WIDE Technical-Report in 2008

WIDE-CNRS 間の交換留学活動報告

wide-tr-mawi-widecnrs-himura-00.pdf

WIDE Project : http://www.wide.ad.jp/

If you have any comments on this document, please contact to ad@wide.ad.jp

WIDE-CNRS 間の交換留学活動報告

肥村 洋輔

平成 20 年 12 月 1 日

概要

WIDE プロジェクトおよび CNRS 間の学生交換として，東京大学大学院 江崎研究室 修士 1 年の 肥村 洋輔 が ÉNS Lyon に 一ヶ月間，Patrice Abry らの下で研究活動を行った．研究内容はインターネットトラフィック分類手法の性能向上であり，この問題に対するアプローチとして，コネクションパターンを考慮した特徴量を用い，MST (Minimum Spanning Tree) クラスタリングによる分類を行った．その結果，コネクションパターンを考慮しないクラスタリングに比べて独立性の高いクラスタが得られ，また，未知のトラフィックグループを新たに発見することができ，コネクション構造を考慮した特徴量によって既存の分類手法の性能向上を行うことができた．

1 はじめに

WIDE プロジェクトの研究活動のひとつに，フランス国立科学センター (CNRS : Centre National de la Recherche Scientifique) との協力活動があり，その一環として両者は学生交換を行っている．本年度は，東京大学大学院 江崎研究室 修士 1 年の 肥村 洋輔 が対象者の一人として，2008 年 9 月 1 日から 2008 年 9 月 30 日までの一ヶ月間，フランスのリヨンにおいて研究活動を行った．受け入れ先は，École Normale Supérieure de Lyon の Patrice Abry らの研究室である．Patrice は，同大学の Pierre Borgnat，Guillaume Dewaele とともに，WIDE プロジェクトメンバの 長 健二朗 (IIJ 技術研究所)，福田 健介 (国立情報学研究所) との協力体制でインターネットトラフィック解析の研究を行っている．その中でも特に異常トラフィック検出に焦点を当て，統計的なモデルに基づく異常トラフィックの検出を行う手法 [1] を提案し，同手法の評価および検出トラフィックの解析を行っている．

本学生交換における研究トピックは，異常検出手法の性能評価・性能比較をさらに高精度に行うための，トラフィック分類手法に関する調査および実装・解析である．異常検出手法を評価するためには予め分類された (ラベル付けをされた) データが必要であるが，現在使用されている分類手法はポート番号および TCP フラグを用いた経験的な方法であり，総ホスト数 (本レポートにおいてはホスト単位でトラフィック分類を行う) のうち 30% 程度が未分類である．これらの未分類ホストは評価結果に多かれ少なかれ影響を与えるため，早急な性能向上が望まれている．この問題に対するアプローチとして，コネクションパターンを考慮した特徴量を採用し，MST に基づくクラスタリング [2] を行うことで，新たな分類を試みた．この手法はポート番号に依存しない分類を行うため，現在使用している手法と相互に性能向上を行うことができる．結果として，コネクションパターンを考慮したトラフィック分類手法の有効性を確認することができ，既存の分類手法の性能向上を行うことができた．

2　成果

本研究において，分類性能向上のために 2 つの知見を利用した．それらは，(1) コネクションパターンを考慮した特徴量 および (2) MST クラスタリング である．

2.1　コネクションパターンを考慮した特徴量の発見

現在，提案されている統計的トラフィック分類手法が用いている多くの分類手法は，フローサイズや平均パケットサイズなどの “一次元的” な特徴量を主に採用しているが，高い分類精度や未知トラフィックの発見などにおいて，十分な結果を出していないと考えられる．そこで，我々は “二次元的” な特徴量の発見のため，コネクションパターンを構造化して調査を行った．

