Summarizing Significant Changes in Network Traffic Using Contrast Pattern Mining

Elaheh Alipour Chavary, Sarah M. Erfani, Christopher Leckie
ealipourchav@student.unimelb.edu.au,{sarah.erfani,caleckie}@unimelb.edu.au
Department of Computing and Information Systems, The University of Melbourne

ABSTRACT

Extracting knowledge from the massive volumes of network traffic is an important challenge in network and security management. In particular, network managers require concise reports about significant changes in their network traffic. While most existing techniques focus on summarizing a single traffic dataset, the problem of finding significant differences between multiple datasets is an open challenge. In this paper, we focus on finding important differences between network traffic datasets, and preparing a summarized and interpretable report for security managers. We propose the use of contrast pattern mining, which finds patterns whose support differs significantly from one dataset to another. We show that contrast patterns are highly effective at extracting meaningful changes in traffic data. We also propose several evaluation metrics that reflect the interpretability of patterns for security managers. Our experimental results show that with the proposed unsupervised approach, the vast majority of extracted patterns are pure, i.e., most changes are either attack traffic or normal traffic, but not a mixture of both.

CCS CONCEPTS

• **Theory of computation → Unsupervised learning and clustering**;

KEYWORDS

contrast patterns; dataset summarization; closed patterns

1 INTRODUCTION

The continual growth in Internet traffic has created major challenges for network managers who need to understand the usage of their networks. They require a compact but accurate summary report so they can quickly recognize what is happening in their network. An important form of summarization is finding the significant changes in traffic. While summarization methods such as frequent itemset mining and clustering can summarize a single dataset, summarizing the differences between multiple datasets is an open challenge.

Relevant methods for summarizing change include extracting items whose support *differs significantly* from one time window to another [3], using the *group testing* concept to randomly divide network data streams into groups and find at most one frequent item in each group. The study in [2] used two approaches for data compression: clustering and association pattern analysis. The Krimp algorithm [9] encodes high quality patterns by mining candidate sets of frequent itemsets, and accepting a candidate if it improves compression. In [5], two new metrics called Interestingness and Intelligibility are proposed to evaluate *data summarization*. The authors of [6, 7] used the concept of *closed patterns* for lossless compression to summarize data. Although these methods are useful for data compression, either they do not consider finding *important* changes in a dataset, or their scope is just a *single* dataset, rather than finding a summarized report of significant changes between multiple datasets.

In this paper, we focus on *(i)* how to provide a concise and meaningful report of significant changes in multiple datasets, *(ii)* how to evaluate the quality of generated patterns, and *(iii)* how to select the best set of patterns.

We propose the use of contrast pattern mining [1], which finds patterns whose *support* differs significantly from one dataset to another. The key idea of our method is to find a set of high quality contrast patterns, based on a representation known as *closed patterns*, which provide a lossless summarization of the data, and can be used to generate a concise and meaningful report of the significant changes for network managers. We also propose several evaluation metrics that reflect the interpretability of patterns for security managers.

Our experimental results show that our proposed unsupervised approach can extract significant changes between two datasets, because the vast majority of extracted patterns are pure, i.e., most change patterns correspond to either attack traffic or normal traffic, but not a mixture of both. The results also show that most of the pure patterns are attack traffic. Therefore, we conclude that contrast patterns have strong discriminative power that make them suitable for data summarization and finding meaningful and important changes between different traffic datasets.

2 BACKGROUND

Here, we introduce some notation from [1]. Let $U_D = \{i_1, i_2, \ldots, i_m\}$ be the set of all distinct items in dataset D. A *transaction* T is a non-empty set of items. A pattern or an itemset I is contained in a transaction T if $I \subseteq T$. We

CIKM'17 , November 6–10, 2017, Singapore, Singapore

ACM ISBN 978-1-4503-4918-5/17/11…$15.00
https://doi.org/10.1145/3132847.3133111

define $f_D(I) = \{T \in D \mid I \subseteq T\}$ as all transactions in D containing pattern I. The *count* of transactions in D that contain pattern I is given by $count(I, D) = |f_D(I)|$, and the *support* of I in D, which is the percentage of transactions in D that contain I, is given by $supp(I, D) = \frac{count(I,D)}{|D|}$. An itemset X is frequent in a dataset D if $supp(X, D)$ is greater than or equal to a pre-defined threshold ms. An itemset X is *closed* [10] in a dataset D if it is frequent and there is no superset Y of X satisfying $count(Y, D) = count(X, D)$.

