데이터 마이닝을 이용한 무선 인터넷 서비스 분류기법

(Wireless Internet Service Classification using Data Mining)

이 성 진 † 송 종 우 †† 안 수 한 ††† 원 유 집 †††† 장 재 성 †††††
(Seongjin Lee) (Jongwoo Song) (Soohan Ahn) (Youjip Won) (Jae-Sung Chang)

요 약 오늘 날 다양한 플랫폼을 기반으로 한 무선 네트워크 위에 실행되고 있는 수 많은 응용 프로그램은 서비스 운영자 입장에서 정확히 분류해내는 것은 중요하다. 이 연구는 WiBro 상용망에서 임의로 생성한 트래픽 데이터에서 다양한 응용프로그램들을 분류하는 것을 목적으로 한다. 분류기를 개발하는데 있어서 기존에 Flow기반으로 분류를 하는 대신 세션이라는 단위로 실험을 진행하였다. 이 단위를 사용하여 두 가지 분류 기법을 사용하였다: Classification and Regression Tree와 Support Vector Machine. 각 판별기는 생성된 변수들을 기반으로 판별을 시도하였을 때 CART의 경우 0.85%, SVM의 경우 0.94%의 오차를 보여 우수한 성능을 보였지만, 판별기의 구현과 결과 해석이 용이한 CART를 이용하여 판별 시스템을 구축하는 것이 유리함을 보였다.

키워드 : 판별 시스템, CART, SVM, 인터넷 서비스 분류

Abstract It is a challenging work for service operators to accurately classify different services, which runs on various wireless networks based upon numerous platforms. This works focuses on design and implementation of a classifier, which accurately classifies applications, which are captured from WiBro Network. Notion of session is introduced for the classifier, instead of commonly used Flow to develop a classifier. Based on session information of given traffic, two classification algorithms are presented, Classification and Regression Tree and Support Vector Machine. Both algorithms are capable of classifying accurately and effectively with misclassification rate of 0.85%, and 0.94%, respectively. This work shows that classifier using CART provides ease of interpreting the result and implementation.

Key words : Traffic Classification, CART, SVM, Internet Services Classification

· 이 논문은 대학 IT연구센터 육성지원사업의 연구결과로써 HY-SDR연구센터의 연구비 지원으로 수행되었습니다.

† 비 회 원 : 한양대학교 전자컴퓨터통신공학부
james@ece.hanyang.ac.kr
†† 비 회 원 : 이화여자대학교 통계학과 교수
josong@ewha.ac.kr
††† 비 회 원 : 서울시립대학교 통계학과 교수
sahn@uos.ac.kr
†††† 종신회원 : 한양대학교 전자컴퓨터통신공학부 교수
yjwon@ece.hanyang.ac.kr
††††† 비 회 원 : SKTelecom
jsjang@sktelecom.com

논문접수 : 2008년 6월 20일
심사완료 : 2009년 4월 16일

정보과학회논문지: 정보통신 제36권 제3호(2009.6)

1. 서 론

무선망의 경우 억세스 네트워크를 포함한 네트워크 자원 증설은 큰 비용을 요구하기 때문에 정교한 용량 계획, QoS 관리, 그리고 인터넷 트래픽 공학의 방법론 및 지원 인프라가 필요로 하다. 특히, 무선 인터넷의 특성상 셀 내의 가입자들은 하나의 광대역을 공유하고 여러 환경 변수를 고려한 스케쥴링 알고리즘에 따라 자원이 배분되므로 가입자의 QoS 및 네트워크 성능은 트래픽 변화에 민감하다. WiBro 망 기반 서비스와 EV-DO/HSDPA 망을 이용한 무선 모뎀 서비스가 출시되었다. 이러한 무선 망을 이용한 데이터 서비스는 기존의 핸드셋을 통한 자사 컨텐츠 위주로 서비스하고 있었다. 그러나 점차 외부 컨텐츠를 활용하여 다양성을 증가시키고 궁극적으로 여러 소형 기기 플랫폼들을 지원하고

있어 유선 인터넷 서비스와 비슷한 환경으로 발전하고 있다. 그렇기 때문에 트래픽의 변화와 패턴 예측을 위한 네트워크 상에서 발생하는 서비스 별 트래픽 패턴에 대한 분석이 필요하다. 이러한 분석을 통하여 네트워크 엔지니어링 및 서비스 기획 단계에 유용한 도구를 제공할 수가 있다.

