The Korean Journal of Applied Statistics (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.1.001
28(1), 1–8

Network Classification of P2P Traffic with Various Classification Methods

Seokwan Han[a] · Jinsoo Hwang[a,1]

[a]Department of Statistics, Inha University

(Received August 21, 2014; Revised December 23, 2014; Accepted December 24, 2014)

Abstract

Security has become an issue due to the rapid increases in internet traffic data network. Especially P2P traffic data poses a great challenge to network systems administrators. Preemptive measures are necessary for network quality of service(QoS) and efficient resource management like blocking suspicious traffic data. Deep packet inspection(DPI) is the most exact way to detect an intrusion but it may pose a private security problem that requires time. We used several machine learning methods to compare the performance in classifying network traffic data accurately over time. The Random Forest method shows an excellent performance in both accuracy and time.

Keywords: Traffic, classification, P2P, network, learning.

1. 서론

전통적인 네트워크 트래픽의 분류는 포트(port)기반 혹은 페이로드(payload)기반의 두 가지로 구분된다. 그러나 각각의 방법은 최근의 발전하는 해킹 기술, 방화벽을 우회하거나 포트번호를 암호화하여 숨기거나 정상적인 포트번호로 위장을 하거나 등의 방법으로 그 효율성이 많이 떨어지고 있는 실정이다. 따라서 이를 해결하는 방법 중의 하나로서 트래픽을 이루는 패킷들의 크기나 시간 또는 방향 등의 정보만을 이용하여 네트워크의 트래픽을 분류하는 방법인 기계학습방법이 많이 연구되고 있다.

네트워크의 트래픽은 예전에는 메일, FTP, TELNET, NEWS, WEB 등의 대부분이었지만 요즘은 대부분이 P2P와 WEB 트래픽이고 이메일 등의 트래픽이 적은 일부를 구성하고 있다. 특히 P2P 트래픽의 엄청난 증가는 네트워크의 속도와 보안에 커다란 문제를 제기하고 있다. 많은 P2P 트래픽이 방화벽을 건너뛰기 위해 정상적인 WEB 트래픽 포트인 80번을 이용한다는 것은 이미 널리 알려진 사실이다.

본 연구에서는 초기의 몇 패킷의 크기와 방향만을 이용하여 트래픽을 통계적으로 분류하는 방법, 특히 P2P 트래픽을 잘 구별해 내는 방법을 찾으려고 기존의 여러 방법들을 잘 알려진 공개된 트래픽 데이터베이스 자료를 바탕으로 비교실험을 하였다. 비교실험에 사용되는 분류 알고리즘으로는 LDA, QDA, Naive Bayes, KNN, KD-TREE, Random Forests, SVM을 이용하였다.

2장에서는 관련된 연구와 실험에 사용하고자 하는 알고리즘을 간단히 소개하고, 3장에서는 연구방법에 대한 간단한 소개를 하고 4장에서는 비교실험 결과 그리고 마지막으로 5장에서는 결론과 추후 연구에 대하여 논의하고자 한다.

This work was supported by the National Research Fund(NRF-2013R1A1A2059335).
[1]Corresponding author: Department of Statistics, Inha University, 100 Inha-ro, Nam-gu, Incheon 402-751, Korea. E-mail: jshwang@inha.ac.kr