図 1 に構造化の例 (1 つの送信元ホストに着目し，コネクションパターンの構造化を行った) を示す．ここで，srcIP は送信元 IP アドレス，proto はトランスポート層プロトコル，srcPort は送信元ポート，dstPort は宛先ポート，dstIP は宛先 IP アドレスを意味する．用いたデータは MAWI トラフィックレポジトリ [3] の 2007 年 09 月 16 日における 14:00 から 14:15 までのトラフィックであり，先頭の 100 パケット分のみを抽出して構造化および視覚化がなされている．また，付加的な情報として，線の太さおよびノードの大きさ (フローサイズ) およびフローの色 (コネクションの状態[1]) も視覚化した．ここで，図 1(a) におけるホスト (srcIP) は，P2P アプリケーションを用いていると考えられる．なぜならば，同ホストは 2 つのプロトコルを用い，通信相手は複数存在し，ポート番号は通信相手によって異なり，各フローはほぼ独立しているためである．一方，図 1(b) におけるホストは明らかに，送信元ポートを適宜変更しつつ水平ポートスキャン [4] を行っていると考えられる．

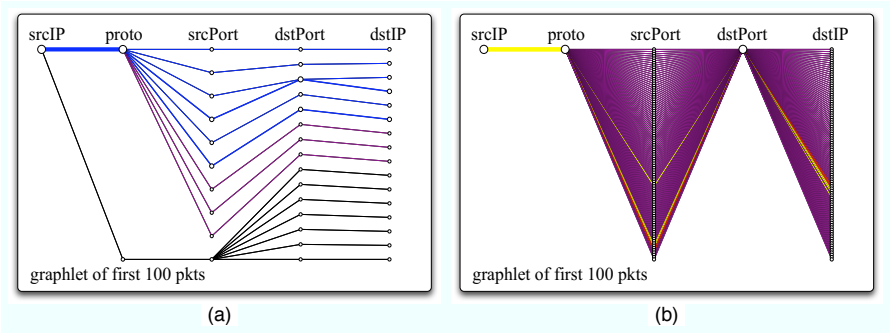

図 1: コネクションパターン構造化の例：(a) P2P トラフィック，(b) ポートスキャン

このように，各ホストのコネクションパターン構造を視覚化し調査を行った結果，コネクションパターンはホスト分類における重要な特徴量になり得ることが発見された．一方，これらの特徴は非常に直感的ではあるが，機械的分類用の特徴量としての抽出は未解決問題である．本報告書においては，機械的分類を行うための出だしとして，次の特徴量を選択する．

[1] 青：TCP シーケンス番号の整合性が保たれている，黄：TCP シーケンス番号の整合性が保たれていない，赤：TCP コネクションが明らかに不成立である，黒：UDP フローである，橙：ICMP フローである，紫：1 パケットのみで構成されるフローである

- (送信元ポート数) / (宛先 IP アドレス数)：この特徴量は，クライアント・サーバ型通信でのクライアントとしては 1 より非常に大きな値をとり，サーバとしては 0 に近い値をとる．一方，P2P 型通信では，1 に近い値をとると考えられる．
- (宛先ポート数) / (宛先 IP アドレス数)：この特徴量は，クライアント・サーバ型通信でのクライアントとしては 0 に近い値をとり，サーバとしては 1 より非常に大きな値をとる．一方，P2P 型通信では，1 に近い値をとると考えられる．
- (宛先 IP アドレス数) / (総パケット数)：この特徴量は，ポートスキャンであれば 1 に近い値をとり，それ以外では 0 に近い値をとると考えられる．

2.2 MST クラスタリング

クラスタリングは，特徴量の類似度が高いもの同士をグループに分ける操作であり，特徴の隠れた構造を発見することができる．本学生交換においては，MST (Minimum Spanning Tree) を用いるクラスタリング手法 [2] に着目した．この手法は，代表的なクラスタリング手法である K-means 法 [5] などとは異なり，クラスタ内のデータ数およびクラスタの形状に依存せず，ノイズに強いロバストなクラスタリングを行う．そのため，特徴量空間に複雑な構造を持つインターネットトラフィックに対するクラスタリング手法として，本手法が適切であると考えられる．