Contrast patterns are patterns whose *support* is significantly different from one dataset to another. They describe differences between datasets and they are effective at extracting strong discriminating features between datasets. To measure whether there is a significant difference in support between two datasets the growth rate [4] concept can be used. The growth rate of a pattern X for dataset D_p is $gr(X, D_p) = \frac{supp(X,D_p)}{supp(X,D_n)}$. It is defined that $gr(X, D_p) = 0$ if $supp(X, D_p) = supp(X, D_n)$ and $gr(X, D_p) = \infty$ if $supp(X, D_p) > 0$ and $supp(X, D_n) = 0$.

Definition 2.1. Given a growth rate threshold $\rho > 0$, pattern X is a contrast pattern for dataset D_p if $gr(X, D_p) \geq \rho$. A contrast pattern whose support is non-zero in the positive dataset but zero in the negative dataset is called a jumping emerging pattern, and its growth rate is ∞.

3 OUR APPROACH

In this section, we first explain how we mine datasets to extract contrast patterns. Then we propose an algorithm to select the first K high quality contrast patterns and keep the most significant changes. Finally, in Section 4.3, we propose several metrics to evaluate our approach.

3.1 Contrast Pattern Mining

Our unsupervised approach for generating contrast patterns is based on a technique for data compression. We use *closed patterns* to generate a lossless compression of the data. The GC-growth algorithm [6] was used for this purpose. GC-growth uses a special prefix tree to simultaneously mine frequent patterns and closed patterns. In the next step, we mine contrast patterns from the generated closed patterns. To do so, we compare two datasets with each other, although it can be easily applied to more than two datasets.

We generate closed patterns from the joint dataset of $D = D_{pos} \cup D_{neg}$, so the *support* of each closed pattern is the *joint support*. For the *growth rate* calculation, we need the *distinct support* of each pattern per dataset. Therefore, first, for each closed pattern we find its *support* in D_{pos} and D_{neg} separately, and then calculate its *growth rate*. Finally, according to Definition 2.1 if $gr(X, D_p) \geq \rho$ we consider the closed pattern to be a contrast pattern. After extracting the complete list of contrast patterns, the next step is to remove similar contrast patterns, by keeping the top-k high quality patterns as described below.

ALGORITHM 1: Select Best Set of Contrast Patterns

Data: $\mathcal{X} = \{X_1, X_2, \ldots, X_k\}$ is the set of contrast patterns, $gr(X_i)$ is growth rate of each contrast pattern, $cd(X_i)$ is class coverage differentiation of each contrast pattern, Z is a cut-off constant

Result: set of best contrast patterns

1 $Z \longleftarrow Z * |\mathcal{X}|$;
2 sort $\mathcal{X}$ according to growth rate in a descending order;
3 **for** *all item* $X_i \in \mathcal{X}; i = 1, 2, \ldots, k$ **do**
4 | $gr(X_i) \longleftarrow \log_2(gr(X_i))$;
5 **end**
6 $GrMargin \longleftarrow Z * \frac{(gr(X_1) - gr(X_k))}{|\mathcal{X}|}$;
7 add X_1 to the list of best pattern C_{bp};
8 **for** *all item* $X_i \in \mathcal{X}; i = 2, 3, \ldots, k$ **do**
9 | $bp \longleftarrow$ *get the last pattern of* C_{bp};
10 | **if** $(X_i \subset bp$ *or* $bp \subset X_i)$ ***and*** $(gr(bp) - gr(X_i)) < GrMargin$ **then**
11 | | **if** $cd(X_i) > cd(bp)$ **then**
12 | | | replace bp in C_{bp} by X_i;
13 | | **end**
14 | **else**
15 | | add X_i to C_{bp};
16 | **end**
17 **end**

3.2 The Best Set of Contrast Patterns

To provide a concise report of significant changes for network administrators, we require a method for selecting a set of high quality contrast patterns while removing patterns that are slight variations of each other. Our approach is that among patterns with similar differentiating power, keep those with the highest coverage and purity based on two metrics: growth rate and class coverage differentiation.