2. 관련연구 및 분류 알고리즘

전통적인 분류방법 중에서 포트기반방법은 패킷에 적혀있는 포트번호를 보고 트래픽의 종류를 판단하는 방법으로서 P2P와 같은 대부분의 트래픽을 분류하는 데에는 적합하지 않다. 왜냐하면 P2P 트래픽은 동적인 포트번호를 사용하거나 암호화된 패킷 사용 또는 기존 WEB 트래픽의 대표적인 프로토콜인 HTTP가 사용하는 80번 포트로 위장하여 방화벽을 통과하는 방법을 사용하고 있다. 이러한 P2P에 대한 문제점과 분석은 Karagiannis 등 (2004)에 정리되어 있다. 또한 페이로드기반 분류기법은 Deep Packet Inspection(DPI) 접근법이라고도 불리며 패킷 속의 내용을 자세히 조사하여 어떤 트래픽인지를 판단하는 방법이다. 이 방법은 가장 정확하게 패킷을 분류할 수 있는 방법이지만 시간과 노력이 많이 소요되면 패킷 내용 조사에 따른 사적인 정보 노출의 위험이 있어서 사용하기에 어려움이 있을 수 있다. 이처럼 인터넷에서의 트래픽 분류에 수반되는 여러 문제점을 Nguyen 등 (2008)이 각 방법별로 장단점을 조사한 결과를 제시하였다.

다른 방법으로는 패킷들의 통계적인 특성이외에 시계열적인 특성을 은닉마르코프체인(Hidden Markov Chain)을 이용하여 분석한 Dainotti 등 (2008)과 Mu 등 (2011)의 연구가 있다. 그리고 Munz 등 (2010)은 8개의 상태공간과 4 단계를 이용한 단순한 마르코프모델을 사용하였으며 Zhang 등 (2013)은 인접한 패킷들의 상관관계를 이용하여 분류의 효율을 높이는 방법을 연구하였다.

다음은 사용된 분류알고리즘 중 일부에 대한 간단한 정리이다.

2.1. KD-TREE(K-Dimensional Tree)

일반적으로 KNN의 문제점은 자료의 개수가 많아질수록 시간이 많이 걸리게 된다. 따라서 이를 해결하기위한 방법 중의 하나로 사용되는 방법이 KD-tree 방법이다. Kd-tree 분류방법은 k차원의 데이터 공간에 존재하는 데이터들을 구조화 하기위해 공간을 분할하여 데이터 개체를 구조화 시켜가는 방법이다. 데이터 집합 S를 포함하는 hyperspace에서 시작하고, 이 최초의 공간은 더 작은 hyperspace로 반복 분해된다. 공간을 분할하는 방향은 미리 정의된 규칙에 따르고, 분할로 생성되는 hyperspace에 포함된 개체가 미리 정해진 할당량만큼 작아질 때까지 반복된다. 분할 규칙은 데이터 공간을 나누는 기준이 되는데, Standard-split, midpt-split, fair-split이 가장 널리 알려진 방법이다. Standard-split은 분할 차원에 대한 최솟값과 최댓값의 중앙값에 수직방향으로 공간을 분할하고, aspect ratio(node들에 해당하는 공간의 가장 긴 방향의 차원 길이와 가장 짧은 차원길이의 비율)가 높은 값을 가진다. midpt-split은 분할되는 차원의 중간 값을 기준으로 분할하지만 트리의 size가 커질 수 있다. fair-split은 aspect ratio가 항상 3 : 1을 넘지 않도록 분할하며 aspect ratio도 적절하고 트리의 모양 또한 균형을 이루고 있다.

탐색방법은 표준탐색(standard search)과 우선탐색(priority search)이 있다. 전자는 주어진 대상이 어느 리프노드에 속하는지를 루트 노드에서부터 아랫방향으로 탐색하여 리프노드에 도착한다. 우선탐색기법은 트리의 루트노드로부터 리프노드까지 차례로 탐색하며 쿼리개체와 트리노드의 경계가 가장 가까운 순서대로 대기열을 형성하여 인접한 이웃을 탐색한다. 쿼리와 대기열에 남은 미방문 노드와의 거리가 지금까지 찾은 KNN과의 거리보다 크면 탐색이 종료된다.

2.2. Random Forests

Random Forests는 Breiman (2001)에 의해 고안된 앙상블 기법 중 하나로 나무들 사이의 상관관계를 감소시킴으로써 분산 감소를 향상시키기 위한 방법이다. 이는 입력 변수를 랜덤하게 선택하여 나무를 성장시킴으로써 이루어진다. Random Forests 알고리즘은 다음과 같다.

