

다중 클래스 SVM을 이용한 계층적 인터넷 애플리케이션 트래픽의 분류

Hierarchical Internet Application Traffic Classification using a Multi-class SVM

유재학* · 이한성** · 임영희* · 김명섭* · 박대희**

Jaehak Yu*, Hansung Lee**, Younghye Im*, Myung-Sup Kim* and Daihee Park**

* 고려대학교 컴퓨터정보학과

** 한국전자통신연구원

요 약

본 논문에서는 인터넷 애플리케이션 트래픽 분류방법으로 대표되는 포트 번호 및 페이로드 정보를 이용하는 방법론의 한계점을 극복하는 대안으로서, SVM을 기반으로 한 계층적 인터넷 애플리케이션 트래픽 분류 시스템을 제안한다. 제안된 시스템은 이진 분류기인 SVM과 단일클래스 SVM의 대표적 모델인 SVDD를 계층적으로 결합한 새로운 트래픽 분류 모델로서, 학내에서 수집된 양방향 트래픽 플로우 데이터에 대한 최적의 속성 부분집합을 선택한 후, P2P 트래픽과 non-P2P 트래픽을 빠르게 분류하는 첫 번째 계층, P2P 트래픽들을 파일공유, 메신저, TV로 분류하는 두 번째 계층, 그리고 전체 16 가지 애플리케이션 트래픽별로 세분화 분류하는 세 번째 계층으로 구성된다. 제안된 시스템은 인터넷 애플리케이션 트래픽을 coarse 혹은 fine하게 분류함으로써 효율적인 시스템의 자원 관리, 안정적인 네트워크 환경의 지원, 원활한 대역폭의 사용, 그리고 적절한 QoS를 보장할 수 있다. 또한, 새로운 애플리케이션 트래픽이 추가되더라도 전체 시스템을 재학습시킬 필요 없이 새로운 애플리케이션 트래픽만을 추가 학습함으로써 시스템의 점증적 갱신 및 확장성도 가능하다. 실험을 통하여 제안된 시스템의 성능을 검증한다.

Abstract

In this paper, we introduce a hierarchical internet application traffic classification system based on SVM as an alternative overcoming the uppermost limit of the conventional methodology which is using the port number or payload information. After selecting an optimal attribute subset of the bidirectional traffic flow data collected from the campus, the proposed system classifies the internet application traffic hierarchically. The system is composed of three layers: the first layer quickly determines P2P traffic and non-P2P traffic using a SVM, the second layer classifies P2P traffics into file-sharing, messenger, and TV, based on three SVDDs. The third layer makes specific classification of the entire 16 application traffics. By classifying the internet application traffic finely or coarsely, the proposed system can guarantee an efficient system resource management, a stable network environment, a seamless bandwidth, and an appropriate QoS. Also, even a new application traffic is added, it is possible to have a system incremental updating and scalability by training only a new SVDD without retraining the whole system. We validate the performance of our approach with computer experiments.

Key Words : Internet application classification, Support vector machine, Attribute subset selection

1. 서 론

인터넷의 급속한 발전과 이를 이용하는 사용자의 수가 급증함에 따라, 보다 다양하고 새로운 네트워크 서비스들이 개발되어 상용화되고 있다. 특히 네트워크 서비스 중, P2P(Peer-to-Peer) 관련 트래픽은 2009년 독일 Ipoque사의 조사에 의하면, 지난 몇 년 동안 전체 인터넷 트래픽에서 60%이상을 차지하고 있으며 앞으로도 상당한 비중을 차지할 전망이다[1-2].

인터넷 총 트래픽의 상당수를 차지하는 P2P는 대용량의 파일을 포함하는 엄청난 트래픽과 대칭적 성격으로 인하여 네트워크상에서 혼잡한 상황을 유발할 뿐만 아니라, 고비용의 하부구조 업그레이드 및 네트워크 재분할 등의 추가적 비용을 요구한다[2]. 따라서 다양한 네트워크 서비스에 보다 적합한 QoS(Quality of Service) 및 안전한 네트워크 환경 제공을 목적으로 하는 네트워크 망 관리시스템에서, P2P를 포함하는 인터넷 애플리케이션 트래픽의 보다 빠르고 정확한 분류가 최근 학계의 중요한 이슈 중 하나이다[2-4].

인터넷 애플리케이션 트래픽을 분류하는 전통적인 방법론은 포트 번호를 이용하는 방법과 페이로드 정보를 이용하

접수일자 : 2009년 4월 30일

완료일자 : 2010년 1월 12일

+ 교신저자

는 방법으로 나눌 수 있다[2-3]: 1) 포트 번호 기반의 분류 방법은 IANA에서 할당된 잘 알려진 포트(well-known port) 번호를 분석하는 비교적 단순하면서도 실용적인 방법이지만, 동적으로 포트 번호를 할당하여 패킷을 발생하거나 기존에 사용한 잘 알려진 포트 번호를 다른 목적으로 이용하는 최근의 애플리케이션들이 증가함에 따라 정확한 트래픽 분류가 어렵다; 2) 페이로드 정보에 기반한 분류방법은 페이로드의 특징을 추출하고 이를 패킷과 비교하는 방법으로서, 패킷을 발생할 때 페이로드를 암호화할 경우 이를 사용할 수 없다. 또한 최근에 불법침입의 차단 및 고비용의 접근 등을 이유로 페이로드의 접근을 금지하고 있는 실정이 페이로드 정보에 기반한 방법론의 사용을 더욱 어렵게 한다. 따라서 기존의 전통적인 방법에서 벗어난 새로운 방법론의 대안이 요구된다.