図 2 に MST クラスタリングの例を示す．クラスタリングの流れは次のようになる：

(a) 特徴量空間にホストの持つ特徴量をプロットする．

(b) 全てのホストをエッジで結合する．このとき，エッジの総距離が最小かつ木にループがないように結合する．

(c) 距離が閾値以上のエッジを切断する．その結果，複数のホスト群 (クラスタ) を見つけることができる．

図 2 では 2 次元特徴量空間による例であるが，実際には以下の 8 つの特徴量を用いてクラスタリングを行った．

(1) (送信元ポート数) / (宛先 IP アドレス数)
(2) (宛先ポート数) / (宛先 IP アドレス数)
(3) (宛先 IP アドレス数) / (総パケット数)
(4) スモールサイズパケットの割合
(5) ラージサイズパケットの割合
(6) H (ミディアムサイズパケットのパケットサイズ)
(7) (H (IP[3])) / (H (IP[4]))
(8) (H (IP[2])) / (H (IP[4]))

ここで，H (x) は “変数 x の分布のエントロピー”， IP[n] は “IP アドレスの n オクテット目” を表す．また，スモールサイズパケットおよびラージサイズパケットとは，全長がそれぞれ 144 バイト未満および 1400 バイト以上のパケットを指し，ミディアムサイズパケットは先述した条件に当てはまらないパケットを指す．なお，距離としてユークリッド距離を採用するが，特徴量の値域はそれぞれ異なるため，適宜正規化を行っている．

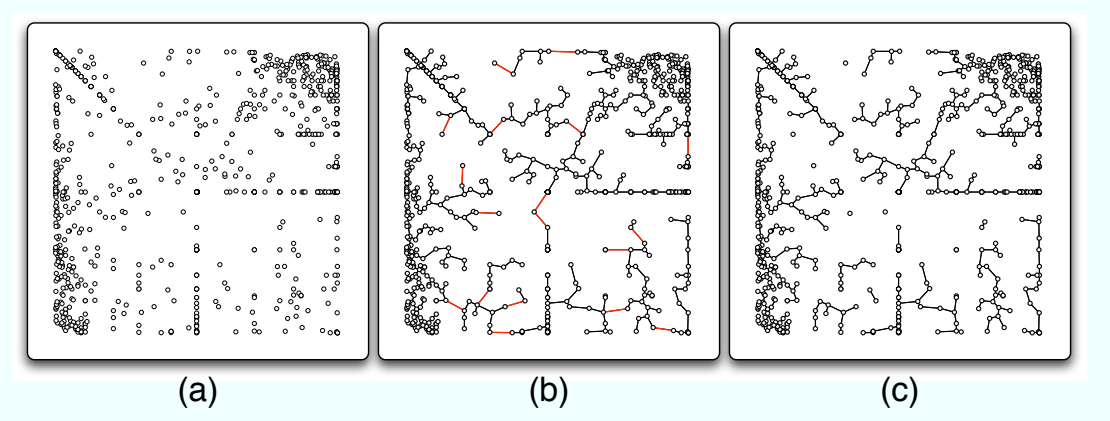

図 2: MST クラスタリングの例：(a) 特徴量空間へのプロット，(b) MST の決定 (赤線は距離が一定値以上のエッジ)，(c) クラスタの決定

8 次元特徴量空間におけるクラスタリング結果の一例を，表 1 に示す．列はポート番号に基づく分類結果，行はクラスタリングによる分類結果である．用いたデータセットは，2007 年 9 月 16 日のデータ (15 分) である．このクラスタリングにより得られた結果は，以下の 2 点である：

- 2 手法間のクロスバリデーション：表 1 は，ポート番号に基づく分類手法およびコネクションパターンに基づく分類手法による分類結果を比較したものである．表 1 によると，これらの手法は概念の大きく異なる手法であるにもかかわらず，各クラスタの独立性が高いことが理解できる．これは，両手法の妥当性・信頼性を高めるだけではなく，各種法の分類ミスを検出することにおいても有効である．
- 未知トラフィックの新たな識別：ポート番号に基づく分類手法では分類できなかったホスト (表 1 における未分類カテゴリのホスト) は，クラスタリングによって複数のクラスタに分かれている．未分類ホストを特徴に応じて分類することで，人間による精査を効率的にするだけではなく，同クラスタの他のカテゴリのホストとの比較により，類似したアプリケーションを判断する際に有効である．