Table 3.1. Traffic data size per port

port	21	22	25	80	110	119	143	443	995	BT
size	22445	25000	25000	25000	25000	4556	25000	25000	2919	21279

1. training data로부터 N개의 붓스트랩 샘플을 추출한다. 이 때, 복원추출방법을 이용한다.
2. 설명변수 p개에서 랜덤하게 m개의 변수를 선택한다. m개의 변수들 사이에서 최적의 변수와 분리점을 선택, 그리고 노드를 두 개의 자식으로 분리한다. 이 때, m은 보통 $\sqrt{p}$가 된다.
3. 위 두 과정을 나무의 각각 끝 노드에서 최소 노드 크기 k에 도달할 때가지 반복한다.
4. 위의 세 과정을 총 B번 귀납적으로 반복한다.
5. 회귀나무의 경우, B개의 트리에서의 평균값이 예측값이 되고, 분류나무의 경우, B개의 트리의 예측값들에서 다수결에 의해 예측값이 결정된다.

Random Forests는 multi-class classifier로 매우 큰 데이터베이스를 다루는 데 효과적일 뿐 아니라 속도 또한 빠르다. 특히 높은 차원의 특성(feature)벡터를 입력받는 것에 유리하고 특성벡터 중 분류에 중요한 영향을 미치는 특징을 고를 수도 있는 장점을 가지고 있다.

2.3. SVM(Support Vector Machine)

1995년에 Vapnik 등 (1977)에 의해 제안된 SVM은 통계적 학습 이론에 기반하여 분류분석과 회귀분석을 수행한다. SVM의 분류 원리는 각 그룹에서 데이터 간의 거리를 측정하여 중심을 구한 후, 그 가운데에서 최적의 초평면(hyperplane)을 구함으로써 그룹을 예측하는 것이다. SVM의 가장 큰 장점은 비선형분류에 효과적인 성능을 나타낸다. 커널 함수를 사용하여 Feature Space라는 새로운 공간에 선형판별을 수행함으로써, 마치 실제 데이터에는 매우 복잡한 비선형 판별을 수행한 결과로 나타난다. 그러나 SVM은 $O(M^3)$의 time complexity(M: 학습 데이터의 수)를 갖고 있어 학습 데이터가 많아질수록 학습 시간이 급격히 늘어나는 단점이 있다.

3. 데이터 구성과 연구 방법

여러 분류방법을 비교하는 기준이 되는 실제 네트워크 트래픽 데이터는 공개된 패킷 트레이스 자료(http://www.simpleweb.org/wiki/Traces)를 수집하여 해당 트레이스로부터 pcap library 함수를 이용하여 적절한 TCP 연결 자료를 추출하여 사용하였다. 여기에서 추출한 트래픽 자료는 모두 포트번호가 이미 알려진 non-P2P로서 9개의 포트에 대한 특정기간동안의 실제 자료이다. 각 포트에서는 서버와 클라이언트 간의 handshake 후의 처음 4개의 패킷의 크기와 방향 자료만을 자료로 이용하였다. 방향자료는 +는 서버에서 클라이언트로의 패킷을 나타내고 −는 클라이언트에서 서버로의 패킷을 나타낸다. P2P 패킷자료로서는 pcap library에서 정해진 시간대에 호스트들 간에 충분하게 서로 패킷교환이 일어난 호스트들과의 BitTorrent 자료를 이용하였다. 이 자료와 위에서 수집한 일반포트의 패킷자료를 가지고 분류방법을 서로 비교하는데 사용하였다. 분류방법으로는 전통적인 선형판별법(LDA), 이차판별법(QDA), 베이즈분류법(Naive Bayes; NB)과 KNN방법을 사용하였으며 최근 방법으로는 KNN의 시간을 단축하기위한 KD-tree법(KD), 랜덤포리스트(RF)방법, 그리고 SVM방법을 사용하였다.