최근의 연구문헌 조사에 의하면, 애플리케이션의 변화에 대처할 수 있는 새로운 해결책으로써 데이터마이닝 및 기계학습 기법을 인터넷 애플리케이션 트래픽 분류에 적용하려는 시도가 성공적으로 진행 중이다[3-8]. 기계학습 기법은 동적으로 변하는 포트 번호와 암호화된 페이로드에 독립적인 데이터의 특징 벡터로부터 교사학습 혹은 비교사학습 방법을 통하여 중요한 패턴들을 찾아낸다는 점이 최근의 애플리케이션 변화에 대처할 수 있음을 시사한다. 또한 네트워크 트래픽 데이터의 성격이 대용량의 스트림 데이터임을 고려할 때, 대용량의 데이터 처리를 위한 데이터마이닝 기법의 적용은 매우 적절하다. 이러한 연구 동향 중, 특히 패턴분석을 위한 자동화 알고리즘의 역사적 진화과정 중, 가장 강력하다고 이미 검증된 SVM(support vector machine)을 인터넷 애플리케이션 트래픽 분류에 적용하려는 연구가 주목을 받고 있다[5-8]. 현재까지의 SVM을 이용한 인터넷 애플리케이션 트래픽 분류는 패턴 분류 및 함수 근사 등의 문제에서 매우 우수한 성능을 보이는 SVM을 사용하여 인터넷 애플리케이션 트래픽 분류 문제에 적용하는 가능성을 검증하는 비교적 초기의 시도로서 다음의 두 가지 방법론을 취하고 있다: 첫 번째 방법[5-6]에서는 이진 분류기인 SVM을 이용하여 P2P 트래픽과 non-P2P 트래픽을 단순히 이분 분류하는 방법으로 P2P 트래픽을 탐지하고 있으며, 두 번째 방법[7-8]에서는 일대일 방법(one-against-one method)과 같이 여러 개의 이진 분류기인 SVM을 유기적으로 결합하여 다중 클래스 SVM(multi-class SVM)을 구축함으로써 P2P 트래픽의 식별 및 분류를 수행한다. 그러나 SVM을 이용하여 다중 클래스 SVM을 설계할 경우, 학습 시간이 많이 소요될 뿐만 아니라 학습데이터의 질에 따라 시스템의 성능이 차이가 나는 등 여러 가지 문제점을 가지고 있다. 따라서 현실적으로 실무에 적용하기 어렵다[9]. 특히, 인터넷 애플리케이션 트래픽 분류의 경우 각 트래픽 유형별 데이터의 크기는 차이가 있을 수 있다. 결국 학습 데이터 크기의 불균형으로 인하여 학습 시, 한 트래픽 유형의 학습 결과가 다른 트래픽 유형의 데이터로부터 영향을 받을 가능성이 높다. 또한 한 트래픽 클래스에 속하는 새로운 트래픽 유형들이 계속적으로 생성될 수 있기에 현재의 학습 데이터가 클래스 전체를 대표한다고 말하기도 어렵다. 따라서 이진 분류기 SVM은 관측되지 않은 영역을 포함하여 결정 경계면을 생성함으로써 새로운 학습 데이터에 대해서 오분류할 가능성이 크다. 그러므로 해당 클래스만을 독립적으로 표현하는 단일 클래스 분류기(one-class SVM)로서 결정 경계면을 선택하는 것이 보다 유리하다.

본 논문에서는 위에서 언급된 두 번째 기법을 계승, 발전

시킨 보다 성숙한 모델을 제안하는 차원에서 출발하여, SVM을 기반으로 한 새로운 계층적 인터넷 애플리케이션 트래픽 분류 시스템을 제안한다. 제안된 시스템은 이진 분류기인 SVM과 단일클래스 SVM의 대표적인 모델인 SVDD(support vector data description)을 계층적으로 결합한 새로운 트래픽 분류 모델로써, 학내 인터넷 트래픽의 실시간 수집 및 분석이 가능한 시스템(KU-MON)으로부터 수집된 양방향 플로우 데이터에 대한 속성 부분집합의 선택 방법을 사용하여 특징선택 및 축소를 실시한 후, 이진 분류기인 SVM으로 P2P 트래픽과 non-P2P 트래픽을 빠르게 분류하는 첫 번째 계층, 3개의 SVDD를 기반으로 P2P 트래픽들을 파일공유, 메신저, TV로 분류하는 두 번째 계층, 그리고 16개의 애플리케이션 트래픽별로 16개의 SVDD를 사용하여 세분화 분류하는 세 번째 계층으로 구성된다. 본 논문에서 제안하는 인터넷 애플리케이션 트래픽 분류 시스템은 그 구조의 성격상 실시간으로 P2P와 non-P2P 트래픽을 분류함으로써 보다 안정적인 네트워크 환경을 지원할 수 있을 뿐만 아니라, P2P 트래픽의 대표적인 3가지 유형별 분류가 가능함으로써 보다 적합한 QoS의 보장이 가능하다. 또한 트래픽의 세분화된 분류로 원활한 대역폭(bandwidth)의 사용 및 효율적 시스템 자원관리의 지원도 가능하다. 더욱이, 새로운 애플리케이션 트래픽이 추가될 때, 전체 시스템의 재학습이 아닌 해당 애플리케이션 트래픽 클래스에 해당하는 모듈만을 추가 학습하는 점증적 갱신이 가능하다.

본 논문의 구성은 다음과 같다. 2장에서는 애플리케이션 트래픽의 분류를 위한 고려사항들을 서술하고, 3장에서는 본 논문에서 제안하는 SVM 기반의 계층적 애플리케이션 트래픽 분류 모델에 대해 기술한다. 4장에서는 실험결과 및 성능 분석을, 마지막으로 5장에서는 결론 및 향후 연구과제에 대해 논한다.