3 評価

コネクションパターンを考慮した特徴量 (1), (2), (3) を取り入れたことによるクラスタリングの有効性を示すために，これらの特徴量がある場合とない場合のクラスタリング結果を比較する．ここで，特徴量数を増加させると必ずしも精度が上がるわけではなく，逆に性能を低下させる恐れがあることに注意されたい．表 1 は 8 特徴量におけるクラスタリング結果，表 2 は上記特徴量を除いた 5 特徴量におけるクラスタリング結果である．評価に公平性を持たせるために，総クラスタ数が同程度になるようにクラスタリングを行い，ホストが多い順にクラスを並べ替えた．なお，現在は定性的な評価のみを行うこととする．なぜならば，クラスタリング評価において一般的に用いられている指標 (F 尺度，エントロピー，相互情報量など) は本結果に用いることは不適切である．これらの尺度はクラスタの独立性・完備性に主眼が置かれているが，本研究の目的は (1) 未分

類ホストを分類するための解析 および (2) 同カテゴリ内の新たな構造発見 であるため，クラスタの独立性および完備性は必ずしも必要とならない．

表 1 と表 2 を比較した時，特にクラスタ 2 において表 1 が直感的にも適切なクラスタリングが行われていることが分かる．なぜならば，表 2 では，HTTP クライアント，DNS サーバ，ポートスキャナという，比較的容易に分類できるホストが同一クラスタに属しているためである．また，同クラスタの未分類カテゴリを精査したところメールサーバのトラフィックが多く見られた．これら 4 種類のホストは，表 1 では，クラスタ 2，4，7，12 にそれぞれ表れている．これらは，コネクションパターンを考慮した特徴量を追加したことにより，分類が可能となったといえる．図 3 に，各ホストの代表的なコネクションパターンを示す．各トラフィックは異なるコネクション構造を持つことが明らかである．

この分類方法は，コネクションパターンを考慮するという点において BLINC[6] がとっている方法に類似している．しかし，BLINC はルールに基づく分類方法に主眼が置かれているため，(1) ロバストな分類 および (2) 未知トラフィックの新たな識別 を行うことができないと考えられる．これに対して，本手法はクラスタリングに基づく教師なし学習を行うことで，上記二点の問題を解決できている．

表 1: 8 特徴量によるクラスタリング結果

クラスタ No.	合計	HTTP サーバ	HTTP クライアント	SCAN	DNS	その他	未分類
1	173	0	104	0	3	5	61
2	58	0	52	0	2	2	2
3	46	0	23	0	2	4	17
4	42	0	0	0	37	1	4
5	39	37	0	0	0	0	2
6	24	0	0	0	24	0	0
7	18	0	0	18	0	0	0
8	16	0	0	12	0	2	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

表 2: 5 特徴量によるクラスタリング結果

クラスタ No.	合計	HTTP サーバ	HTTP クライアント	SCAN	DNS	その他	未分類
1	243	1	134	1	10	15	82
2	146	0	39	14	46	8	39
3	55	52	0	0	0	0	3
4	20	0	0	0	0	1	19
5	17	0	0	17	0	0	0
6	17	12	0	0	0	0	5
7	16	0	4	0	0	9	3
8	9	1	1	1	6	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

4 今後の課題

定量的評価：本報告書において定性的評価を行った理由は，F 尺度およびエントロピーなどの広く一般的に用いられているクラスタリングの評価指標は我々の意図に即さず，本研究の評価として

