

계층적 다중 클래스 SVM을 이용한 인터넷 애플리케이션 트래픽 분류

유재학[○], 김성윤, 이한성, 김명섭, 박대희

고려대학교 컴퓨터정보학과

{dbzzang[○], adayslife, mohan, tmskim, dhpark}@korea.ac.kr

Internet Application Traffic Classification using a Hierarchical Multi-class SVM

Jaehak Yu[○], Sungyun Kim, Hansung Lee, Myungsup Kim, Daihee Park

Dept. of Computer & Information Science, Korea University

요 약

P2P를 포함하는 인터넷 애플리케이션 트래픽의 보다 빠르고 정확한 분류는 최근 학계의 중요한 이슈 중 하나이다. 본 논문에서는 기존의 전통적인 분류방법으로 대표되는 port 번호 및 payload 정보를 이용하는 방법론의 구조적 한계점을 극복하는 새로운 대안으로써, 이진 분류기인 SVM과 단일클래스 SVM을 계층적으로 결합한 다중 클래스 SVM을 구축하여 인터넷 애플리케이션 트래픽 분류를 수행하였다. 제안된 시스템은 이진 분류기인 SVM으로 P2P 트래픽과 non-P2P 트래픽을 빠르게 분류하는 첫 번째 계층, 3개의 단일클래스 SVM을 기반으로 P2P 트래픽들을 파일공유, 메신저, TV로 분류하는 두 번째 계층, 그리고 전체 16가지의 애플리케이션 트래픽별로 세분화 분류하는 세 번째 계층으로 구성된다. 제안된 시스템은 flow 기반의 트래픽 정보를 수집하여 인터넷 애플리케이션 트래픽을 coarse 혹은 fine하게 분류함으로써 효율적인 시스템의 자원 관리, 안정적인 네트워크 환경의 지원, 원활한 bandwidth의 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.

1. 서 론

인터넷의 급속한 발전과 이를 이용하는 사용자의 수가 급증함에 따라 보다 다양하고 새로운 네트워크 서비스들이 개발되어 상용화되고 있다. 따라서 다양한 네트워크 서비스에 보다 적합한 Quality of Service (QoS) 및 안전한 네트워크 환경의 제공을 목적으로 하는 네트워크 망 관리 시스템에서, 인터넷 애플리케이션 트래픽 분류는 가장 중요한 기본 기능 중 하나이다. 특히 네트워크 서비스 중, Peer-to-Peer (P2P) 관련 트래픽은 2004년 CacheLogic사의 조사에 의하면, 전체 인터넷 트래픽에서의 비중이 60%에 이르고 있으며 현재는 그 비중이 보다 커지고 있다 [1-2]. 인터넷 총 트래픽의 대부분을 차지하는 P2P는 대용량의 파일을 포함하는 엄청난 트래픽과 대칭적 성격으로 인하여 네트워크상에서 혼잡한 상황을 유발할 뿐만 아니라, 고비용의 하부구조 업그레이드 및 네트워크 세분화 등의 추가적 비용을 요구하고 있다. 따라서 P2P를 포함하는 인터넷 애플리케이션 트래픽의 보다 빠르고 정확한 분류가 최근 학계의 중요한 이슈 중 하나이다 [2-4].

인터넷 애플리케이션 트래픽을 분류하는 전통적인 방법론은 port 번호를 이용하는 방법과 payload 정보를 이용

하는 방법으로 나눌 수 있다 [2-3]: 1) port 번호 기반의 분류방법은 IANA에서 할당한 well-known port 번호를 분석하는 비교적 단순하면서도 실용적인 방법이지만, 동적으로 port 번호를 할당하여 packet을 발생하거나 기존에 사용한 well-known port 번호를 다른 목적으로 이용하는 최근의 애플리케이션들이 증가함에 따라 정확한 트래픽 분류가 어렵다; 2) payload 정보에 기반 한 분류방법은 payload의 특징을 추출하고 이를 packet과 비교하는 방법으로, packet을 발생할 때 payload를 암호화할 경우 이를 사용할 수 없다. 또한 최근에 불법침입의 차단 및 고비용의 접근 등을 이유로 payload의 접근을 금지하고 있는 실정이 본 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 [3-6]. 기계학습 기법은 동적으로 변하는 port 번호와 암호화된 payload에 독립적인 데이터의 feature vector로부터 교사학습 (supervised learning) 혹은 비교사학습 (unsupervised learning)을 통하여 중요한 패턴들을 찾아낸다는 점이 최근의 애플리케이션 변화에 대처할 수 있음을 시사한다. 또한 네트워크 트래픽 데이터의 성격이 대용량의 스트림