2. 애플리케이션 트래픽 분류를 위한 고려 사항

애플리케이션 트래픽의 분류를 위해서는 다음의 두 가지 사항을 고려해야 한다. 첫째, 애플리케이션 트래픽 분류의 기준을 규정하는 것이다. 응용레벨 프로토콜을 기준으로 분류할 것인지, 스트리밍이나 파일 다운로드와 같이 트래픽의 타입을 기준으로 할 것인지, 아니면 개별 애플리케이션 트래픽으로 할 것인지에 대한 선택이다. 둘째, 애플리케이션 트래픽 분류를 위한 트래픽 데이터의 처리 단위를 결정해야 한다. IP 네트워크에서 트래픽 데이터의 최소 단위인 패킷을 학습과 테스트의 단위로 할 수도 있고, 패킷의 집합인 플로우를 처리 단위로 할 수도 있다. 또한, 처리 단위를 결정함에 있어 각각의 분류 기준에 적합한 트래픽의 특징이 잘 나타나도록 특징 값들을 결정해야 한다.

애플리케이션 트래픽 분류 규정은 분석 목적에 따라 달라질 수 있지만 본 논문에서는 다음과 같은 원칙으로 애플리케이션 트래픽을 분류한다. 기본적으로 애플리케이션의 관점에서 해당 애플리케이션이 발생한 모든 트래픽을 그 애플리케이션의 트래픽으로 규정한다. 예외적인 사항으로 여러 애플리케이션에서 공유하여 사용하는 공개된 응용레벨 프로토콜의 경우는 응용레벨 프로토콜의 단위로 분리하여 분석한다. 이렇게 분류하는 것은 QoS와 같은 트래픽 제어 시스템의 기준과도 일치함으로써 향후 분석결과를 그대로 또는 최소한의 수정으로 활용할 수 있다.

애플리케이션 트래픽의 정확한 분석을 위해서는 데이터의 처리 단위와 트래픽 레코드로부터 추출할 특징 값들을 결정해야 한다. 먼저 데이터의 처리단위는 많은 연구들[10-13]에서 트래픽 분류의 최소단위로 사용되고 있는 플로우를 사용한다. 본 논문에서 사용되는 플로우는 기존 연구들[14-16]에서 조금씩 다르게 정의하고 있는 플로우의 특징들을 대부분 포괄하는 형태로서, 5-tuple 정보(source IP, source port, destination IP, destination port, protocol number)를 공유하는 양방향 패킷들의 집합으로 정의한다. 양방향 패킷들의 집합으로 정의되는 플로우는 하나의 연결에서 인-플로우와 아웃-플로우의 특징이 서로 다르게 나타나는 데, 그림 1과 같이 TCP의 경우, 클라이언트에서 서버로의 패킷을 인-플로우로, 서버에서 클라이언트로의 패킷들을 아웃-플로우로 결정하여 SYN 패킷에 의한 연결시작에서부터 FIN/RST 패킷이 발생하는 연결 끝 사이에 발생한 모든 양방향 패킷들을 하나의 플로우로 정의한다. 또한 UDP의 경우에는 클라이언트/서버의 관계를 정의할 수 없기 때문에 최초 패킷 발생 시점에서 마지막 패킷이 발생한 시점 사이의 모든 양방향 패킷들을 하나의 플로우로 정의한다.

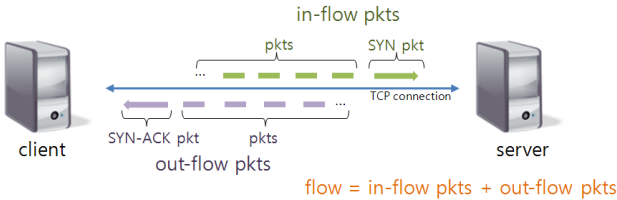


그림 1. TCP 연결의 트래픽 플로우에서 인-플로우, 아웃-플로우

Fig. 1. In-flow and Out-flow in traffic flow of TCP connection

본 논문에서는 표 1과 같이 transport 프로토콜, 인-플로우, 그리고 아웃-플로우에 대한 통계적 정보를 플로우의 특징 값들로 정하였다. 이때, 5-tuple 정보에서의 IP와 포트 정보는 실험에서 수집된 트래픽 데이터의 대부분이 특정 값으로 제한될 가능성[12]으로 인하여 실험 결과가 IP와 포트 정보에 크게 영향을 받을 수 있으므로 플로우의 특징 값에서 제거하였다.

3. 계층적 SVM 기반 인터넷 애플리케이션 트래픽 분류 시스템

본 장에서는 속성 부분집합의 선택에 관한 소개 및 계층적 SVM기반의 인터넷 애플리케이션 트래픽 분류 모델을 소개한다.

3.1 속성 부분집합의 선택

패턴 분류를 위한 효율적인 속성 부분집합의 선택은 중요한 연구 이슈 중 하나이다[17-20]. 특히, 고차원의 패턴 분류 문제에서 데이터의 차원을 줄여 패턴 인식기의 수행 시간을 단축시키고 분류 성능을 향상시키는 것은 필수적 요소이다. 최적의 속성 부분집합 선택은 최초의 속성 집합 D 로부터 거의 사용되지 않거나 중복된 성질을 갖는 특징을 사전에 제거한 속성 부분집합 d 를 찾는 문제이며[17-19],

가능한 적은 성능 저하를 유지하면서 정확한 분류기를 제공하는데 그 목적이 있다. 본 논문에서는 플로우 데이터에 대한 속성 부분집합의 선택 방법 중 그 성능이 이미 검증된 Hall[17]의 방법을 사용한다. 이는 최적우선탐색(best first search) 방법과 속성(attribute or feature) 값 Y 에 대한 엔트로피, 목표 클래스와 속성들 간의 피어슨 상관 계수(Pearson's correlation coefficient)를 이용한 조건부 확률을 계산하여 전체 속성들의 확률 분포도를 가능한 가깝게 표현할 수 있는 최소 개수의 속성집합을 찾는 방법이다.