트래픽을 분류하는 방법에는 가장 초기에 사용된 방법인 패킷의 페이로드를 직접 읽어 서비스를 구분하는 방법과 IP와 Port를 잘 알려진 Port와 비교를 하여 사용된 응용 프로그램을 판단하는 방법이 있다. 기존의 운영 컨텐츠의 경우, IP/Port 서비스들이 관계 데이터베이스로 구축이 되어 있어 IP/Port 검색에 의해서 사용자가 생성한 트래픽을 분류할 수 있었다. 하지만, WiBro와 같은 오픈망의 경우, 웹 응용 프로그램들이 정규 포트들을 사용하지 않는 경우가 늘어나고 있기 때문에 서비스의 정확한 분류가 쉽지가 않다. 또한, 통계적 방법을 적용하여 분류하는 대부분의 기법들[1,2]은 유선망을 기반으로 한 연구이기 때문에 무선 인터넷 망에 적용했을 때의 예측 성능은 불확실 하다. 현재 구축되어 있는 시스템의 제한 사항과, 무선 인터넷의 특성, 사업자가 필요로 하는 정보 요건 등을 고려하여 구축하고자 하는 시스템에 적합한 분류 방법론의 개발이 필요로 하다. 또한, 실제로 서비스를 제공하는 사업자의 경우 무선망에서의 과금 체계의 확립을 위한 기초 자료를 제공하기 위하여 서비스 별 분류 기법의 개발과 무선망에서의 분석 기본 단위의 정립이 필요하다.

초기에는 패킷 기반 분류 기법[3,4]을 사용하였다. 이 기법은 모든 패킷들을 직접 분석하여 분류를 하는 방식이기 때문에 시스템 자원을 많이 소모하는 가장 큰 단점이 있다. 본 연구에서는 데이터 마이닝 기법을 이용하여 WiBro 무선망에서의 서비스별 트래픽 패턴의 분류를 하고자 한다. 모든 서비스들은 실시간성이나 신뢰성의 보장과 같이 다른 요구 조건들을 갖고 있다. 각 서비스들은 네트워크의 한정 된 자원을 이용하여 서비스를 하기 때문에 서비스들은 서로 다른 통계적 특성을 갖게 된다. 데이터 마이닝 기법을 이용을 하면 각 서비스들이 갖고 있는 통계적 특성을 군집 지을 수가 있다. 본 연구에서는 WiBro 망을 사용하는 서비스들의 통계적 특성을 Classification and Regression Tree(CART[5,6])와 Support Vector Machine(SVM[7])에 적용하여 분류한다.

본 연구를 통하여 얻을 수 있는 몇 가지 장점은 다음과 같다. 실측 트래픽 모니터링 장치를 적용하여 대용량 트래픽을 실시간 처리 할 수 있는 새로운 서비스 별 분류기법을 생성할 수 있다. WiBro 망의 서비스 별 트래픽 분류를 통하여 각 서비스 별 비율 정보를 이용하여 트래픽의 예측과 망의 배분 등의 전략 수립에 필요한 도구를 제공할 수 있다. 또한 트래픽의 서비스 별 분류 연구를 통하여 가입자 별 서비스 이용 패턴을 분석을 가능하게 하고 이것을 통해 트래픽 공학과 자원 용량 계획을 효율적으로 할 수 있는 기반을 제공한다. 더 나아가 패킷 스케쥴링, 정책 기반의 네트워크 관리의 기술 개발과 서비스 개발과정에서의 네트워크 영향의 분석 그리고 서비스 별 무선망 과금 체계의 기초자료를 제공할 수 있다.