* 본 연구는 산업자원부 및 한국산업기술평가원의 성장동력 기술개발사업의 연구결과로 수행되었습니다

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 support vector machine (SVM)을 인터넷 애플리케이션 트래픽 분류에 적용하려는 연구가 주목을 받고 있다 [5-6]. 현재까지의 SVM을 이용한 인터넷 애플리케이션 트래픽 분류는 패턴 분류 및 함수 근사 (function approximation) 등의 문제에서 매우 우수한 성능을 보이는 SVM을 사용하여 인터넷 애플리케이션 트래픽 분류 문제에 적용하는 가능성을 검증하는 비교적 초기의 시도로써 다음의 두 가지 방법론을 취하고 있다: 첫 번째 방법 [5]에서는 이진 분류기 (binary classifier)인 SVM을 이용하여 P2P 트래픽과 non-P2P 트래픽을 단순히 이진 분류하는 방법을 취하고 있으며, 두 번째 방법 [6]에서는 다중 클래스 SVM (multi-class SVM)을 구축함으로써 P2P 트래픽의 identification 및 트래픽의 분류를 수행한다.

본 논문에서는 위에서 언급된 두 번째 기법을 계승, 발전시킨 보다 성숙한 모델을 제안하는 차원에서 출발하여, SVM을 기반으로 한 새로운 계층적 인터넷 애플리케이션 트래픽 분류 시스템을 제안한다. 제안된 시스템은 이진 분류기인 SVM과 단일클래스 SVM (one-class SVM)의 대표적인 모델인 support vector data description (SVDD)을 계층적으로 결합한 새로운 트래픽 분류 모델로써, 이진 분류기인 SVM으로 P2P 트래픽과 non-P2P 트래픽을 빠르게 분류하는 첫 번째 계층, 3개의 단일클래스 SVM을 기반으로 P2P 트래픽들을 파일공유 (file-sharing), 메신저, TV로 분류하는 두 번째 계층, 그리고 전체 16가지 애플리케이션 트래픽별로 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 보다 원활한 bandwidth의 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 갱신 (incremental updating)이 가능하다.

본 논문의 구성은 다음과 같다. 2장에서는 본 논문에서 제안하는 SVM 기반의 계층적 인터넷 애플리케이션 트래픽 분류 모델에 대해 기술한다. 3장에서는 실험결과 및 성능 분석을 기술하며, 마지막으로 4장에서는 결론 및 향후 연구과제에 대해 논한다.

2. 계층적 SVM 기반의 인터넷 애플리케이션 트래픽 분류 시스템

본 장에서는 P2P와 non-P2P 분류를 위한 SVM의 기본 개념과 각 트래픽을 세분화 분류하는 다중 클래스 SVM의 기본 요소인 단일 클래스 SVM을 소개하고, 이를 주요 구성요소로 하는 새롭게 제안된 계층적 SVM기반의 인터넷 애플리케이션 트래픽 분류 모델을 소개한다.

2.1 이진 클래스 SVM

통계적 학습이론 (statistical learning theory)에 기반을 둔 SVM은 주어진 문제를 항상 전역적 최적해가 보장되는 convex quadratic problem으로 변환하여 해를 구하기 때문에 패턴인식 분야에서 매우 우수한 성능을 보여주고 있다 [5-6]. SVM의 기본 원리는 선형 분리 (linearly separable)가 가능한 문제에서부터 출발한다. d -차원에서 입력데이터 x_i 가 주어졌을 때 학습 데이터의 출력으로 $\{-1, +1\}$ 처럼 이진 값으로 구분되는 문제를 고려한다 [5]. 두 집합을 분류하기 위한 모델을 정의하기 위하여 그림 1과 같은 선형 식별함수인 초평면 (hyperplane)을 정의할 수 있다.