표 1. 플로우의 특징 집합
Table 1. Feature set of flow

Features		Description	
protocol		TCP or UDP	
in	dPkts	number of packets	
	dOctets	number of bytes	
	timeval(last - first)	flow duration	
	syn	number of SYN packets	
	ack	number of ACK packets	
	rst	number of RST packets	
fin	number of FIN packets		
out	"		"
in	state of packet	min, max, avg, mdev	minimum, maximum value, average, standard deviation
	state of window	"	"
	state of jitter	"	"
out	state of packet	"	"
	state of window	"	"
	state of jitter	"	"

3.2 다중 클래스 SVM

통계적 학습이론에 기반을 둔 SVM은 주어진 문제를 항상 전역적 최적해가 보장되는 convex quadratic problem으로 변환하여 해를 구하기 때문에 패턴인식 분야에서 매우 우수한 성능을 보이고 있다[5-6]. 그러나 이진 분류기라는 SVM의 기능적 한계점으로 인하여, 주어진 문제가 현재 우리가 다루고자 하는 트래픽 분류와 같이 다중 분류 문제에는 SVM을 직접적으로 적용할 수가 없다. 따라서 여러 개의 이진 분류기인 SVM을 유기적으로 결합하여 다중 클래스 SVM을 설계하는 것이 일반적인 연구방법론이다[7-9, 21]. 본 논문에서는 단일 클래스 SVM의 대표적인 알고리즘인 SVDD를 기반으로 설계된 다중 클래스 SVM[9]을 일부 변형하여 인터넷 애플리케이션 트래픽 유형을 분류하는 시스템에 사용한다.

3.3 인터넷 애플리케이션 트래픽 분류 시스템

본 논문에서 제안하는 인터넷 애플리케이션 트래픽 분류

시스템은 총 4개의 모듈로 구성된다. 즉, 2개의 온라인 처리 모듈인 플로우 데이터 수집, 애플리케이션 트래픽 분류 모듈과 2개의 오프라인 처리 모듈인 속성 부분집합 선택, SVM 학습 모듈로 구성된다(그림 2 참조). 1) 플로우 데이터 수집 모듈에서는 실시간으로 수집된 패킷 데이터로부터 트래픽 분류의 최소단위인 플로우를 생성한다. 2) 속성 부분집합 선택 모듈은 각 계층별 분류를 위한 최적의 플로우 속성 집합을 선택함으로써, 전체적인 분류 시스템의 정확도 및 분류 속도를 향상시킨다. 3) SVM 학습 모듈은 각 계층별 분류를 위한 플로우의 속성 집합을 기반으로 학습을 실시한다. 이때, P2P 트래픽과 non-P2P 트래픽 분류에는 이진 분류기인 SVM을, P2P 트래픽들을 파일공유, 메신저, TV로 분류할 때에는 3개의 SVDD를, 각 애플리케이션 단위로 트래픽 분류 시에는 애플리케이션의 개수와 동일한 개수의 SVDD로 학습시킨다. 4) 애플리케이션 트래픽 분류 모듈에서는 SVM 학습 모듈에서 각 계층별로 학습이 완료된 분류 모델을 사용하여 실시간으로 유입되는 플로우의 분류를 수행한다.

그림 3에 도식화된 SVM 기반의 애플리케이션 트래픽 분류 모듈은 플로우 기반의 트래픽 정보를 입력으로 이진 분류 SVM을 이용하여 P2P 트래픽과 non-P2P 트래픽으로 분류하는 첫 번째 계층, P2P 트래픽을 대표적인 3가지 유형인 파일공유, 메신저, TV로 분류하는 두 번째 계층, 그리고 본 논문의 실험에서 사용한 전체 16가지 애플리케이션 트래픽별로 세분화 분류하는 세 번째 계층으로 구성된다. 각 계층별로 속성 부분집합의 선택 방법을 사용하여 특징 선택 및 축소를 각각 실시한다.

테스트 데이터에 대한 실제 트래픽 분류 절차는 다음과 같다: 첫 번째 계층에서는 네트워크 망에서 수집된 트래픽 플로우 정보를 이진 분류기인 SVM을 이용하여 P2P 트래픽과 non-P2P 트래픽으로 빠르게 분류한다. 따라서 보다 효율적인 시스템의 자원관리 및 안전한 네트워크 환경의 지원이 가능하다. 두 번째 계층에서는 P2P 트래픽을 3개의 SVDD를 이용하여 파일공유, 메신저, TV로 분류한다. 따라서 보다 안정적이고 적합한 QoS의 보장 및 과부하 등의 소비를 유발시키는 해당 트래픽 유형의 대역폭을 관리함으로써 효율적인 시스템의 자원관리가 가능하다. 특히 제한적이고 복잡한 포트 회피와 네트워크상에서 혼잡한 상황을 유발시키는 P2P 트래픽을 유형별로 분류하고 관리함으로써 보다 안전한 네트워크 환경의 구축을 지원할 수 있다. 마지막 계층에서는 SVDD를 기반으로 전체 16가지의 트래픽별로 세분화 분류를 수행한다. 각 트래픽별로 배정된 SVDD를 독립적으로 학습함으로써 보다 빠르고 효율적인 학습 및 갱신이 가능하다. 또한, 새로운 애플리케이션 트래픽이 추가되더라도 전체 시스템을 재학습 시킬 필요 없이 새로운 애플리케이션 트래픽 클래스에 해당하는 SVDD만을 추가 학습함으로써 시스템의 점증적 갱신 및 확장에 필요한 비용을 줄일 수 있다. 결과적으로 본 시스템은 데이터마이닝의 개념계층(concept hierarchy)과 같이 네트워크 망 관리자의 관리 목적에 따라 일반화(generalization)와 세분화(specialization) 작업을 통하여 추상화 정도(levels of abstraction)를 조절함으로써 유연한 망 관리가 가능하다.