본 논문의 2장에서는 관련 연구들을 다루고 3장에서는 데이터의 수집과 사용된 테이블의 설명 그리고 세션의 정의에 대해서 소개한다. 4장은 수집한 데이터를 이용하여 분석 변수의 생성과 CART와 SVM의 분석 기법에 대해 설명하고 5장은 이렇게 개발된 두 분석 기법을 통해 얻은 결과를 토대로 성능 평가를 한다. 6장에서는 판별기를 실제로 구현하여 사용하는 시스템을 도식화 하여 최종적으로 구축된 판별 시스템을 설명한다. 그리고 끝으로 7장에서는 본 연구의 결론을 맺는다.

2. 관련연구

기존의 분류 기법에 관한 많은 연구들은 유선랜 환경에서의 서비스 별 분류에 중점으로 연구가 진행 되어왔었다. 분류를 하는 많은 기법들 중 클러스터링을 사용하는 기법[8]으로 K-Means, DBSCAN, 그리고 AutoClass 등이 있다. K-means는 클러스터링은 주어진 오브젝트들을 특성이나 성질을 기반으로 K개의 묶음으로 분류 또는 그룹을 하는 것을 말한다. 이때 각 그룹은 클러스터의 중심에서 오브젝트 간의 거리의 제곱의 합이 최소가 되도록 구성이 된다. DBscan은 Density-Based Spacial Clustering of Application with Noise의 약자로 노드의 밀도의 분포를 통해서 클러스터링을 하는 알고리즘이다. 이러한 클러스터링을 이용한 판별은 연산속도와 정확도 면에서 좋은 성능을 갖고 있다.

이와 달리 A.I의 기법으로 기계 학습을 사용한 연구들도 있다. 특히 주목할만한 연구로 Naïve Bayes Estimator를 사용한 것과 이 판별 함수의 변형을 사용한 것이다[9-11]. 학습을 과정과 검증을 하는 부분으로 나누어 특성을 배워서 그 결과 값에 유동적으로 분류를 하는 방식으로 프로파일링이나 시그니쳐를[8,11-13] 이용한 분류보다 자동화가 잘 되어 있다고 할 수 있다. 프로파일링이나 시그니쳐를 사용하는 분류 방법은 패킷에 특정 위치에 있는 단어 또는 문자열을 의미 있는 시그니쳐로 사용하거나 또는 문자열들의 집합으로 이루어진 임의의 단위를 시그니쳐로 사용하여 분류를 하는 방법이다. 하지만 이러한 분류 방법의 장점이자 단점은 특정 서비스 어플리케이션에 대한 분석에 매우 좋은 성능을 보이지만, 이러한 정보를 수집하기 위해서는 많은 수의

응용프로그램의 시그니쳐 프로파일, 또는 핑거 프린트 정보를 갖고 있어야 한다. 더군다나 버전이 바뀌거나 포맷이 바뀌는 경우에는 방법론에 의해 생성된 분류법이 무효화될 수도 있다. 특히 많은 응용프로그램들이 독점적인 어플리케이션 프로토콜을 사용하므로 이는 점점 더 어려워지고 있다.

어플리케이션 분석 차원을 떠나서 하나의 사용자가행사하는 사회과학적 역할, 예를 들어 그 클라이언트가 생성자인지 소비자인지와 같은 역할 규명이나, 사용하는 어플리케이션들은 어떤 것이 있는지를 규명하려는 연구가 있다[14]. 짧은 순간의 데이터를 분석하는 과정을 통해서 응용프로그램의 분석뿐만 아니라 사용자 패턴까지 분석을 시도한 다차원적인 분석으로서 의미 있는 연구라 할 수 있다. 이 외에도 한 연구에서는 분류 분석에서 플로우의 크기를 정확히 측정하는 것에는 신뢰도가 높은 연구들이 많이 진행되어 왔지만, 그 플로우의 바이트 크기를 정확히 예측하지는 못했다는 점을 지적하고 있는 연구가 있다[15].