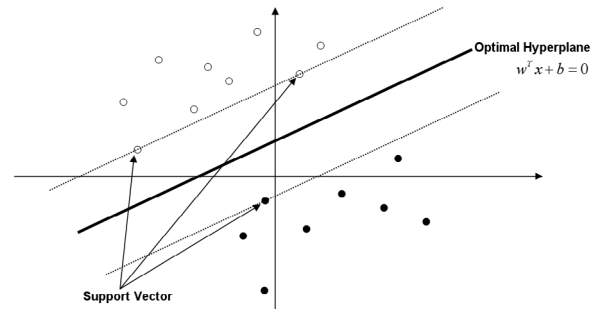


그림 1. 선형 분리가 가능한 최적 초평면과 support vector.

최적 경계 초평면이 학습 데이터의 분류 함수로 주어질 때, 초평면에서 학습 데이터와의 거리 r 은 $r = \frac{|w^T x + b|}{\|w\|} \geq \frac{1}{\|w\|}$ 의 관계가 되고, 이 하한 값의 거리에 있는 데이터는 최적 경계 초평면과 가장 가까운 거리에 위치하게 된다. 이 데이터들을 support vector라고 부른다. 따라서 최적 경계 초평면에 의해 분류되는 두 클래스간의 거리는 $\rho = 2r = \frac{2}{\|w\|}$ 가 되고, 이때 ρ 를 분류 한도 (margin of separation)라 정의한다.

최적의 초평면을 구하는 문제는 margin을 최대화하는 문제로 정의될 수 있으며, 이 경우 SVM의 학습은 다음의 최적화 문제로 정의된다.

$$\text{minimize } \phi(w) = \frac{1}{2} \|w\|^2$$

$$\text{subject to } d_i(w^T x_i + b) \geq 1 \quad \text{for } i=1, \dots, N \quad (1)$$

식(1)에 관한 쌍대 문제(dual problem)를 구하기 위하여 라그랑제 함수 (Largrange function) L 을 도입한다.

$$L(w, b, \alpha) = \frac{1}{2} w^T w + \sum_{i=1}^N \alpha_i [1 - d_i(w_i^T x_i + b)] \quad (2)$$

학습 종료 후, 주어진 테스트 데이터가 어떤 클래스인지의 소속 여부를 판정하는 결정함수 f 는 다음과 같이 정의된다.

$$f(x) = \text{sgn} \left(\sum_{i=1}^N y_i \alpha_i k(x, x_i) + b \right) \quad (3)$$

2.2 다중 클래스 SVM

이진 분류기라는 SVM의 기능적 한계점으로 인하여, 주

이진 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 연구 방법론이다 [7-8]. 그러나 SVM을 이용하여 다중 클래스 SVM을 설계할 경우, 각 SVM은 관측되지 않은 영역을 포함하여 결정 경계면을 생성함으로써 새로운 데이터에 대하여 오분류 (misclassification)할 가능성이 높다. 그러므로 해당 클래스만을 독립적으로 표현하는 단일 클래스 분류기 (one-class SVM)로서 결정 경계면을 선택하는 것이 다중 클래스 SVM의 설계 시 보다 유리하다. 따라서 본 논문에서는 단일 클래스 SVM의 대표적인 알고리즘인 support vector data description (SVDD) [8]을 기반으로 다중 클래스 SVM를 설계하여 인터넷 애플리케이션 트래픽 유형들을 분류하는 새로운 시스템을 제안한다.

d -차원의 입력공간상에 존재하는 K -개의 데이터 집합 $D_k = \{x_i^k \in R^d \mid i=1, \dots, N_k; k=1, \dots, K\}$ 이 주어졌을 경우, 각각의 클래스를 분류하기 위한 분류기는 각 클래스의 학습 데이터를 최대한 많이 포함하면서 동시에 체적을 최소화하는 구체 (sphere)를 구하는 문제로 정의되며, 다음의 최적화 문제를 통하여 수식화 된다.