Definition 3.1. Class coverage differentiation is the percentage of transactions in the positive dataset covered by a given contrast pattern with regard to each class of traffic:

$$\begin{aligned} Pattern\ coverage &= \frac{|f_A(X)|}{|D_{pos}(att)|} - \frac{|f_N(X)|}{|D_{pos}(nrm)|} \\ &= supp(X, D_{pos}(att)) - supp(X, D_{pos}(nrm)). \end{aligned}$$

Algorithm 1 shows the pseudo code for generating the best set of contrast patterns. It contains two main steps. In the first step, the contrast patterns are grouped according to their growth rate similarity. All patterns in a group are a subset or superset of each other. To formulate a similarity measure, we define a "gap margin" according to line 6. For the gap margin calculation, a cut-off value Z needs to be determined. The value of Z can be tuned based on the percentage of contrast patterns we would like to represent in the final report. In each group, the first member is a pattern with the highest growth rate and the other members are selected if their growth rate differentiation with the first member is smaller than the gap margin. The second step keeps the pattern with the highest

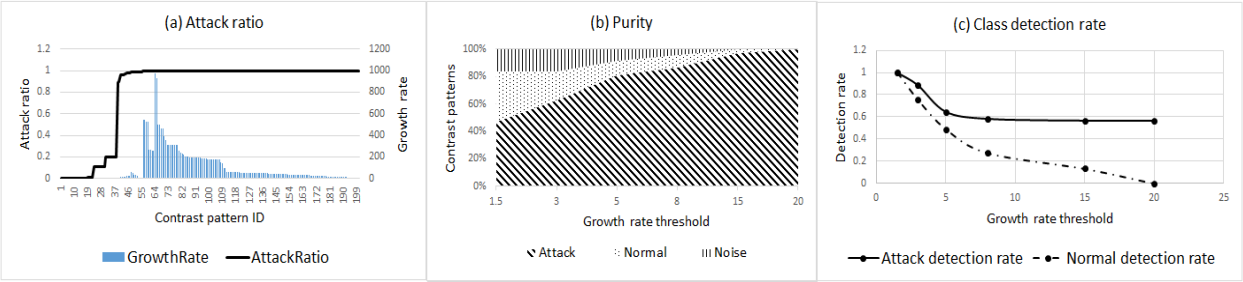

Figure 1: Attack ratio, purity and detection rate

class coverage differentiation, giving the pattern that covers more transactions and is more pure.

4 EXPERIMENTAL EVALUATION

Our aim is to investigate the following questions:

1. Are a high proportion of generated contrast patterns pure patterns? Purity means that contrast patterns belong to either an attack class or a normal class, and few patterns are mixed/noisy. If so, then contrast pattern mining would be an effective method for detecting significant changes that are mainly pure patterns.

2. Among the pure patterns, are a high proportion attack patterns? If so, then contrast patterns can have a strong discriminative power in detecting attack traffic.

3. How does the proportion of attack patterns change with an increase in the growth rate threshold? In particular, with an increase of the growth rate threshold, are similar patterns filtered and attack patterns preserved? If so, then we could conclude that strong contrast patterns are attack patterns.

In the next section, first we explain the datasets. Then after pre-processing, in Section 4.3 we provide some criteria for analyzing the generated contrast patterns. Finally, we present and discuss our results.

4.1 Network Traffic Representation

We focus on the Kyoto 2006+ datasets[1] to evaluate the quality of generated contrast patterns. It contain 24 features of attack and normal traffic. Among them, 14 features are based on the KDD Cup 99 dataset.[2] In the Kyoto 2006+ dataset all traffic is labeled as either normal or intrusion. We considered the traffic of 15 July 2009 as a positive dataset and the traffic of 01 July 2009 as a negative dataset. The total number of transactions is 244900, with 117730 as attack sessions and 127170 as normal sessions. For generating closed patterns, we experimentally set the minimum support threshold to 0.1%.

4.2 Preprocessing

Feature selection: To select the most relevant features we used those in [8], namely: **Service, Source bytes, Destination bytes, Count, Same srv rate, Dst host same src port rate, Dst host srv serror rate**.[1]

[1]http://www.takakura.com/Kyoto_data/
[2]http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

Discretization: For contrast pattern mining we need to transform continuous feature values to discrete values. We transformed the vector dataset to a transaction dataset using equal-frequency unsupervised discretization.