인터넷 트래픽을 분류하는 것을 통해서 과금을 하거나 서비스 QoS를 다르게 제공하는 등 여러 이점이 있기 때문에 여러 종류의 분류 기법들이 생겨났다. 여러 분류 기법을 분석하고 장단점을 서로 비교한 연구가 있다[16]. 이 연구에서는 각 분류 기법들을 체계적으로 그리고 특징 별로 나누어 구분해 놓았고, 특히 트리 기반의 판별 함수들을 심도 있게 다루었다.

3. 데이터 생성 및 수집

본 연구에서 분석한 자료의 수집은 그림 1에 나타나 있고 각 서비스 별로 SKT에서 임의로 생성한 패킷 흐름 자료(packet trace) 자료를 이용하였다. 자료의 생성 일시는 표 1과 같다.

WiBro 망은 이동 중에도 높은 전송속도를 제공하여 언제, 어디서나 무선 인터넷에 접속이 가능하도록 하는 서비스로서, 그림 1에 나타난 것과 같이 억세스 제어 라우터(ACR) 시스템이 라디오 억세스 스테이션(RAS)를

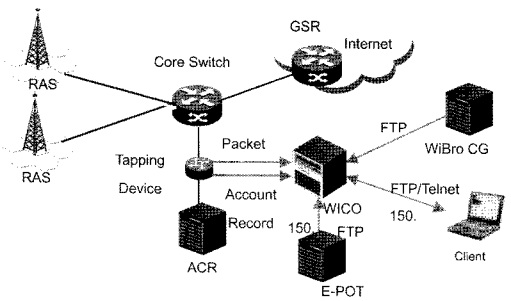

그림 1 WiBro 망 구조와 데이터 수집 장치의 구조

표 1 Dates of Captured Traces

날짜	비고
2007년 9월 27~28일	10분씩 3번
2007년 10월 1~31일	10분씩 3번, 3, 22일 제외
2007년 11월 1~30일	10분씩 6번, 2, 30일 제외

제어하는 구성에서 Tapping device는 라우터를 통과하는 패킷 정보를 WICO 라는 장치에 저장하도록 구성되어 있다. 이 장치는 WiBro 트래픽을 수집하고 분석하는 장치로서 libpcap 라이브러리를 사용하는 TCPdump [17]와 Wireshark[18] 같은 기능을 하면서 추가적으로 과금을 위한 Call Data Record(CDR) 정보도 수집을 한다.

각 Trace의 길이는 10분이며 하루에 각 세 번에 걸쳐 패킷 흐름 자료를 생성하는데, 2007년 11월 15일부터는 하루에 여섯 번 트래픽을 생성하였다. 약 900개의 패킷 흐름 자료 파일을 분석을 위해 제공받았다. 본 연구를 위한 자료의 생성은 여섯 가지 서비스 군으로 나누어서 진행하였다. 서비스 군은 각각 Download, Game, Streaming, Upload, VoD, VoIP, Web 별로 이루어졌으며, 각 서비스 별 사용된 응용 프로그램의 종류는 표 2에 나타나 있다. 다운로드 서비스에서는 웹하드 서비스 업체인 XTOC[19]과 넷폴더[20]에서 대용량 파일들을 다운을 받았다. 온라인 게임 서비스로 3D 액션 게임인 쿵파[21]와 MMORPG 게임인 메이플 스토리[22]를 사용하였고, 업로드는 웹 기반 메일 서비스와 넷폴더를 이용하여 대용량 파일 업로드를 하였다. VoD 서비스의 경우 다음 UCC[23]와 YouTube[24] 그리고 FM 라디오[25]를 사용하였다. Gil Et al. [26]과 Cha Et al.[27]의 연구에서 VoD 서비스의 트래픽 특성에 대한 다각적인 분석이 나타나 있다. VoIP 서비스에서는 SkyPe[28]와 네이트온[29]을 이용하였고, Bonfiglio Et al.[30]에서는 SkyPe의 트래픽의 특성과 PC로 통신을 하는 경우와 PSTN 망을 이용하는 경우에서의 분류가 나타나 있다. 그리고 Naver[31], Daum[32], 그리고 Empas[33] 서비스를 사용하여 웹 서비스 트래픽을 발생하였다.