$$\min L_0(R_k^2, a_k, \xi_k) = R_k^2 + C \sum_{i=1}^{N_k} \xi_i^k \quad (4)$$

$$s.t. \|x_i^k - a_k\|^2 \leq R_k^2 + \xi_i^k, \xi_i^k \geq 0, \forall i.$$

여기에서, a_k 는 k -번째 클래스를 표현하는 구체의 중심이며, R_k^2 은 구체 반경의 제곱, ξ_i^k 는 k -번째의 클래스에 속한 i -번째 학습 데이터 x_i^k 가 구체에서 벗어나는 정도를 나타내는 벌점 항이며, C 는 상대적 중요성을 조정하는 상수 (trade-off constant)이다.

학습 종료 후 적용 과정에서, 각 클래스의 결정함수는 다음과 같이 정의된다.

$$f_k(x) = R_k^2 - \left[1 - 2 \sum_{i=1}^{N_k} \alpha_i^k k_k(x_i^k, x) + \sum_{i=1}^{N_k} \sum_{j=1}^{N_k} \alpha_i^k \alpha_j^k k_k(x_i^k, x_j^k) \right] \geq 0 \quad (5)$$

서로 다른 특징 공간상에서 정의되는 SVDD의 출력 $f_k(x)$ 값은 각 클래스의 특징 공간상의 경계로부터 해당 테스트 데이터와의 절대 거리를 의미함으로써, 서로 다른 특징 공간상의 절대 거리를 비교하여 소속 클래스를 결정하는 것은 바람직하지 않다. 따라서 특징 공간상의 절대 거리 $f_k(x)$ 를 특징 공간상에서 정의되는 구형체의 반경 R_k 로 나눔으로써 상대적 거리 $\hat{f}(x) = f_x(x)/R_k$ 를 계산하고, 상대거리가 가장 큰 클래스를 입력 데이터 x 의 소속 클래스로 결정한다.

$$\text{Class of } x \equiv \arg \max_{k=1, \dots, K} \hat{f}_k(x)$$

$$\equiv \arg \max_k \left[\left\{ R_k^2 - \left(1 - 2 \sum_{i=1}^{N_k} \alpha_i^k k_k(x_i^k, x) + \sum_{i=1}^{N_k} \sum_{j=1}^{N_k} \alpha_i^k \alpha_j^k k_k(x_i^k, x_j^k) \right) \right\} / R_k \right] \quad (6)$$

2.3 인터넷 애플리케이션 트래픽 분류 시스템

그림 2에 도식화된 SVM 기반의 계층적 인터넷 애플리케이션 트래픽 분류 시스템은 flow 기반의 트래픽 정보를 입력으로 P2P 트래픽과 non-P2P 트래픽을 분류하는 이진 분류기 SVM 계층, P2P 트래픽을 대표적인 3가지 유형인 파일공유 (file-sharing), 메신저, TV로 분류하는 계층, 그리고 전체 16가지 애플리케이션 트래픽별로 세분화 분류하는 세 번째 계층으로 구성된다. 테스트 데이터에 대한 실제 트래픽 분류 절차는 다음과 같다: 첫 번째 계층에서는 네트워크 망에서 수집된 트래픽 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 SVDDs를 이용하여 파일공유, 메신저, TV로 분류한다. 따라서 보다 안정적이고 적합한 QoS의 보장 및 과부하 등의 소비를 유발시키는 해당 트래픽 유형의 bandwidth을 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 데이터마ining의 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. 아래의 그림 2는 본 논문에서 제안하는 SVM과 SVDD를 계층적으로 결합한 인터넷 애플리케이션 트래픽 분류 시스템의 구성도이다.

