}4$ do
    $\{c_1, c_2\} \leftarrow$ k-means$(D_i, 2)$
end
$C_n = \bigcup_1^{2i} |c_i|$
End
```

ALGORITHM 3: ***MCNTS(Multiview Clustering based Network Traffic Summarization)***

```
Input  : $D$, Dataset
Output: $S$, The summary of $D$
Begin
$\{C_1, C_2, ....C_n\} \leftarrow$ Multiview Clustering $(D)$
for each cluster $C_i$, $i = 1{:}n$ do
    Calculate the summary size (3)
    $S_i$ = Systematic Sample of $C_i$
end
$S = \bigcup_1^n \ |S_i|$
End
```

In the **Algorithm 2**, the multiview clustering process for network traffic is given. As discussed in Section 3, the network traffic dataset considered in the scope of this paper has four kinds of attributes. So, for multiview clustering, we separated each of the attribute category and applied partitional clustering on them. The essence of multiview here is the application of partitional clustering on different types of attributes of the same dataset. It is expected that clustering different attribute types will result in different clustering solution. Also, we chose the number of cluster as two based on the assumption made by Portnoy et al[17], '*The majority of the network connections are normal traffic, only a small percentage of traffic are malicious*'. As a result, from four different types of attributes in the dataset, eight different clusters are produced. These clusters are the result of multiview clustering as well as applying partitional clustering on different types of attributes such as *Basic*, *Time*, *Host*, *Content*.

Next, we present our proposed approach for network traffic summarization in **Algorithm 3**. Once the multiview clustering (**Algorithm 2**) is applied and the resulting clusters are ready, from each of the clusters, we calculate the sample/summary size using the statistical theories discussed in previous Section 4.2. Once the summary size of the cluster is calculated, we take representative sample from the cluster having original data instances using systematic sampling. The representative sample has the minimum difference between the cluster centroid and mean of the selected sample. Finally, we merge all the representative samples from all the clustering solutions produced to create the final summary. Our proposed approach overcomes the problems with the existing summarization techniques where the sample size and the representation of original data in the summary are the main constraints. Additionally, the summary produced by our approach can be used as an input to anomaly detection techniques.

6. Experimental Analysis

For our experimental analysis, we used a variant of benchmark KDD cup 1999 dataset. NSL-KDD dataset[11] is a short form KDD cup 1999 which is derived from DARPA 1998 data from Licoln Laboratory at MIT. KDD 1999 is the most widely utilized dataset for the evaluation of the anomaly detection methods on network traffic. NSL-KDD is a dataset suggested to solve some of the inherent problems of the KDD 1999 dataset as mentioned in[16]. For sample/summary size calculation, we considered 95% significance level which corresponds to $\alpha = 0.05$ and $z_{\alpha/2} = 0.475$. In the table of standard normal distribution, an area of 0.475 corresponds to a z value of 1.96[20]. Thus, we used $z_{\alpha/2}$ = 1.96 in our experimental analysis and $E = 1$.

Table 2. A Summary of the dataset in Table 1

Source IP	Destination IP	Source Port	Destination Port	Protocol
192.168.12.32	192.168.1.1	53	443	TCP

6.1. Summarization Metrics

To simplify the understanding of a good network traffic summary, here in this section we explain the existing summary evaluation metrics [15]. Additionally, we also discuss two of our proposed metrics for network traffic summary evaluation hinted beforehand. First we discuss the existing metrics as follows

- **Conciseness:** Conciseness expresses how compact a summary is with respect to the original dataset. It is the ratio of input dataset size and the summarized dataset size. Then conciseness is represented in equation (4), where N is the number of data instances in input dataset and S denotes the number of data instances in summary. For example, the conciseness of the summary (Table 2) of the dataset in Table 1 is $\frac{10}{1} = 10$.

$$Conciseness = \frac{N}{S} \tag{4}$$

- **Information Loss:** A general metric used to describe the amount of information lost from the original dataset as a result of the summarization. Loss is defined as the ratios of number of cells not present by cells present in the summary [15]. Equation (5) states the information loss, where T_i is the number of unique cells represented by summary i and L_i defines the number of cells not present in summary i. For example, information loss of summary in Table 2, where $L_i = 25$ and T_i = 30. So, the information loss of summary in Table 2 will be $\frac{27}{30} = 0.90$.

$$Information\ Loss = \frac{L_i}{T_i} \tag{5}$$

- **Interestingness:** It is a new summarization metric proposed in [15] which focuses on the objective measures of interestingness with applicability to summarization, emphasizing diversity. Equation (6) defines the interestingness, where n_i states how many of the data instances in the original dataset are covered by the summary i, m is the number of individual summaries and N is the total number of data instances in original dataset. For example, the interestingness of summary in Table 2 is $\frac{(1(1-1))}{(10(10-1))} = 0$. Since, the original data has 10 data instances and the summary has one data instance, so the 1 tuple in summary represents 10 data points of the original data in Table 1.