각 서비스와 해당 응용프로그램에서 발생된 패킷 자료를 바탕으로 하여 표 3의 데이터베이스 테이블을 생

표 2 Scenarios of Services

ID	Services	Description
S1	Download	XTOC, NetFolder
S2	Game	Koongpa, Maple Story
S3	Upload	Mail Upload, NetFolder
S4	VoD	Daum UCC, YouTube, FM Radio
S5	VoIP	SkyPe, Nateon
S6	Web	Naver, Daum, Empas

표 3 Description of DB Tables

Table names	Description
FlowNumberInc	Flow Number와 한 Flow에 속한 패킷들의 정보가 포함되어 있다. 패킷 간의 도착간격시간과 헤더 정보와 크기 그리고 어느 서비스 군에 속한 패킷인지에 대한 정보가 포함되어 있다.
FlowInfoALL	한 세션에 속한 Flow들의 방향정보와 기본 헤더 정보 그리고 속해 있는 서비스 군을 설명하고 도착간격시간과 패킷에 대한 기초 통계 정보가 포함되어 있다.
SessionInfo	한 세션이 갖는 Flow의 기초정보로 Rx와 Tx로 구분이 되어 있고 Flow에 대한 기초 통계정보로 이루어져 있다.

성하였다. 여기서 언급된 플로우는 방향을 갖고 있으며 Source IP, Source Port, Destination IP, Destination Port, Layer 4 Protocol 그리고 Types와 Scenarios와 같은 패킷들의 집합이며, 10초라는 시간 동안 집합에 속하는 패킷이 도착하지 않을 때 플로우는 끝이 난다. 위의 자료 테이블에는 장치에서 얻을 수 있는 정보 외에 실험하고자 한 여섯 가지 네트워크 응용 프로그램 정보가 데이터베이스에 시나리오 항목에 추가되어 있다. 이와 같이 시나리오 정보를 첨가한 이유는 판별기에서 구분한 서비스 군이 올바르게 예측하였는가에 대한 사후 검정을 하기 위함이다.

세션은 하나의 응용 프로그램을 이용해 두 대의 컴퓨터가 양방향 통신을 하는 경우, 이들이 주고받은 패킷들의 집합이다. 즉, 세션 은 방향성을 갖는 플로우가 서로 반대 방향으로 메시지를 주고받으면 하나의 세션이라고 한다. 예를 들어 플로우 넘버가 1이고 SrcIP=주소1, SrcPort=포트1, DstIP=주소2, DstPort=주소2이고 플로우 넘버가 2이가 SrcIP= 주소2, SrcPort=포트2, DstIP=주소1, DstPort=포트1일 때 하나의 세션으로 간주한다.

세션의 도식화한 정의는 그림 2에 나와 있다. 이런 세션 정보가 플로우에 대해 갖는 장점은 한 사용자의 서비스 사용 패턴을 알 수가 있다. 플로우의 경우 본 연구에서 추가적으로 정의한 10초라는 시간 동안 사용자가 아무런 명령을 내리지 않을 때는 종료가 된다. 그렇지만 사용자가 뉴스를 읽거나 그림을 보거나 잠시 자리를 비운다 하더라도 10초 이후에 일어나는 일은 해당 사용자의 연속된 행위로 인식할 수 있다. 그렇기 때문에 하나의 세션은 한 사용자가 하나의 서비스 어플리케이션을 사용하기를 시작하여 다른 서비스로 전환하기까지의 모든 정보를 집약적으로 갖고 있는 것이다.