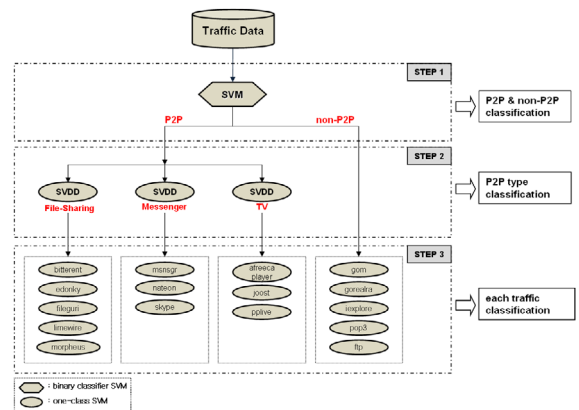


그림 2. 계층적 인터넷 애플리케이션 트래픽 분류 시스템의 구조도

3. 실험 결과 및 분석

3.1 데이터 수집 및 데이터 집합

본 논문에서는 각각 다른 시간대에 수집한 트래픽 데이

터를 이용하여 학습 데이터 셋과 테스트 데이터 셋을 구성하였고, 트래픽별로 데이터 셋의 크기는 일정하게 200 개씩 수집하였다. 데이터 셋을 구성하는 레코드는 양방향 flow 정보를 기반으로 하고 있다. 본 논문에서는 flow의 정의를 패킷의 5-tuple 정보(source IP, source port, destination IP, destination port, protocol number)를 공유하는 양방향 패킷들의 집합으로 정의한다. 즉 5-tuple 정보는 동일한 단방향 패킷들과 이의 역방향 패킷들의 합이다. TCP의 경우 SYN 패킷에 의한 연결시작에서부터 FIN 패킷이 발생하는 연결의 끝 사이에 발생한 모든 양방향 패킷들이 하나의 flow가 되고, UDP의 경우는 최소 패킷 발생 시점에서 마지막 패킷이 발생한 시점 사이의 모든 양방향 패킷들이 하나의 flow가 된다. 또한 실험 데이터로써 아래의 [표 1]에서와 같이 현재 국내·외에서 가장 널리 사용되고 있는 P2P 트래픽 11가지와 non-P2P 트래픽 5가지를 포함하는 16가지의 트래픽을 선정하였다.

[표 1] P2P와 non-P2P별 트래픽 종류

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

아래의 [표 2]에서는 본 논문의 실험에서 사용된 flow 기반의 feature set들을 정리하였다.

[표 2] 실험에 사용된 flow 기반의 features

Features		Description
port type		IP protocol (TCP, UDP,...)
in	dPkts	packets sent in duration
	dOctets	octets sent in duration
	timeval(last - first)	time interval
	syn	#of SYN packet in flow
	ack	#of ACK packet in flow
out	rst	#of RST packet in flow
	fin	#of FIN packet in flow
out	"	"
in	state of packet	min, max, avg, mdev
	state of window	"
	state of jitter	"
out	state of packet	"
	state of window	"
	state of jitter	"

3.2 실험결과 및 분석

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

$$recall = \frac{\sum_{i=1}^N (TP)_i}{\sum_{i=1}^N (NT)_i} \times 100 \quad (7)$$

$$precision = \frac{\sum_{i=1}^N (TP)_i}{\sum_{i=1}^N (TN)_i} \times 100 \quad (8)$$

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.

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 테스트 하였다. 실험 결과 만족스러운 정확도 결과를 얻었으며, 실험결과는 [표 3]에 정리하였다.

[표 3] P2P와 non-P2P 분류 성능 측정 표

평가 항목	recall	precision
P2P	96.6	95.8
non-P2P	90.6	92.4

두 번째 실험은 11가지 P2P 트래픽을 대표적 유형인 파일공유, 메신저, TV로 분류하는 실험으로써, 파일공유 클래스에는 5가지의 P2P 트래픽 (bitterrent, edonky, fileguri, limewire, morpheus), 메신저 클래스는 3가지 P2P 트래픽 (msnsg, nateon, skype), 그리고 TV 클래스는 3가지 P2P 트래픽 (afreeca player, joost, pplive)으로 구성하였다. 각 트래픽별로 랜덤하게 100개씩 데이터를 추출하여 각각의 SVDD로 학습하였으며, 학습에 참여하지 않은 데이터들을 테스트 데이터로 사용하였다. 실험결과는 [표 4]에 정리하였다. 여기서 조정상수 C 는 0.1, 커널 함수인 가우시안 함수의 상수 σ 값은 파일공유: 0.12, 메신저: 0.4, TV: 0.48로 고정하였다. [표 4]에서 보는 바와 같이 P2P의 대표적인 3가지 유형들이 모두 recall과 precision에서 만족스러운 성능을 보여주고 있음을 확인하였다.