$$Interestingness = \frac{\sum_{i=1}^{m} n_i(n_i - 1)}{N(N-1)} \tag{6}$$

- **Intelligibility:** This metric is used to measure how meaningful a summary is, based on the attributes present in the summary. Intelligibility is defined and displayed in equation (7), where m is the number of summary, a_i is the number of attributes present in the original dataset that is covered by summary i and q_i is the number of attributes present in the summary i. The intelligibility of summary in Table 2 is $\frac{1}{1}\frac{5}{5} =$ 1.0. The term attribute here in case of Table 1 means the *Source IP*, *Destination IP*, *Source Port*, *Destination Port*, *Protocol*.

$$Intelligibility = \frac{1}{m}\sum_{i=1}^{m}\frac{q_i}{a_i} \tag{7}$$

Summary size is considered as a constraint in summarization algorithms. Summary size, which defines conciseness is an important metric and has influence on the other metrics. When the summary is empty it has maximum information loss and when conciseness is 1 meaning the summary contains the whole dataset has no information loss.

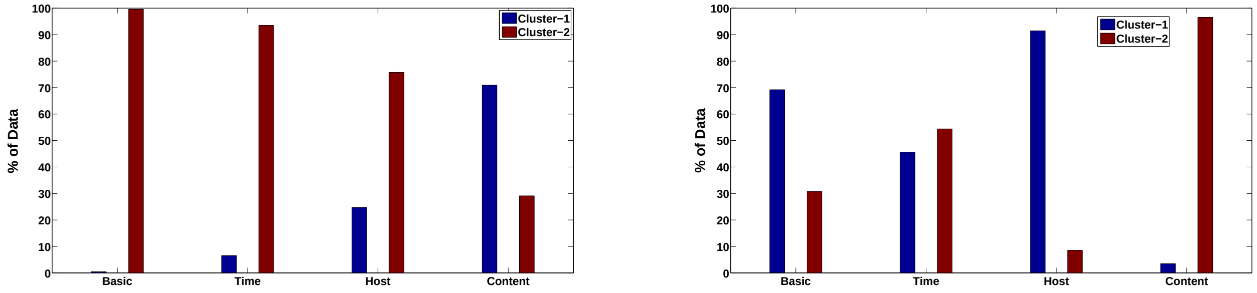
6.2. Discussion on Experimental Results

Table 3. Multiview Clustering Results

Dataset	Basic	Host	Time	Content
Cluster-1	32.47%	55.57%	24.76%	39.48%
Cluster-2	67.53%	44.43%	75.24%	60.52%

Table 4. Regular Clustering Results

Dataset	Number of Instances
Cluster-1	35.06%
Cluster-2	64.94%



(Normal Data Distribution)

(Anomalous Data Distribution)

Figure 6. Data distribution of multiview clustering

Table 5. Experimental results of the **MCNTS** algorithm

Techniques	Conciseness	Information Loss	Interestingness	Intelligibility
MCNTS	47.62	0.90	0.04	1.0
RCNTS	169.07	0.94	0.003	1.0
Random	169.07	0.94	0.003	1.0

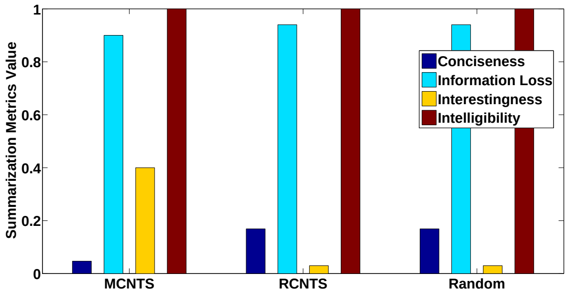


Figure 7. Comparison among the summarization techniques

Table 3 displays the clustering solutions over different views (on different attribute types). It is clearly visible that the multiview clustering (*k-means* on different attribute types of the given dataset) produces different clustering results. Figure 6 displays the data distribution of multiview clustering solutions. For each of the attribute type of network traffic, the clustering solution reflects a different data assignment. For example, the basic attributes clustering shows that, cluster 1 contains almost no normal traffic instances, whereas the content attributes clustering yields 70% normal traffic instances in cluster 1. This scenario is also visible in case of the anomalous traffic instances, each of the attribute types yield different clustering solutions. Table 4 contains the clustering solution of regular *k-means* algorithm, which means clustering on the dataset considering all the attributes types together and that is why the Table 3 and Table 4 is different.

