講演番号：136

ダークネットの UDP パケット分類法の一検討

D－19

Study of Classification method of UDP packet of darknet

山本 貴幸† 笠間 貴弘‡ 宮保 憲治†

Takayuki YAMAMOTO† Takahiro KASAMA‡ Noriharu MIYAHO†

† 東京電機大学大学院 情報環境学研究科

† Graduate School of Information Environment, Tokyo Denki University

‡国立研究開発法人 情報通信研究機構

‡ National Institute of Information and Communications Technology

1. はじめに

ダークネットには常時，多数のパケットが到達し，その殆どは攻撃に関連したパケットである可能性が高い．1 日に 100 万パケット数以上の UDP パケットが東京電機大学に到達するが，分析者が全パケットを手動で分析するのは困難であり，膨大な時間を要する．パケット分析を効率的に行うためには，宛先ポート番号で当該パケットを分類・整理する方法が考えられるが，同一種類の通信の中に，複数のポート番号を使用するパケットが存在する．本稿では，同じプロトコルのペイロードには，視覚的類似性に関わる共通部分が観測されることを活用する分類手法を検討した．当該のダークネット宛に到達した UDP パケットのペイロードに対して，宛先ポート番号を利用しない分類法に独自性がある．

2. 実験に使用したデータ

本学がダークネットにおけるパケット収集を開始した 3 月 8 日のデータをクラスタリング用に活用し，翌日の 3 月 9 日のデータを分類精度評価用として実験評価を行った．

3. UDP ペイロードのグレースケール画像化

ペイロードを 1Byte ごとに二次元配列として編成した．ペイロードサイズが 100Byte 以下の場合は 10×10，101～400Byte の場合は 20×20，401～900Byte の場合は 30×30，901Byte 以上の場合は 40×40 の2次元配置を取った．ペイロードサイズが画像サイズに満たない分に対しては，0xFF=255(10 進数)(白色)でパディングを行った．

4. クラスタリング

実際のパケット分析環境と同じクラスタ数が未知の場合を想定し，事前にクラスタ数を指定する必要のない DBSCAN[1]を用いて，グレースケール画像から抽出した GIST[2]特徴量のクラスタリングを行った．

5. 分類

各クラスタ内の各サンプルの GIST と分類対象の GIST のユークリッド距離を計算した．最小ユークリッド距離が閾値未満の場合は同じクラスタに属するとし，クラスタと分類対象のラベルが一致した場合には正解とした．最小ユークリッド距離が閾値以上となったとき，分類対象のラベルがどのクラスタのラベルにも存在しない場合には，割当てられるべきクラスタは存在しないため，正解とした．

6. 最小ユークリッド距離による分類の精度評価

分類結果を表 1 に示す．ここで 2 つの文字列 x1 と x2 に対するレーベンシュタイン距離を標準化した，標準化レーベンシュタイン距離(NLD: Normalized Levenshtein Distance)を用いて文字列類似度を計算した．NLD は，2 つの文字列の類似度を0%～100%の間で示す．NLD の定義式を以下に示す．

$$NLD(\%) = \left(1 - \frac{LevenshteinDistance(x1, x2)}{max\left(len(x1), len(x2)\right)}\right) \times 100$$

表 1. 最小ユークリッド距離による分類結果

割り当て条件	分類精度	
最小ユークリッド距離 0.4 未満	総合分類精度	84.3%
	特定ポートにのみ到達	87.7%
	複数ポートに到達	70.5%

分類を行った結果，分類が成功したデータにおいて，割当先クラスタが存在したペイロードの多くは文字列類似度 30%を上回った．最小ユークリッド距離の閾値を 0.6 として分類を行った際は，最小ユークリッド距離の閾値が 0.4 のときに誤分類となったプロトコルの内 2 つが分類に成功し，新たに 2 つのプロトコルについて誤分類が発生した．しかし，新たに誤分類となった 2 つのプロトコルについては，両方とも文字列類似度が 30%以下となった．

以上述べたことから，最小ユークリッド距離の閾値を 0.6，文字列類似度の閾値を 30%とし，分類精度を再評価した．

7. 文字列類似度を併用した分類の精度評価

文字列類似度を併用した分類結果を表 2 に示す．全体的な分類精度が向上し，特に複数ポートに到達する同一プロトコルのパケットの分類精度が向上することが判明した．

表 2. 文字列類似度を併用した分類結果

割り当て条件	分類精度	
・最小ユークリッド距離 0.6 未満 ・文字列類似度(NLD)30%以上	総合分類精度	86.8%
	特定ポートにのみ到達	88.6%
	複数ポートに到達	79.7%

8. まとめ

最小ユークリッド距離と文字列類似度を併用することで，複数ポートに到達する同一プロトコルのパケットを 79.7%の精度で分類できた．今後は，分類精度の一層の向上を図るため，DNS 通信の場合のように，多様なペイロードが多数存在する同一プロトコルの検討やサイズが非常に小さいペイロードについて分類精度の検討を進める予定である．

参考文献

[1]Martin Esteret.al. “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise“, AAAI Press, pp. 226-231, 1996.

[2]Oliva,A, et.al. “Modeling the shape of a scene: a holistic representation of the spatial envelope”, IJCV, Vol. 42(3), 145-175, 2001.