[표 4] P2P 유형별 성능 측정 표

평가 항목	recall	precision
type (σ)		
파일공유 (0.12)	97.2	97.8
메신저 (0.4)	95.3	96.3
TV (0.48)	91.3	92.9

세 번째 실험은 전체 16가지 트래픽별로 세분화 분류하는 실험으로써 각 트래픽별로 랜덤하게 데이터를 100개씩 추출하여 각각의 SVDD로 학습하였으며, 학습에 참여하지 않은 데이터를 테스트 데이터로 사용하였다. 실험결과

[표 5]에 정리하였다. 여기서 조정상수 C는 0.1, σ 값은 각 트래픽별로 고정하였다. [표 5]에서 보는 바와 같이 전체적으로 만족스러운 성능을 보이고는 있으나, morpheus와 joost 트래픽의 recall과 fileguri 트래픽의 precision 성능이 상대적으로 낮다는 것을 확인할 수 있다. 이는 morpheus와 joost 트래픽이 feature의 성격이 유사한 fileguri 트래픽으로 잘못 분류되었기 때문이다. 따라서 fileguri의 recall 성능은 높지만 precision 성능은 상대적으로 떨어진다.

[표 5] 전체 애플리케이션 트래픽별 성능 측정 표

평가 항목 application (σ)	recall	precision
bittorrent (0.3)	94	95.9
edonky (0.65)	82	98.8
fileguri (0.07)	100	79.4
limewire (0.65)	94	96.9
morpheus (0.6)	64	100
msnmsgr (0.5)	84	94.4
nateon (0.6)	88	91.7
skype (0.4)	84	95.5
afreecaplayer (0.6)	89	90.8
joost (0.6)	78	96.3
pplive (0.5)	84	93.3
gom (0.25)	90	90.1
gorealra (0.5)	96	86.5
iexplore (0.55)	89	91.8
pop3 (0.5)	97	86.6
ftp (0.3)	98	85.5

4. 결론

본 논문에서는 인터넷 애플리케이션 트래픽 분류가 네트워크 망 관리시스템에서 가장 중요한 기본 기능 중 하나라는 문제의식과 기존의 전통적인 분류 방법으로 대표되는 port 번호 및 payload 정보를 이용하는 방법의 구조적 문제점을 극복하는 차원에서 SVM에 기초한 새로운 인터넷 애플리케이션 트래픽 분류 시스템을 제안하였다. 제안된 시스템은 flow 기반의 트래픽 정보를 수집하여 인터넷 애플리케이션 트래픽을 coarse 혹은 fine하게 분류함으로써 효율적인 시스템의 자원 관리, 안정적인 네트워크 환경의 지원, 원활한 bandwidth의 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 트래픽의 특성을 보다 적절히 반영하는 feature set에 관한 추가적인 연구(feature set selection or/and feature set reduction)가 요구된다.

참고 문헌

[1] CacheLogic homepage, Available in <http://www.cache logic.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 classification," 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] R. Wang, Y. Liu, Y. Yang, and X. Zhou, "Solving the app-level classification problem of P2P traffic via optimized support vector machines," IEEE Sixth Int. Conf. on Intelligent Systems Design and Applications, Vol 2, pp. 534-539, 2006.
 [7] T. Ambwani, "Multi class support vector machine implementation to intrusion detection," Proc. of the Int. Joint Conf. on Neural Networks, Vol. 3, pp. 2300-2305, 2003.
 [8] H. Lee, J. Song, and D. Park, "Intrusion detection system based on multi-class SVM", LNAI, Vol. 3642, pp. 511-519, 2005.