

An Efficient Approach for Complex Data Summarization Using Multiview Clustering

Mohiuddin Ahmed^(✉), Abdun Naser Mahmood, and Michael J. Maher

School of Engineering and Information Technology,
UNSW Canberra, Canberra, ACT 2600, Australia
Mohiuddin.Ahmed@student.unsw.edu.au, {A.Mahmood,M.Maher}@unsw.edu.au

Abstract. There is significant interest in the data mining and network management communities to efficiently analyse huge amount 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 summarization techniques are based on clustering and frequent itemset mining which lacks the ability to create summary for further data mining tasks such as anomaly detection. Additionally, for complex and high dimensional network traffic dataset, 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 meaningful summary from network traffic data in an efficient manner. We develop a mathematically sound approach to select the summary size using a sampling technique. The main contribution of this paper is to propose a summarization technique for use in anomaly detection. Additionally, we also propose a new metric to evaluate summary based on the presence of normal and anomalous data instances. We validate our proposed approach using the benchmark network traffic dataset.

Keywords: Scalable data mining · Network traffic summarization · Multiview clustering

1 Introduction

Summarization is considered as a key knowledge discovery approach that produces a concise, yet informative version of the original dataset [3]. Clustering, which groups together similar data instances, is often used for summarization [4–7]. Among the large pool of clustering algorithms [8], *k-means* [9] 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 may not be an actual member of the dataset.

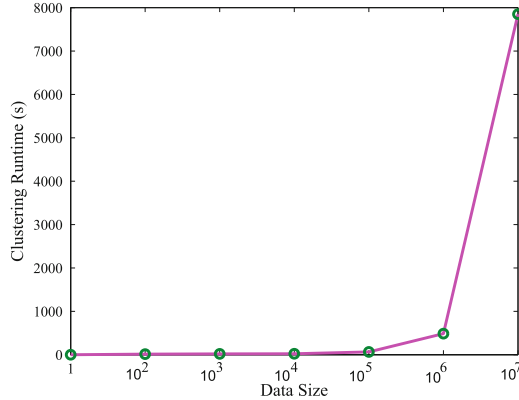


Fig. 1. Run time complexity

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 another, 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* [9] 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.

Rest of the paper contains the related works in Sect. 2, Multiview clustering and its relevance to complex data analysis is discussed in Sect. 3. We discuss our proposed approach in Sect. 4 and experimental results in Sect. 5. Section 6 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. [5] proposed a quality-threshold data summarization method modifying the *k-means* algorithm. The number of cluster is determined using

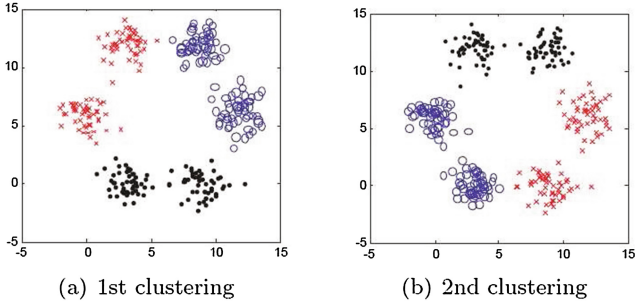


Fig. 2. Two alternative clusterings of the same dataset, each with 3 clusters. Point shapes show cluster membership, adapted from [1].

the characteristics of dataset and a threshold. The algorithm partitions a dataset until the distortion or sum of squared error (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. New centroid is introduced which is closer to the larger cluster centroid. This process is repeated until all the clusters SSE exceeds the given threshold as input. They did not explain the method to choose the threshold and how the characteristics of datasets are analysed. Patrick et al. [6] 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. [7] 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. [4] 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.

3 Multiview Clustering

Exploratory data analysis aims to identify and generate multiple views of the structure within a dataset. Conventional clustering techniques [8], 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 [2], where goal is to explore different partitions, in order to describe different grouping aspects for a given dataset. For example, consider the data given in Fig. 2 and assume the number of clusters to be uncovered is 3. It is clear that both of the clustering solutions found in two Figs. 2a and 2b 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.

3.1 Theoretical Background

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 $X = (x_1, x_2, \dots, x_N)$, the task is to find a set of alternative clustering solution, $C = (c_1, c_2, \dots)$, where the clustering quality in terms of 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 $c_1 \neq c_2$. Entropy is an important information theoretic measure to reflect uncertainty of information. For example, for a random variable R with probability distribution $p(r)$, the entropy can be defined using Eq. (1).

$$H(R) = - \int p(r) \log p(r) dr \quad (1)$$

For a pair of random variable (R, S) their joint entropy can be estimated using Eq. (2).

$$H(R, S) = - \int \int p(r, s) \log p(r, s) dr ds \quad (2)$$

Now, mutual information can be defined as the relative entropy between the joint distribution $p(r, s)$ and the product of two marginal distributions $p(r)p(s)$ as given in Eq. (3).

$$I(R, S) = \int \int p(r, s) \log \frac{p(r, s)}{p(r)p(s)} dr ds \quad (3)$$

3.2 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. The relationship among the attributes is not always significant. Moreover, network traffic dataset contains mixed attributes and thus the relationship among the attributes is quite insignificant.