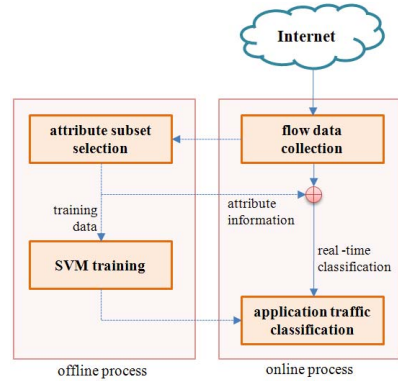


그림 2. 시스템의 전체 구조도
fig. 2. The overall architecture of system

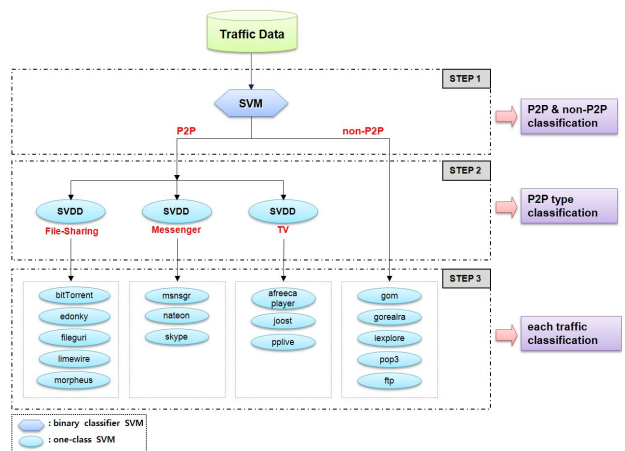


그림 3. 계층적 인터넷 애플리케이션 트래픽 분류 시스템의 구조도

fig. 3. The architecture of hierarchical internet application traffic classification system

4. 실험 결과 및 분석

4.1 데이터 수집 및 데이터 집합

본 실험에서는 KU-MON[16]과 TMA[22]를 기반으로 학내 인터넷 트래픽을 대상으로 분류의 정확성을 평가할 수 있는 검증 네트워크를 구축하였고, 이를 통하여 다양한 트래픽분류 결과들의 분석 및 정확성을 평가한다. 먼저, 데이터 수집에 사용한 KU-MON은 기존의 NG-MON을 개선한 시스템으로서, 학교와 같은 엔터프라이즈 네트워크에서 발생하는 인터넷 트래픽의 실시간 수집 및 분석이 가능한 시스템이다. KU-MON은 아래의 그림 4와 같이 구성되며, 호스트에서 발생하는 패킷을 저장하고 이를 기반으로 플로우 정보로 생성하는 플로우 생성 모듈, 생성한 플로우를 1분 단위로 저장하는 플로우 저장 모듈, 플로우 저장 모듈에 저장된 플로우를 기반으로 네트워크 관리자의 목적에 따라 분석하는 트래픽 분석 모듈로 구성된다.

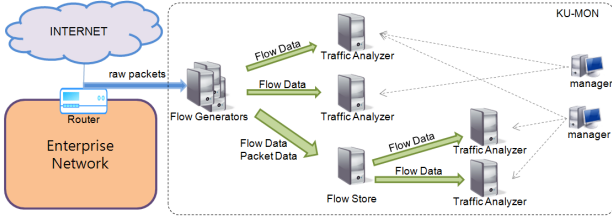


그림 4. KU-MON: 실시간 트래픽 수집 및 분석 시스템
fig. 4. KU-MON: Real-time traffic collection and analysis system

KU-MON에서 생성된 플로우 정보에는 애플리케이션 트래픽의 이름이 포함되어 있지 않다[16]. 이를 해결하기 위하여 본 실험에서는 중단 호스트(end-host)의 TMA에서 수집한 소켓-프로세스 정보와 플로우 생성 모듈에서 패킷 정보로부터 생성한 플로우 정보를 이용하여 정확히 애플리케이션 트래픽의 종류를 알 수 있는 검증 네트워크를 구성하였다[22](그림 5). TMA는 중단 호스트에 설치되는 에이전트 프로그램으로 중단 호스트에서 네트워크 통신을 위해 생성되는 TCP/UDP 소켓의 정보와 이 소켓을 생성한 프로세스 정보를 실시간으로 수집하는 기능을 수행한다. 각각의 중단 호스트에서 수집된 소켓-프로세스 정보는 중앙 수집 서버인 TMS(traffic measurement server)로 전송된다. TMS에서는 각 중단 호스트에서 수집된 프로세스-소켓 정보와 패킷정보로부터 생성된 플로우 정보를 바탕으로 실험에 사용할 플로우의 애플리케이션 이름을 확정하여 정답지 트래픽 데이터를 생성한다.

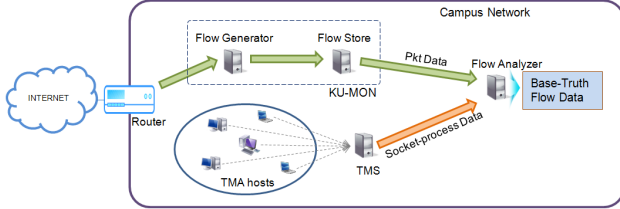


그림 5. 정답지 트래픽 데이터의 생성 방법
fig. 5. The generation method of ground truth traffic

본 논문에서는 아래의 표 2에서와 같이 현재 국내·외에서 가장 널리 사용되고 있는 P2P 트래픽 11가지와 non-P2P 트래픽 5가지를 포함하는 16가지의 트래픽을 선정하였다. 또한 KU-MON 및 TMA를 기반으로 각각 다른 시간대에 수집한 플로우 데이터를 가지고 학습 데이터 셋과 테스트 데이터 셋으로 구성하였으며, 각 애플리케이션 트래픽별로 200개씩의 플로우 데이터를 수집하였다.