Table 3.1에서는 비교연구에서 사용하는 non-P2P 포트별로 수집한 자료의 크기와 P2P의 경우는 전부 BitTorrent(BT) 항목으로 묶어 크기를 나타내었다.

Table 4.1. Precision and Recall of P2P/P80

방법	P2P		P80	
	precision	recall	precision	recall
LDA	0.897±0.006	0.755±0.009	0.752±0.012	0.895±0.008
QDA	0.732±0.078	0.910±0.074	0.939±0.085	0.805±0.045
Naive Bayes	0.061±0.350	0.510±0.198	0.950±0.004	0.543±0.139
KNN	0.931±0.458	0.985±0.230	0.988±0.017	0.944±0.218
KD-Tree	0.931±0.031	0.985±0.003	0.988±0.002	0.944±0.025
Random Forests	0.947±0.023	0.993±0.004	0.994±0.003	0.957±0.019
SVM	0.916±0.038	0.972±0.011	0.977±0.009	0.932±0.031

분류방법별 성능비교는 트래픽 분류 성능에서 보편적으로 사용되는 다음과 같은 척도를 사용하였다.

- $\text{recall}_k = \Pr(\text{classified as } k \,|\, \text{true class } k)$
- $\text{precision}_k = \Pr(\text{true class } k \,|\, \text{predicted as } k)$

위의 recall과 precision의 조화평균으로 계산한 F-measure가 하나의 값으로 간혹 제시되기도 한다. 성능비교에서 소요되는 시간도 중요하므로 각 방법 당 소요되는 시간도 동시에 계산하였다. 각 포트별로 자료 개수가 차이가 나므로 이를 감안한 시간도 계산하였다.

분류의 성능 비교는 10 fold cross-validation을 수행하여 그 결과의 평균과 표준편차를 계산하였다. 분류 방법별로 tuning parameter들은 사전실험을 통하여 적절한 값을 가지고 실험하였다. KNN과 KD-tree에서는 $k = 5$를 사용하였고 Random Forests의 나무 수는 R 프로그램을 이용한 데이터 분류를 할 때, 실험이 가능하였던 최소의 나무 수인 20부터 시작해 변화를 시켜본 결과 적절한 20을 이용하였다.

4. 결과

첫 번째 비교실험은 이진분류실험으로서 일반 정상 트래픽인 non-P2P들과 BT 트래픽간의 여러 분류방법에 따른 분류성능을 비교하였고 두 번째 비교실험은 다중분류실험으로서 9개 포트의 일반정상 트래픽과 BT 트래픽을 여러 분류 방법에 따라서 분류성능을 비교하였다.

4.1. 첫 번째 실험결과

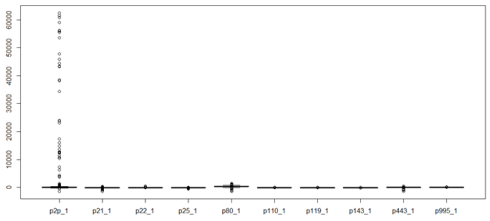
이진 분류실험결과는 P2P 트래픽이 가장 많이 위장하여 사용하는 포트 80의 정상 트래픽과의 결과에 대한 성능을 먼저 다음 Table 4.1에서 보여주고 있다.

나머지 포트들과 P2P 트래픽과의 결과는 F-measure의 평균값만을 다음 Table 4.2에서 보여준다. 실제 자료 중 앞의 4개 패킷의 크기의 분포를 알아보기 위하여 첫 패킷인 패킷1 크기의 분포를 상자그림으로 나타낸 결과가 Figure 4.1에 있으며 그 옆에 일부의 요약통계량을 정리하였다. 패킷2, 3, 4의 자료도 유사한 패턴을 보이므로 그 결과는 생략한다. 그림을 보면 P2P 자료는 정상포트의 자료에 비하여 이상치가 많아서 정상포트 분포들의 비교가 어렵다. 따라서 P2P를 제외하고 나머지 정상포트들의 패킷1의 분포를 요약한 상자그림은 Figure 4.2에 있다.