4 Proposed Multiview Clustering Based Network Traffic Summary

In this Section, we describe our proposed method for network traffic summarization. At first we discuss about the necessity of sampling and the statistical approach to calculate the summary size. Then we explain our algorithm and the metric we propose for network traffic summarization.

4.1 Sampling Methods

The rationale behind integrating the sampling methods for summarization is based on the need to represent actual data instances in the summary unlike other existing methods discussed in Sect. 2 that may have average or some other representative of the data in the summary. Sampling is a popular choice for reduction of input data in data mining and machine learning techniques.

For the network traffic summarization purpose, systematic sampling is advantageous over the simple random sampling and stratified sampling 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. 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 (discussed in next Sect. 4.2).

4.2 Sample Size Calculation

Sample size determination is a very important issue because large sample size is a wastage of time and resource; on the other hand smaller sample may lead to wrong results [12]. 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 original dataset mean. According to Walpole et al. [12] view point, this error E , can be defined using the following Eq. (4). Where, $z_{\alpha/2}$ is the critical value; σ is the dataset standard deviation and n is the sample size. After rearranging Eq. (4), the sample size (summary size) can be calculated (5)

$$E = z_{\alpha/2} * \frac{\sigma}{\sqrt{n}} \quad (4)$$

$$n = \left[\frac{z_{\alpha/2} * \sigma}{E} \right]^2 \quad (5)$$

4.3 Multiview Clustering Based Network Traffic Summarization

In this Section, we describe our proposed algorithm for creating summary using the aforementioned data mining and statistical theories.

<i>MCNTS (Multiview Clustering based Network Traffic Summarization)</i>	
Input:	D , Dataset.
Output:	S , the summary of D .
Method:	
	<i>Begin</i>
1.	$Multiview\ Clustering(D) = C_1, C_2, \dots, C_k$
2.	for each clustering solution $C_i, i = 1:k$
3.	Calculate the summary size (5)
4.	$S_i = Representative\ Sample\ of\ C_i$
5.	end
6.	$S = Union\ i=1\dots k\ \{S_i\}$
	<i>End</i>

In the MCNTS algorithm, our proposed framework for network traffic summarization is presented. At first, we apply k -means clustering on the network traffic dataset [13] which has four different attribute types. For multiview clustering, we apply k -means clustering on each of the attribute types of the dataset assuming that, the dataset contains only normal and attack traffic. So, the number of clusters in the dataset is considered as two. Next, from each of the clustering solution, we calculate the sample/summary size using the statistical theories discussed in previous Sect. 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.

4.4 New Summarization Metric: Adaptability

$$\begin{aligned} Adaptability &= a1/A + n1/N + a2/A + n2/N + \dots + as/A + ns/N \\ &= (a1 + a2 + \dots + as)/A + (n1 + n2 + \dots + ns)/N \end{aligned} \quad (6)$$

Our aim is to create summaries that can be useful for anomaly detection and such summary may contain two types of data instances, one belonging to normal behaviour and the other belonging to attacks. In addition to existing summarization metrics, such as conciseness, information loss, in this paper we propose a new metric *Adaptability*; that reflects the amount of normal and attack data instances present in the summary. Adaptability can be defined as follows (6), where s represent the number of individual summary elements S_i and a Summary $S = \sum_{i=1}^s (S_i)$. Here a is the number of anomalous data in summary and A is the number of anomalous data in the original dataset, n/N represents the proportion of normal data in summary with respect to original data. Consequently, higher values of adaptability index refer to a summary’s suitability as an input to anomaly detection technique.

5 Experimental Analysis

For our experimental analysis, we used a variant of benchmark KDD cup 1999 dataset. NSL-KDD dataset [13] 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 [11].

5.1 Summarization Metrics

The summarization metrics discussed here were recently proposed and used specifically for network traffic summarization (For more details, please see [10]). **Conciseness** defines how compact a summary is with respect to the original dataset. It is the ratio of input dataset size and the summarized set size or ratio of the number of elements in the both sets (original and summarized). **Information Loss** is 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 sum of all the ratios of attributes not present by attributes represented in the summary. **Interestingness** is a new summarization metric, which focused on the objective measures of interestingness with applicability to summarization, emphasizing on diversity. **Intelligibility** is used to measure how much meaningful a summary is based on the attributes present on the summary.