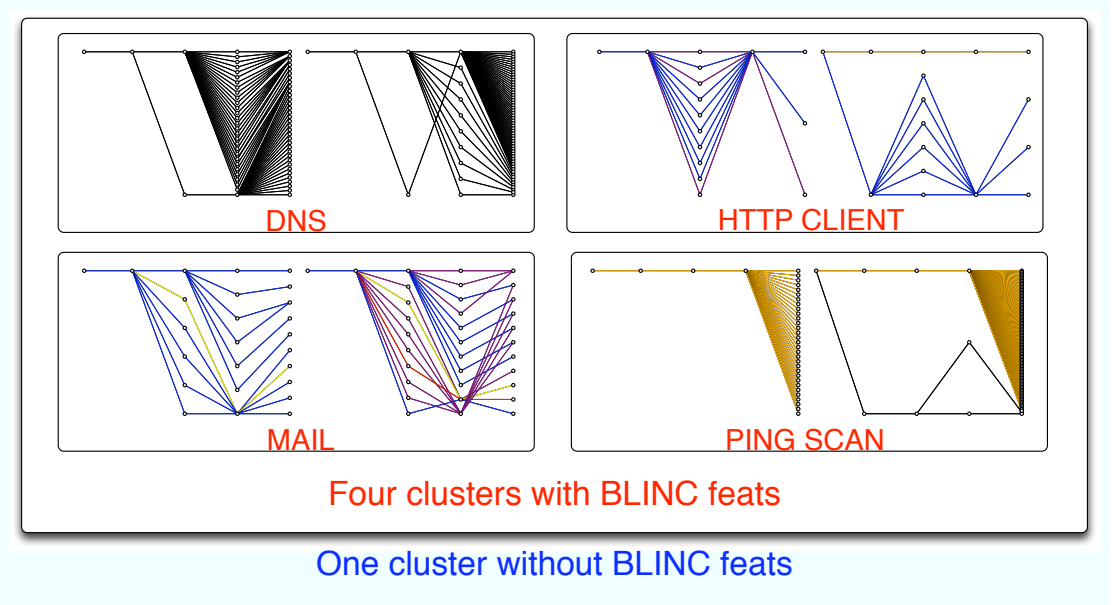

図 3: コネクションパターン特徴量による出現した新たなクラスタ

適切ではないためである．今後は的確な評価を与える指標を開発し，定量的評価を行う必要があると考えられる．

特徴量の再検討：今回はコネクションパターン構造を表す特徴量として，(1) 送信元ポート数と宛先ホスト数の比，(2) 宛先ポート数と宛先ホスト数の比，(3) 宛先ホスト数と総パケット数の比を用いた．これらの特徴量は統計的分類を行う上で強力であることを確認できたが，コネクションパターンを的確に記述する上では不十分である．なぜならば，フローサイズおよび TCP コネクションの成立・不成立などを考慮していないためである．今後は，これらの状態を的確に記述する特徴量の調査および分類への応用を行う必要があると考えられる．また，特徴量の正規化方法についても検討を行う必要がある．

クラスタの更なる精査：クラスタリングにより，統計的特徴の類似したホストに分類することができた．今後は，特に未分類クラスタのホストを調査することにより，未知ホストの解明および新たな経験則の追加について研究を行う．

トラフィックデータベースの構築：トラフィック分類を行う上で発見できる多種の知見をデータベースとして体系的にまとめ，トラフィック解析に携わる研究者を主として広く公開する．

5　まとめ

本学生交換で行った研究活動は，インターネットトラフィック異常検出手法のより信頼性の高い評価のためのトラフィック分類手法の向上である．この問題に対するアプローチとして，(1) コネクションパターンを考慮した特徴量を発見し，(2) MST (Minimum Spanning Tree) クラスタリングによる分類を行った．その結果，コネクションパターンを考慮しないクラスタリングに比べて，独立性の高いクラスタが得られ，また，未知のトラフィックグループを新たに発見することができ，コネクション構造を考慮した特徴量の有効性を確認するとともに，分類手法の性能向上を行うことができた．

参考文献

[1] G. Dewaele, K. Fukuda, P. Borgnat, P. Abry, and K. Cho. Extracting Hidden Anomalies using Sketch and Non Gaussian Multiresolution Statistical Detection Procedure. *ACM SIGCOMM'07 LSAD workshop*, pages 145–152, August 2007.

[2] C. Zahn. Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Transaction on Computers Volume C-20*, pages 68–86, January 1971.

[3] K. Cho, K.Mitsuya, and A. Kato. Traffic Data Reposity at the WIDE Project. *USENIX 2000 FREENIX Track*, June 2000.

[4] J. Mirkovic and P. Reiher. A taxonomy of DDoS attack and DDoS defense mechanisms. *ACM SIGCOMM Computer Communication Review Volume 34 Issue 2*, pages 39–53, April 2004.

[5] J. Erman, M. Arlitt, and A. Mahanti. Traffic Classification Using Clustering Algorithms. *ACM SIGCOMM'06 MINENET workshop*, pages 281–286, September 2006.

[6] T. Karagiannis, K. Papagiannaki, and M. Faloutsos. BLINC: Multilevel Traffic Classification in the Dark. *ACM SIGCOMM'05*, pages 229–240, August 2005.