이진분류에서는 각 포트별 자료의 개수가 다르므로 분류방법과 P2P와 비교하는 정상 포트 트래픽들의 특성에 따른 분류의 평균소요시간(초)과 변동계수(CV)를 다음 Table 4.3에서 볼 수 있다. 소요시간을 보면 SVM이 개당 분류시간이 가장 많이 걸리며 가장적은 LDA에 비해서는 약 40배정도 더 시간이

Table 4.2. Results of average F-measures

방법/포트	p21	p22	p25	p80	p110	p119	p143	p443	p995
LDA	0.680	0.739	0.761	0.830	0.761	0.000	0.749	0.825	0.597
QDA	0.797	0.849	0.853	0.879	0.919	0.857	0.937	0.775	0.982
Naive Bayes	0.777	0.837	0.855	0.694	0.837	0.546	0.914	0.752	0.974
KNN	0.993	0.991	0.998	0.985	0.998	0.983	0.999	0.992	0.992
KD-Tree	0.993	0.991	0.998	0.985	0.998	0.983	0.999	0.992	0.992
Random Forests	0.998	0.994	0.998	0.991	0.999	0.988	0.996	0.998	0.998
SVM	0.995	0.994	0.999	0.978	0.997	0.986	0.999	0.995	0.996



	Min	Q_1	Q_2	Q_3	Max
p2p-1	−1460	12	18	161	62520
p21-1	−1411	−89	−59	−43	320
p25-1	−480	−99	−99	−98	42
p80-1	−1460	286	377	509	1460
p143-1	−291	−268	−149	−112	47
p443-1	−1460	0	67	67	1460

Figure 4.1. The first packets of P2P and normal ports

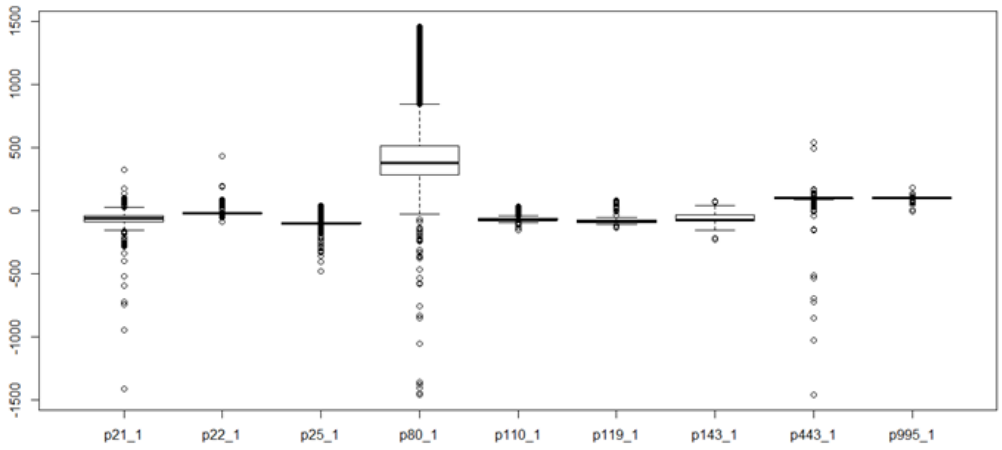


Figure 4.2. Size distributions of the first packet of normal ports

Table 4.3. Average classification time per methods and CV

방법	LDA	QDA	NB	KNN	KD	RF	SVM
평균(CV)	1.0(0.05)	1.1(0.11)	2.2(0.09)	3.0(0.23)	1.8(0.02)	1.6(0.12)	40.1(0.38)

소요된다. 하지만 분류의 정확성을 고려한 KNN, KD, RF, SVM을 비교해보면 특정포트 P80에서의 precision과 recall 뿐 아니라 모든 정상포트들의 F-measure에서도 RF가 가장 우수한 분류성능을 보여주고 있으며 SVM에 비해 25배나 빠른 결과를 제공해 준다.