Next, we propose several metrics for network managers to analyze the generated contrast patterns.

4.3 Evaluation Metrics

A major challenge in our evaluation is how to assess whether the generated contrast patterns are informative to network managers. In this section we propose measures that can be used to assess the discriminative power of contrast patterns. It should be mentioned that for the purposes of evaluation, each Kyoto 2006+ dataset has two classes of traffic: *attack class* and *normal*.

First, we need to calculate the set of attack sessions and normal sessions for each contrast pattern in D_{Pos}. If X is a contrast pattern, $D_{Pos}(att) = \{T \in D_{pos} \mid Class = Attack\}$ is the set of all attack transactions in D_{Pos}, and similarly for $D_{Pos}(nml)$. We define $f_A(X) = \{T \in D_{pos}(att) \mid X \subseteq T\}$ as all attack transactions in D_{Pos} containing pattern X. Thus $|f_A(X)| = count(X, D_{pos}(att))$. Similarly $f_N(X) = \{T \in D_{pos}(nml) \mid X \subseteq T\}$. Thus $|f_N(X)| = count(X, D_{pos}(nml))$. Finally, we define $f_{D_{Pos}}(X) = \{T \in D_{pos} \mid X \subseteq T\}$, i.e, all transactions in D_{Pos} containing contrast pattern X. We say that $|f_{D_{Pos}}(X)| = |f_A(X)| + |f_N(X)| = count(X, D_{pos})$.

The first measure is *attack ratio*: Given a contrast pattern X, what is the probability that X belongs to the attack class?

Definition 4.1. Attack ratio is the probability that a given contrast pattern X belongs to the attack class:

$$attack\ ratio = \frac{count(X, D_{pos}(att))}{count(X, D_{pos})} = \frac{|f_A(X)|}{|f_{D_{Pos}}(X)|}.$$

Attack ratio evaluates the predictive ability of a specific contrast pattern for the attack class in the positive dataset.

The second measure is the *attack detection rate*. What proportion of attack transactions in the positive dataset are covered by the set of mined attack contrast patterns? Let $\mathcal{X} = \{X_1, X_2, \ldots, X_k\}$ be the set of all mined contrast patterns. We define coverage as $f_{cvg}(att) = \cup_{i=1}^{k} f_A(X_i)$.

Definition 4.2. Attack detection rate is the proportion of attack transactions in the positive dataset covered by

generated contrast patterns:

$$attack\ detection\ rate\ =\ \frac{|f_{cvg}(att)|}{|D_{pos}(att)|}.$$

The *attack detection rate* corresponds to *recall* in binary classification. For a growth rate threshold of 1 the attack coverage is 100%, but some portion may be repeated attacks. When we increase the growth rate threshold, a rapid fall is observed in the *attack detection rate*. The reason is that many repeated attacks are filtered by the growth rate threshold. Thus, we need to alleviate the effect of repeated attacks. Our solution to this problem is to *normalize* the detection rates by dividing by a baseline detection rate. This baseline can be calculated by varying the growth rate threshold between 1 to 2. The *normalized attack detection rate* is defined as *attack detection rate/baseline attack rate*.

The third measure is the detection rate for normal traffic. The aim is to determine the percentage of total normal traffic coverage in a positive dataset. We define the coverage function as $f_{cvg}(nml) = \cup_{i=1}^{k} f_N(X_i)$.

Definition 4.3. Normal detection rate is the proportion of normal transactions covered by the generated contrast patterns:

$$normal\ detection\ rate\ =\ \frac{|f_{cvg}(nrm)|}{|D_{pos}(att)|}.$$

Again, for filtering the repeated normal traffic we normalize *normal detection rate* by dividing by a baseline detection rate. The *normalized normal detection rate* is defined as *normal detection rate/baseline normal rate*.

In the next section, we show some results extracted from our approach. These results show the strong power of contrast patterns for providing a concise report of significant changes.