In Table 5, we show the comparison with two other approaches. Regular clustering based approach performs basic *k-means* and creates two clusters because underlying data has normal and attack data instances. Once the clustering is done, the summary size is calculated according to the methodology discussed in Section 4.2. We applied the sampling technique on regular clustering to compare with our proposed approach. Another approach is based on random scenario, which chooses summary data instances randomly to see whether our proposed technique is actually better than the existing ones. It is clearly stated in Table 5, that our approach has less information loss than the other approaches. The proposed method did not outperform others in terms of conciseness because of the merging of summaries from four different clustering solutions, whereas, the other approaches consider only one clustering solution. Since, all the attributes are present in the summary, intelligibility is equal in all case and interestingness also suggests that our approach is better. The regular clustering approach and random approach results are similar, because both the approaches were clustered in same way. Although, there should be a difference in information loss, however due to the same size of summary and the instances picked might resulted in the similar information loss. Figure 7 showcases the evaluation among the summarization techniques based on the metrics discussed earlier (scaled 0 to 1).

7. Conclusion

In this paper, we addressed two major drawbacks of the existing clustering based summarization techniques. Summary size estimation and representing original data instances in the summary without losing any attribute are the key focus of this paper. Additionally, instead of using regular clustering algorithm for summarization, we use multiview clustering which is theoretically sound and more informative in nature for summarization. Our proposed algorithm uses sampling method pick original data instances to be added in the summary and statistical measure is used to calculate the sample size. Experimental analysis used the state-of-the-art evaluation metrics for summarization. In future, we will focus on real-time network traffic summarization.

References

[1] V. Chandola and V. Kumar, "Summarization- compressing data into an informative representation," *Knowl. Inf. Syst.*, vol. 12, no. 3, pp. 355–378, Aug. 2007.

[2] L. Wagstaff, P. Shu, D. Mazzoni, and R. Castano, "Semi-supervised data summarization: using spectral libraries to improve hyperspectral clustering," in *The Interplanetary Network Progress Report*, vol. 42, 2005.

[3] V. Ha-Thuc, D.-C. Nguyen, and P. Srinivasan, "A quality-threshold data summarization algorithm." in *RIVF*. IEEE, 2008, pp. 240–246.

[4] P. Wendel, M. Ghanem, and Y. Guo, "Scalable clustering on the data grid," in *5th IEEE International Symposium Cluster Computing and the Grid (CCGrid)*, 2005.

[5] P. More and L. Hall, "Scalable clustering: a distributed approach," in *Fuzzy Systems, 2004. Proceedings. 2004 IEEE International Conference on*, vol. 1, 2004, pp. 143–148 vol.1.

[6] M. Ahmed and A. Mahmood, "Clustering based semantic data summarization technique: A new approach," in *Industrial Electronics and Applications (ICIEA), 2014 IEEE 9th Conference on*, June 2014, pp. 1780–1785.

[7] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, Sep. 1999.

[8] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, L. M. L. Cam and J. Neyman, Eds., vol. 1. University of California Press, 1967, pp. 281–297.

[9] M. Ahmed, A. N. Mahmood, and M. R. Islam, "A survey of anomaly detection techniques in financial domain," *Future Generation Computer Systems*, 2015.

[10] M. Ahmed, A. Mahmood, J. Hu, "Outlier detection", *in The State of the Art in Intrusion Prevention and Detection, CRC Press, USA*, pp. 3–23, Jan. 2014.

[11] "NSL-KDD Datasett," accessed: 2014-06-10. [Online]. Available: http://nsl.cs.unb.ca/NSL-KDD/

[12] X. H. Dang and J. Bailey, "Generation of alternative clusterings using the cami approach," in *SDM'10*, 2010, pp. 118–129.

[13] X. Dang and J. Bailey, "A framework to uncover multiple alternative clusterings," *Machine Learning*, pp. 1–24, 2013.

[14] A. Mahmood, C. Leckie, R. Islam, and Z. Tari, "Hierarchical summarization techniques for network traffic," in *Industrial Electronics and Applications (ICIEA), 2011 6th IEEE Conference on*, 2011, pp. 2474–2479.

[15] D. Hoplaros, Z. Tari, and I. Khalil, "Data summarization for network traffic monitoring," *Journal of Network and Computer Applications*, vol. 37, no. 0, pp. 194 – 205, 2014.

[16] Testing intrusion detection systems: A critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory, ACM Trans. Inf. Syst. Secur. 3 (4) (2000) 262–294.

[17] L. Portnoy, E. Eskin, S. Stolfo, Intrusion detection with unlabeled data using clustering, in: In Proceedings of ACM CSS Workshop on Data Mining Applied to Security (DMSA-2001, 2001, pp. 5–8.

[18] W. Cochran, "Sampling Techniques." John Wiley and Sons, Inc.

[19] M. Walpole, "Fundamentals of Probability and Statistics." Prentice Hall.

[20] "Z-table," accessed: 2015-03-15. [Online]. Available: http://www.stat.ufl.edu/~athienit/Tables/Ztable.pdf