표 2. P2P와 non-P2P별 트래픽 종류
Table 2. The P2P and non-P2P traffic types

P2P	File-Sharing	bitTorrent, edonky, fileguri, limewire, morpheus
	Messenger	msnsg, nateon, skype
	TV	afreeca player, joost, pplive
non-P2P		gom, gorealra, iexplore, pop3, ftp

4.2 실험결과 및 분석

애플리케이션 트래픽 분류의 정확도 평가를 위하여 일반적인 평가기준인 recall과 precision을 성능 지표[3]로 사용하였다:

$$recall = \frac{\sum_{i=1}^N (TP)_i}{\sum_{i=1}^N (NT)_i} \times 100 \quad (1)$$

$$precision = \frac{\sum_{i=1}^N (TP)_i}{\sum_{i=1}^N (TN)_i} \times 100 \quad (2)$$

위 식에서 $(NT)_i$ (number of total i)는 i 라는 트래픽 클래스의 전체 원소의 개수, $(TP)_i$ (true positive of i)는 i 클래스 원소 중에서 실제 i 클래스에 속한 원소의 개수, $(TN)_i$ (total number of i)는 i 클래스로 분류된 전체 원소의 개수를 의미한다.

4.2.1 P2P와 non-P2P 분류

첫 번째 실험은 P2P 트래픽과 non-P2P 트래픽을 신속하게 분류하는 실험으로 16가지 트래픽별로 100개씩 랜덤하게 추출한 P2P 트래픽 1,100개와 non-P2P 500개를 이진 분류기 SVM으로 학습하였고, 테스트를 위하여 학습에 참여하지 않은 각 트래픽 별로 100개씩 총 1,600개의 데이터로 테스트하였다. P2P 트래픽과 non-P2P 트래픽의 이진 분류를 위한 최적의 속성 부분집합을 선택하기 위해, java 기반의 기계학습 툴인 weka[23]의 CFS(correlation feature selection)를 사용하였다. CFS에 의해 선택된 속성 부분집합은 {prot, in_fin, in_window_min, in_window_max, in_jitter_avg, out_pkt_max, out_window_min, out_window_avg}이다. 다음의 표 3은 CFS에 의해 선택된 8개의 속성만을 사용한 이진 분류 결과와 전체 39개의 속성을 모두 사용한 이진 분류 결과를 비교한 것이다. 실험 결과, 최적의 속성 부분집합을 사용한 경우가 전체 속성을 모두 사용한 경우보다 분류 정확도가 높은 것으로 나타났다. 이는 분류에 영향을 미치지 못하거나, 오분류를 유발하는 불필요한 속성들이 제거되었기 때문으로 해석된다. 따라서 분류에 사용된 속성을 최소화함으로써 분류 속도를 향상시킬 수 있을 뿐만 아니라, 분류의 정확도까지 향상시킬 수 있음을 보여준다.

표 3. P2P와 non-P2P 분류 성능 측정 표
Table 3. The performance evaluation of classification between P2P and non-P2P

평가 항목	CFS(using 8 features)			using all features		
	σ	recall	precision	σ	recall	precision
P2P	0.88	98.27	96.35	0.8	96.6	95.8
non-P2P		91.8	95.6		90.6	92.4

4.2.2 P2P 트래픽의 유형별 분류

두 번째 실험에서는 SVDD를 이용하여 첫 번째 계층에서 분류한 P2P 트래픽을 대표적인 P2P 트래픽 유형인 파일공유, 메신저, TV로 세분화하였다. 실험을 위하여 5가지 유형의 파일공유 트래픽(bitTorrent, edonky, fileguri,

limewire, morpheus), 3가지 유형의 메신저 트래픽 (msnsg, nateon, skype), 그리고 3가지 유형의 TV 트래픽 (afreeca player, joost, pplive) 등 총 11가지 유형의 P2P 트래픽 데이터를 각 유형별로 랜덤하게 100개씩 추출하였다. 추출된 데이터 집합을 각각의 SVDD(파일 공유 SVDD, 메신저 SVDD, TV SVDD)로 학습하였으며, 학습에 참여하지 않은 데이터들은 테스트 데이터로 사용하였다. CFS를 이용하여 선택된 최적의 속성 부분집합은 {in_dOctets, in_pkt_min, in_pkt_avg, out_pkt_min, out_pkt_avg, out_pkt_mdev, out_jitter_avg}이며, 7개의 속성만을 사용한 분류 결과와 전체 39개의 속성을 모두 사용한 분류 결과는 다음의 표 4와 같다. 여기서 최적 속성 부분집합에서의 조정상수 C는 0.1, 커널 함수인 가우시안 함수의 결정 경계 변수인 σ 값은 파일공유: 0.15, 메신저: 0.43, TV: 0.58로 고정하였다. 실험 결과, 7개의 속성 부분집합의 경우 전체 recall은 95.36%, precision은 95.37%로 측정된 반면에 전체 속성의 경우 recall은 95.09%, precision은 96.0%의 성능을 보였으며, 7개의 속성 부분집합만으로도 P2P의 대표적인 3가지 유형들이 모두 recall과 precision에서 만족스러운 성능을 보여주고 있음을 확인하였다.