4.4 Results

To evaluate the quality of the extracted contrast patterns, we show the graphs in Figure 1. Figure 1(a) illustrates the *attack ratio* and the *growth rate* per contrast pattern. The minimum growth rate is set to 5. It is clear from the attack ratio plot that generated contrast patterns can efficiently distinguish between attack and normal traffic: 76% of contrast patterns can uniquely distinguish between classes with a probability of 100%. If we decrease the class detection probability to 90%, then 90.5% patterns have this distinguishing power. In addition, it shows that a high proportion of contrast patterns are attack patterns. With a probability of 90%, 79.5% of patterns correspond to attack traffic and only 11% are normal traffic. The other plot shown in Figure 1(a) is the growth rate plot. It is interesting to note that all contrast patterns with a high growth rate are attack patterns.

It is clear from the Figure 1(b) that a vast majority of contrast patterns are pure patterns. For example, when the minimum growth rate is 1.5, nearly 80% of contrast patterns are pure, and for growth rates of 15 or higher, all patterns are pure (attack). In addition, Figure 1(b) shows that most of the pure patterns correspond to attacks. At a minimum growth rate of 5, nearly 80% of contrast patterns are attacks, compared to only 11% of normal patterns. The proportion of attack patterns increases significantly with an increase of the growth rate threshold. When the growth rate threshold increases from 1.5 to 15, the attack contrast patterns rise markedly from 47% to 97%.

Figure 1(c) presents the detection rates for attack and normal traffic for different growth rate thresholds. The graphs were normalized based on a baseline detection rate for a growth rate threshold of 1.5. A gradual fall of 40% can be observed in the attack detection rate until the growth rate threshold is around 8. After that the detection rate remains nearly stable at 60%. In contrast, when the growth rate threshold increases, the normal detection rate drops steadily until it reaches zero. So, it demonstrates that strong contrast patterns that correspond to high growth rate thresholds tend to all be attack patterns.

To evaluate the reliability of the proposed method, we repeated the experiment with two other pairs of datasets (July 25 and July 19 2009, July 25 and July 15 2009). Again we find that a high percentage of contrast patterns are pure, and among the pure patterns most are attacks.

5 CONCLUSION AND FUTURE WORK

In this paper, we presented a method for generating a summarized report of significant changes between two traffic datasets. For this purpose we used an unsupervised approach using contrast pattern mining, which yields a set of high quality contrast patterns. We also proposed several evaluation metrics to assess the quality of our contrast patterns. Our experimental results show that most contrast patterns are pure patterns and belong to either attack traffic or normal traffic, but not a mixture of both. In addition, we showed that among the pure patterns most are attacks. In future work, we aim to use the generated contrast patterns as the basis for *clustering traffic* without the use of a distance function.

REFERENCES

[1] James Bailey. 2012. *Contrast Data Mining: Concepts, Algorithms, and Applications*. Chapman and Hall/CRC.

[2] Varun Chandola and Vipin Kumar. 2005. Summarization-compressing data into an informative representation. In *ICDM*.

[3] Graham Cormode and S. Muthukrishnan. 2005. What's new: Finding significant differences in network data streams. *TON* (2005), 1219–1232.

[4] Guozhu Dong and Jinyan Li. 2005. Mining border descriptions of emerging patterns from dataset pairs. *KAIS* (2005), 178–202.

[5] Demetris Hoplaros, Zahir Tari, and Ibrahim Khalil. 2014. Data summarization for network traffic monitoring. *JNCA* (2014), 194–205.

[6] Haiquan Li, Jinyan Li, Limsoon Wong, Mengling Feng, and Yap-Peng Tan. 2005. Relative risk and odds ratio: A data mining perspective. In *PODS*. 368–377.

[7] Jinyan Li, Guimei Liu, and Limsoon Wong. 2007. Mining statistically important equivalence classes and delta-discriminative emerging patterns. In *SIGKDD*. 430–439.

[8] Adetunmbi A Olusola, Adeola S Oladele, and Daramola O Abosede. 2010. Analysis of KDD99 Intrusion detection dataset for selection of relevance features. In *WCECS*. 20–22.

[9] Jilles Vreeken, Matthijs Van Leeuwen, and Arno Siebes. 2011. Krimp: mining itemsets that compress. *DMKD* (2011), 169–214.

[10] Jianyong Wang, Jiawei Han, and Jian Pei. 2003. Closet+: Searching for the best strategies for mining frequent closed itemsets. In *SIGKDD*. 236–245.