Table 4.4. Accuracy of multiple classification of several classification algorithms

	LDA	QDA	NB	KNN	KD	RF
정확도	0.513±0.001	0.609±0.001	0.558±0.001	0.987±0.001	0.989±0.001	0.994±0.001
시간(초)	18.6	18.9	100.5	527.2	75.6	56.3

Table 4.5. Results of multiple classification of KD, KNN, and RF

포트	정확도(KD / KNN / RF)	
	recall	precision
P2P	0.985 / 0.986 / 0.992	0.986 / 0.986 / 0.994
P21	0.972 / 0.974 / 0.989	0.975 / 0.974 / 0.988
P22	0.994 / 0.994 / 0.997	0.992 / 0.992 / 0.997
P25	0.992 / 0.992 / 0.995	0.988 / 0.987 / 0.995
P80	0.987 / 0.986 / 0.992	0.985 / 0.986 / 0.992
P110	0.987 / 0.986 / 0.992	0.990 / 0.990 / 0.995
P119	0.975 / 0.974 / 0.992	0.974 / 0.974 / 0.990
P143	0.996 / 0.995 / 0.998	0.996 / 0.996 / 0.998
P443	0.993 / 0.993 / 0.998	0.993 / 0.993 / 0.997
P995	0.995 / 0.998 / 1.000	0.992 / 0.992 / 0.993

4.2. 두 번째 실험결과

두 번째 비교 분류실험 결과는 다중분류에 대한 결과이다. 9개의 정상 포트 트래픽의 자료와 BitTorrent(BT) P2P자료를 전체 10개의 class로 하여 분류실험을 한 결과가 Table 4.4에 정리되었다. 단, SVM의 경우에는 30분 이상 소요되어 그 결과를 제외하였다. 정확도는 전제 트래픽 중에서 제대로 분류된 트래픽 개수의 비율(recall)로 측정하였으며 소요시간도 함께 측정하였다. 이 결과를 보면 시간과 정확도의 측면에서 RF(Random Forest)방법이 이진분류에서와 마찬가지로 가장 우수한 성능을 보여주었다.

다중분류에서 3가지 우수한 방법들의 포트별 recall과 precision의 결과를 Table 4.5에서 보면 KD와 KNN에서는 p21, p119에서 상대적으로 97%대의 정확도를 보이지만 여기에서도 RF는 99%대의 결과를 주고 있다. 따라서 RF방법의 우수성을 포트별 결과에서도 확인할 수가 있다.

5. 결론

다양한 분류 알고리즘을 이용하여 P2P 데이터와 그렇지 않은 데이터로 분류하는 실험을 해보았다. 이진분류에서나 다중분류에서도 마찬가지로 LDA, QDA, Naive Bayes 알고리즘의 경우, 정확도가 다른 알고리즘에 비해 낮음을 볼 수 있고, SVM 알고리즘의 경우, Random Forests, KD-Tree와 비슷하게, 정확도 면에서는 우수하지만, 측정 시간이 길게 나타나 효율적인 측면에서 단점을 보인다. KNN 역시 정확도가 우수하지만, 데이터의 크기가 커질수록 시간이 많이 소요되어 현실적으로 온라인상에서 적용하기에는 어려움이 따를 것으로 생각된다. 이에 대한 대안으로 KD-Tree와 Random Forests 알고리즘을 고려할 수 있으며 그 중에서도 Random Forests 알고리즘의 높은 성능으로 확인할 수 있었다. 이처럼 데이터의 크기가 커지고, 클래스의 수가 많아지면 Random Forests와 KD-Tree 알고리즘은 P2P 데이터를 포함하는 트래픽 데이터를 분류함에 있어 가장 높은 효율을 보여주고 있음을 알 수 있었다.