표 4. P2P 유형별 성능 측정 표

Table 4. The performance evaluation of P2P traffic classification

평가 항목	CFS(using 7 features)			using all features		
	σ	recall	precision	σ	recall	precision
파일공유	0.15	97.0	95.7	0.12	97.2	97.8
메신저	0.43	94.3	94.3	0.4	95.3	96.3
TV	0.58	93.7	95.9	0.48	91.3	92.9

4.2.3 전체 트래픽 분류

세 번째 실험은 상위계층에서 분류된 세 가지 유형의 P2P 트래픽과 non_P2P 트래픽을 다시 16가지 유형의 트래픽으로 세분화하는 실험이다. 각 트래픽별로 랜덤하게 데이터를 100개씩 추출하여 각각의 유형별 SVDD로 학습하였으며, 학습에 참여하지 않은 데이터는 테스트 데이터로 사용하였다. CFS에 의해 선택된 최적 속성 부분집합은 {in_pkt_max, in_pkt_avg, in_window_min, in_window_max, out_dOctets, out_pkt_min, out_pkt_max, out_pkt_avg, out_jitter_max}이며, 9개의 속성만을 이용한 분류 결과와 전체 39개의 속성을 모두 이용한 분류 결과는 다음의 표 5와 같다. 여기서 조정상수 C는 0.1, σ 값은 각 트래픽별로 고정하였다(표 5 참조). 실험 결과, 9개의 속성 부분집합의 경우 전체 recall은 88.88%, precision은 89.07%로 측정된 반면에 전체 속성의 경우 recall은 88.19%, precision은 92.09%의 성능을 보였으며, 9개의 속성만으로도 만족스러운 성능을 보여주고 있음을 확인하였다. 특히 morpheus 트래픽의 recall이 상대적으로 낮다는 것을 확인할 수 있으며, 이는 morpheus 트래픽이 성격이 유사한 fileguri 트래픽으로 오분류 되었기 때문이다. 따라서 fileguri 트래픽의 recall은 높지만, precision은 recall에 비해 상대적으로 떨어진다.

표 5. 전체 애플리케이션 트래픽별 성능 측정 표

Table 5. The performance evaluation of each subsidiary types of application traffic

application	평가 항목	CFS(using 9 features)			using all features		
		σ	recall	precision	σ	recall	precision
bitTorrent		0.35	98	96.1	0.3	94	95.9
edonky		0.6	88	90.7	0.65	82	98.8
fileguri		0.05	100	90.1	0.07	100	79.4
limewire		0.58	88	85.4	0.65	94	96.9
morpheus		0.6	71	81.6	0.6	64	100
msnmsgr		0.55	85	75.9	0.5	84	94.4
nateon		0.55	83	79.1	0.6	88	91.7
skype		0.45	91	94.8	0.4	84	95.5
afreecaplayer		0.65	88	85.4	0.6	89	90.8
joost		0.5	84	85.7	0.6	78	96.3
pplive		0.5	86	89.6	0.5	84	93.3
gom		0.3	94	91.3	0.25	90	90.1
gorealra		0.55	93	98.9	0.5	96	86.5
iexplore		0.45	84	81.6	0.55	89	91.8
pop3		0.5	93	98.9	0.5	97	86.6
ftp		0.35	96	100	0.3	98	85.5

4. 결 론

본 논문에서는 인터넷 애플리케이션 트래픽의 분류가 네트워크 망 관리시스템에서 가장 중요한 기본 기능 중 하나라는 문제의식과 기존의 전통적인 분류 방법으로 대표되는 포트 번호 및 페이로드 정보를 이용하는 방법의 구조적 문제점을 극복하는 차원에서 SVM에 기초한 새로운 인터넷 애플리케이션 트래픽 분류 시스템을 제안하였다. 제안된 시스템은 학내 인터넷 트래픽의 실시간 수집 및 분석이 가능한 시스템으로부터 수집된 플로우 기반의 트래픽 정보에 대한 속성 부분집합의 선택 방법을 사용하여 특징선택 및 축소를 실시하고, 인터넷 애플리케이션 트래픽을 coarse 혹은 fine하게 분류함으로써 효율적인 시스템의 자원 관리, 안정적인 네트워크 환경의 지원, 원활한 대역폭의 사용, 그리고 적절한 QoS를 보장하였다. 또한, 새로운 애플리케이션 트래픽이 추가되더라도 전체 시스템을 재학습 시킬 필요 없이 새로운 애플리케이션 트래픽만을 추가 학습함으로써 시스템의 점증적 갱신 및 확장성에도 기여하였다. 평가항목인 recall과 precision에서 만족스러운 수치 등을 실험을 통하여 확인함으로써 제안된 시스템의 성능을 검증하였다.

향후 연구 과제로는 인터넷 애플리케이션 트래픽 분류에 관한 메카니즘에 내재되어 있는 유용한 지식의 발견과 분석에 관한 연구를 수행하고자 한다.

참 고 문 헌

- [1] H. Schulze and K. Mochalski, Ipoque Internet Study 2008/2009, <http://www.ipoque.com/>.
- [2] G. Szabo, I. Szabo, and D. Orincsay, "Accurate Traffic Classification," *IEEE Int. Symposium on World of Wireless Mobile and Multimedia*