사용자의 사용패턴에 따른 측면 이외에도 응용 프로그램 차원에서 세션 개념을 사용하는 서비스는 SSH와 Telnet과 같은 것이 있다[34]. 또한, 멀티미디어 서비스

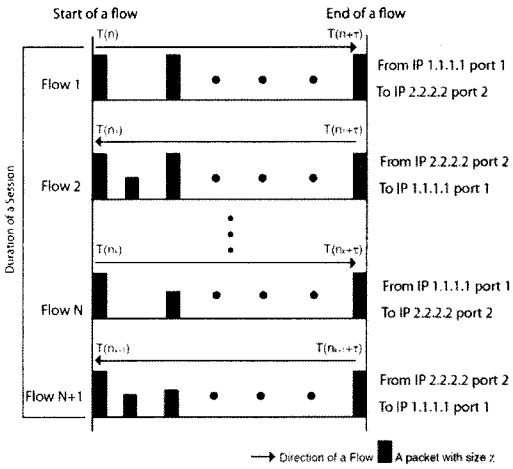

그림 2 Definition of Session

를 사용하는 경우에서는 프로토콜 차원에서 세션을 사용하는 경우이다[35]. 반면 플로우를 기반으로 하는 트래픽 연구는 최근 Origin Destination(OD) 플로우를 기반으로 하고 있다[36]. OD 플로우는 라우터를 통과하는 트래픽을 발생지점과 종착지점으로 구분하여 네트워크 진단과 이상현상[37]을 규명하는 용도로 사용하고 있다. 하지만, 트래픽 분류에선 하나의 모니터링 지점에서 수집한 플로우 데이터를 사용하고 있기 때문에 사용자 패턴과 응용 서비스 그리고 프로토콜 차원에서 사용되는 세션에 대한 분석은 진행되지 않았다[9,38]. 본 연구에서는 세션 정보를 활용하여 트래픽 분류를 한다.

4. 분석 변수의 생성과 분석 기법

본 연구에서 사용된 분석의 단위는 세션이며 이는 그림 2에서 정의되었다. 본 연구에서의 목적은 무선망에서 수집된 세션을 보고 이 세션이 어떠한 응용 프로그램에 해당하는지 판별할 수 있는 판별함수를 만드는 것이다. 따라서 본 연구에서는 제공받은 패킷 자료를 이용하여 세션을 구성하고 구성된 세션으로부터 여러 가지 통계량, 즉, 하나의 세션 에 포함된 패킷 수, 세션 의 지속시간, 패킷들의 크기에 대한 평균, 표준편차, 변동계수 패킷간 도착간격에 대한 평균, 표준편차, 세션을 이루는 업링크, 다운링크 패킷간의 통계량의 비율 등의 변수를 생성하였으며 이들을 판별분석에 이용하였다.

생성된 변수를 이용하여 각 세션에 해당하는 응용 프로그램에 대한 판별분석을 행하였다. 본 연구에서는 이를 위하여 데이터마이닝 기법들 중에서 널리 쓰이고 성능이 우수하다고 알려진 CART[5,6] 기법과 SVM[7] 기법을 이용하여 판별함수를 만들고 이를 통한 판별결과에 대한 성능 검증을 행하였다. 분류를 위해서 사용된

표 4 Table of Features(세션을 기준으로 함)

Notation	Feature Description
Duration	한 세션의 길이
NFlows	플로우 개수
NPackets	패킷의 개수
PcktAVG	패킷 크기의 평균
PcktSTD	패킷 크기의 표준편차
PcktIATAvg	패킷 평균도착간격 시간
sumFlows	플로우 크기의 합
avgFlows	플로우 크기의 평균
secMomentFlows	플로우의 크기의 2차 모멘트
stdFlows	플로우의 크기의 표준 편차
varFlows	플로우의 분산
varDavgt	CSQ로서 분산을 평균의 제곱으로 나눈 값

기법은 CART와 Support Vector Machine(SVM)이다. 데이터마이닝 기법으로서 CART는 간단하고[39] SVM은 정확도가 높은 것으로 알려져 있다[40]. 사용된 통계값들은 표 4에 나타나 있다.