본 연구에서 사용된 트래픽 자료가 아닌 다른 시점에서 추출한 자료로 우리의 알고리즘을 테스트하여 보았으나 그 결과가 크게 다르지 않아 논문에 포함시키지는 않았다. 물론 캡쳐되는 시간대에 따라서 트래

픽의 종류와 크기가 달라지긴 하였다. 그러나 네트워크의 트래픽은 여러 다양한 형태로 진화 발전하므로 여기 실험된 제한된 트래픽 자료로서 제안된 분류 방법의 우수성을 판단하기에는 어려움이 따른다. 따라서 포트번호가 알려져 있지 않은 수많은 새로운 네트워크의 트래픽 자료를 포함한 다양한 환경에서 분류에 대한 연구가 필요하다고 할 수 있다.

References

Breiman, L. (2001). Random forest, *Machine Learning*, **45**, 5–32.

Dainotti, A., Donato W. D., Pescape, A. and Rossi, P. S. (2008). Classification of network traffic via packet-level hidden Markov models, In *Proceedings of IEEE Global Telecommunications Conference, November*.

Karagiannis, T., Broido, A., Brownlee, N. and Claffy, K. Is P2P dying or just hiding?, In *Proceedings 47th annual IEEE Global Telecommunications Conference (Globecom 2004)*, Dallas, Texas, USA, November/December 2004.

Nguyen Thuy, T. T. and Armitage, G. (2008). A survey of techniques for internet traffic classification using machine learning, *IEEE Communications Surveys and Tutorials*, **10**, 56–76.

Mu., X., Wu, W. and Enabled C. (2011). A parallelized Network traffic classification based on hidden Markov model, *Distributed Computing and Knowledge Discovery, October*.

Munz, G., Dai, H., Braum, L. and Carle, G. (2010). TCP traffic classification using Markov models, *TMA'10 Proceedings of the Second International Conference*, 127–140.

Vapnik, V., Golowich, S. and Smola, A. (1977). Support vector method for function approximation, regression estimation and signal processing, *Advances in Neural Information Processing Systems*, **9**, 281–287.

Zhang, J., Xiang, Y., Wang Y., Zhou, W., Xiang, Y. and Guan, Y. (2013). Network traffic classification using correlation information, *IEEE Transactions on Parallel and Distributed Systems*, **24**, 104–117.

http://www.simpleweb.org/wiki/Traces

다양한 분류기법을 이용한 네트워크상의 P2P 데이터 분류실험

한석완[a] · 황진수[b,1]

[a]인하대학교 통계학과

(2014년 8월 21일 접수, 2014년 12월 23일 수정, 2014년 12월 24일 채택)

요 약

인터넷 트래픽의 증가로 인하여 네트워크의 보안 문제가 중요한 문제로 대두되고 있다. 그 중에서도 특히 P2P 트래픽의 증가는 모든 서버의 관리자에게는 해결해야할 중요한 문제로 대두되고 있다. 서버에서 네트워크 트래픽을 조사하여 문제가 있는 트래픽을 미리 차단하는 것은 서비스 품질의 향상과 자원의 효율적인 사용 측면에서 바람직하나 오가는 패킷의 내부정보를 조사하는 것은 개인정보보호 차원에서 문제가 있을 수 있으며 시간과 노력이 많이 소요되므로 요즘은 통계적인 기계학습의 방법을 이용하여 이상 트래픽을 찾아내는 연구가 주를 이루고 있다. 본 연구에서는 최근의 기계학습방법 중에서 널리 쓰이는 방법들을 비교 연구하여 그 결과 랜덤포리스트(random forest)라고 불리는 방법의 우수함을 보였다.

주요용어: 트래픽, 분류, 피투피, 네트워크, 학습.

이 논문은 한국연구재단의 지원에 의한 연구임(NRF-2012R1A1B3003545).

[1]교신저자: (402-751) 인천시 남구 인하로 100, 인하대학교 통계학과. E-mail: jshwang@inha.ac.kr