5.2 Discussion on Experimental Results

Table 1 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 3 displays the data distribution of multiview clustering solutions. For each of the attribute type of network traffic, the clustering solution reflects a different

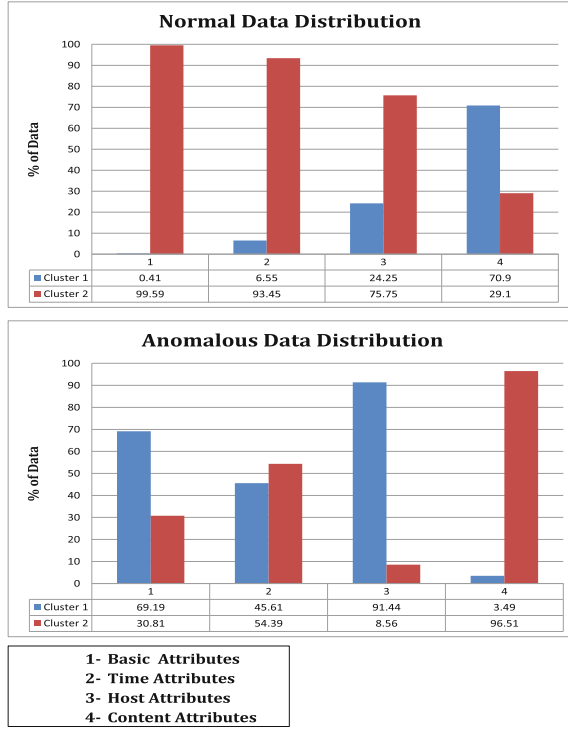


Fig. 3. Data distribution of multiview clustering solutions

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 2 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 Tables 1 and 2 is different.

Table 1. 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 %

In Table 3, 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

Table 2. Regular Clustering Results

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

Table 3. Experimental results of the MCNTS algorithm

Technique	Conciseness	Information loss	Interestingness	Adaptability
MCNTS	47.62	0.90	0.04	4.35
Regular clustering	169.07	0.94	0.003	1.17
RANDOM	169.07	0.94	0.003	1.17

done, the summary size is calculated according to the methodology discussed in Sect. 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 3, that our approach has less information loss and significantly better adaptability than the other approaches. The proposed method also resulted in inferior 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, however, the adaptability is expected to differ but due to the random selection, it reflects similar results.

6 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 and we also proposed a new metric for summarization. In future, we will focus on real-time network traffic summarization.

References

1. Dang, X.H., Bailey, J.: Generation of alternative clusterings using the cami approach, In: SDM 2010, pp. 118–129 (2010)
2. Dang, X., Bailey, J.: A framework to uncover multiple alternative clusterings. *Mach. Learn.* **98**, 1–24 (2013)
3. Chandola, V., Kumar, V.: Summarization- compressing data into an informative representation. *Knowl. Inf. Syst.* **12**(3), 355–378 (2007)
4. Wagstaff, L., Shu, P., Mazzoni, D., Castano, R.: Semi-supervised data summarization: using spectral libraries to improve hyperspectral clustering. *Interplanet. Netw. Prog.* **42**, 1–14 (2005)
5. Ha-Thuc, V., Nguyen, D.-C., Srinivasan, P.: A quality-threshold data summarization algorithm. In: RIVF, pp. 240–246. IEEE (2008)
6. Wendel, P., Ghanem, M., Guo, Y.: Scalable clustering on the data grid. In: 5th IEEE International Symposium Cluster Computing and the Grid (CCGrid) (2005)
7. More, P., Hall, L.: Scalable clustering: a distributed approach. In: Proceedings of the IEEE International Conference on Fuzzy Systems 2004, vol. 1, pp. 143–148 (2004)
8. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. *ACM Comput. Surv.* **31**(3), 264–323 (1999)
9. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Cam, L.M.L., Neyman, J. (eds.) *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 281–297. University of California Press (1967)
10. Hoplaros, D., Tari, Z., Khalil, I.: Data summarization for network traffic monitoring. *J. Netw. Comput. Appl.* **37**, 194–205 (2014)
11. John, M.: A critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory. *ACM Trans. Inf. Syst. Secur.* **3**(4), 262–294 (2000)
12. Walpole, M.: *Fundamentals of Probability and Statistics*. Prentice Hall, Englewood Cliffs (1980)
13. NSL-KDD Datasett. <http://nsl.cs.umb.ca/NSL-KDD/>. Accessed: 10 June 2014