- Networks*, pp. 1-8, 2007.
- [3] J. Erman, A. Mahanti, and M. Arlitt, "Internet Traffic Identification using Machine Learning," *IEEE Conf. on Global Telecommunications*, pp. 1-6, 2006.
- [4] T. Auld, A. Moore, and S. Gull, "Bayesian Neural Networks for Internet Traffic Classifications," *IEEE Trans. on Neural Networks*, Vol. 18, No. 1. pp. 223-239, 2007.
- [5] Y. Liu, R. Wang, H. Huang, Y. Zeng, and H. He, "Applying Support Vector Machine to P2P Traffic Identification with Smooth Processing," *IEEE Int. Conf. on Signal Processing*, Vol. 3, pp.16-20, 2006.
- [6] F. J. Gonzalez-Castano, P. S.Rodriguez-Hernandez R. P. Martinez-Alvarez, A. Gomez, I. Lopez-Cabido, and J. Villasuso-Barreiro, "Support Vector Machine Detection of Peer-to-Peer Traffic," *IEEE Int. Conf. on Computational Intelligence for Measurement Systems and Applications*, pp. 103-108, 2006.
- [7] A. Yang, S. Jiang, and H. Deng, "A P2P Network Traffic Classification Method using SVM," *The 9th Int. Conf. for Young Computer Scientists*, pp. 398-403, 2008.
- [8] X. Zhou, "A P2P Traffic Classification Method Based on SVM," *Int. Symposium Computer Science and Computational Technology*, pp. 53-57, 2008.
- [9] H. Lee, J. Song, and D. Park, "Intrusion Detection System Based on Multi-class SVM," *LNAI*, 3642, pp. 511-519, 2005.
- [10] M. Tai, S. Ata, and I. Oka, "Fast, Accurate, and Lightweight Real-Time Traffic Identification Method Based on Flow Statistics," *LNCS*, 4427, pp. 255-259, 2007.
- [11] T. Karagiannis, K. Papagiannaki, and M. Faloutsos, "BLINC: Multilevel Traffic Classification in the Dark," *In Proc. of ACM SIGCOMM*, Vol. 35, No.4, pp. 229-240, 2005.
- [12] J. Li, S. Zhang, S. Liu, and Y. Xuan, "Active P2P Traffic Identification Technique," *IEEE Int. Conf. on Computational Intelligence and .Security*, pp. 37-41, 2007.
- [13] G. Zhang, G. Xie, J. Yang, Y. Min, Z. Zhou, and X. Duan, "Accurate Online Traffic Classification with Multi-phases Identification Methodology," *IEEE Int. Conf. on Consumer Communications and Networking*, pp. 141-146, 2008.
- [14] P. Phaal, S. Panchen, and N. McKee, InMon Corporation's sFlow: A Method for Monitoring Traffic in Switched and Routed Networks, IETF RFC 3176, 2001.
- [15] Cisco Systems, White Papers, NetFlow Services and Applications, http://www.cisco.com/warp/public/cc/pd/iosw/ioft/neflct/tech/napps_wp.htm.
- [16] S. Han, M. Kim, H. Ju, and J. W. Hong, "The Architecture of NG-MON: A Passive Network Monitoring System," *LNCS*, 2506, pp. 16-27, 2002.
- [17] M. Hall, Correlation-based Feature Selection for Machine Learning, PhD Diss. Department of Computer Science, Waikato University, Hamilton, NZ, 1998.
- [18] I. Seok, J. Lee, and B. Moon, "Hybrid Genetic Algorithms for Feature Selection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 26, No. 11, pp. 1424-1437, 2006.
- [19] F. Fleuret, "Fast Binary Feature Selection with Conditional Mutual Information," *Journal of Machine Learning Research*, Vol. 5, pp. 1531-1555, 2004.
- [20] Y. Sun and J. Li, "Iterative RELIEF for Feature Weighting," *In Proc. of the 23rd Int. Conf. on Machine Learning*, pp. 913-920, 2006.
- [21] T. Ambwani, "Multi Class Support Vector Machine Implementation to Intrusion Detection," *In Proc. of the Int. Conf. on Neural Networks*, Vol. 3, pp.2300-2305, 2003.
- [22] B. Park, Y. Won, M. Kim, and Hong, J. W. Hong, "Towards Automated Application Signature Generation for Traffic Identification," *Network Operations and Management Symposium*, pp. 160-167, 2008.
- [23] Machine Learning Lab in The University of Waikato, <http://www.cs.waikato.ac.nz/ml..>

저 자 소 개



유재학(Jaehak Yu)

2001년 : 건국대학교 전산과학과 학사
 2003년 : 고려대학교 전산학과 석사
 2010년 2월 : 고려대학교 전산학과 박사
 2006년 3월~2008년 2월 : 고려대학교
 컴퓨터정보학과 초빙전임강사
 2008년 3월~현재 : 고려대학교 컴퓨터 정
 보학과 강사

관심분야 : 데이터마이닝, 기계학습, 네트워크 마이닝, 침입탐지
 E-mail : dbzzang@korea.ac.kr



이한성(Hansung Lee)

1996년 : 고려대학교 전산학과 학사
 1996년~1999년 : (주)대우엔지니어링
 2002년 : 고려대학교 전산학과 석사
 2008년 : 고려대학교 전산학과 박사
 2006년 3월~2007년 2월 : 고려대학교
 컴퓨터정보학과 초빙전임강사
 2008년 9월~2009년 2월 : 고려대학교
 BK21 연구교수

2009년 11월~현재: 한국전자통신연구원

관심분야: 멀티미디어 마이닝, 휴먼인식, 네트워크 마이닝, 기계학습, 지능 데이터베이스

E-mail : mohan@etri.re.kr



임영희(Younghee Im)

1994년: 고려대학교 전산학과 학사
1996년: 고려대학교 전산학과 석사
2001년: 고려대학교 전산학과 박사
2001년~2003년: 대전대학교 컴퓨터
정보통신공학부 강의전담교수
2003년~현재: 고려대학교 컴퓨터정보학과
강사

관심분야: 인공지능, 상황인지, 정보검색, 텍스트마이닝,
데이터마이닝

E-mail : yheeim@korea.ac.kr



김명섭(Myung-Sup Kim)

1998년: 포항공과대학교 전자계산학과 학사
2000년: 포항공과대학교 컴퓨터공학과 석사
2004년: 포항공과대학교 컴퓨터공학과 박사
2004년~2006년: Post-Doc., Dept. of ECE,
Univ. of Toronto,
Canada.

2006년~현재: 고려대학교 컴퓨터정보학과
조교수

관심분야: 네트워크 관리 및 보안, 트래픽 모니터링 및 분석,
멀티미디어 네트워크

E-mail : tmskim@korea.ac.kr



박대희(Daihee Park)

1982년: 고려대학교 수학과 학사
1984년: 고려대학교 수학과 석사
1989년: 플로리다 주립대학 전산학과 석사
1992년: 플로리다 주립대학 전산학과 박사
1993년~현재: 고려대학교 컴퓨터정보학과
교수

관심분야: 지능 데이터베이스, 데이터마이닝, 인공지능, 퍼
지이론

E-mail : dhpark@korea.ac.kr