4.1 Classification and Regression Tree(CART)

CART[5,6]는 Classification and Regression Tree의 약자로 분류가 간단하고, 그 결과를 해석하기가 쉬우며 좋은 예측 성능을 보인다. 명칭에서 알 수 있듯이 트리 구조로 되어 있으며 분기가 되는 가지에서 판별식의 분석 변수들을 통해서 이원 분배(Binary Split)를 하고 예시는 그림 3에 나타나 있다. 데이터의 설명 변수들을 이용하여 오차가 최소가 되도록 하는데 분할 반복 횟수는 미리 정해 논 최소 문턱 값이 될 때까지 반복한다. 최소 문턱 값까지 반복하여 더 나눌 수 없을 때 각 터미널 노드에 놓인 변수들이 적합하게 분류가 되었다고 한다. 이 방법의 목적은 분할을 한 결과들이 서로 동차가 되도록 만드는 것이다.

CART 분석 기법은 총 세 단계로 이루어져 있다. 첫 단계에서는 트리를 만든다. 트리는 학습 자료에 의해 생성된 예측된 클라스와 의사 결정 비용 매트릭에 의해 나눠지는 노드들로 구성이 되어 있다. 두 번째 단계는 트리의 신장에 한계를 정하는 것이다. 현재 클라스의 자료로 자식 노드를 구성할 수 없을 때까지 반복이 되기 때문에 멈추는 과정은 중요하다. 멈추는 시점을 정하는 몇 가지 경우의 수가 존재한다. 그 경우의 수는 다음과 같다. 먼저 각 노드에 값이 하나씩인 경우이다. 두 번째는 모든 자식 노드가 같은 확률 분포를 갖고 있어서 더 이상 나눠지지 않는 경우이다. 그리고 세 번째는 처음 설정할 때 트리의 깊이를 설정한 경우까지만 신장할 수 있게 하는 것이다. CART 분석 기법의 세 번째 단계는 나무의 가지치기를 통해서 트리를 간단화 하는 것이다. 터미널 노드에서부터 시작하여 트리의 복잡도가 그 노드를 잘랐을 때 어떤 값 δ 이하로 되지 않을 때까지 잘라낸다. 이 δ에 의해서 트리는 간단화가 될 수 있다. 마지막 단계로는 세 번째 단계를 통해서 얻은 트리를 최적화 하는 것이다. δ 값이 작아져서 Tree가 더 세분화 될 때는 과도하게 분류를 하여 오히려 정확도를 떨어트릴 수가 있기 때문에 문턱 값을 조심스럽게 정하는 것은 매우 중요하다. 이러한 과정을 거치면서 정확도를 높이기 위해서 학습 데이터를 갖고 CART의 세 단계를 수행하기 위해서 데이터를 K등분을 한다. N등분을 중 하나로 학습을 하고 나머지 k-1등분의 데이터를 통해서 검증하는 과정을 k번 시도하여 정확도를 측정한다.

4.2 Support Vector Machine(SVM)

SVM은 Support Vector Machine의 약자로 우수한 성능 때문에 데이터마이닝 기법에서 많이 사용되는 분류 방법 중에 하나이다[7]. 이 방법도 학습과 예측의 두 단계로 나뉘어 연산이 된다. 학습하는 과정에서는 어떤 평면 위에 있는 데이터를 구분하기 위해서 경계면을 그리게 된다. 이 경계면은 무수히 다양한 종류로 표현이 될 수 있는데 최적의 분리 경계면을 찾아내는 것은 SVM의

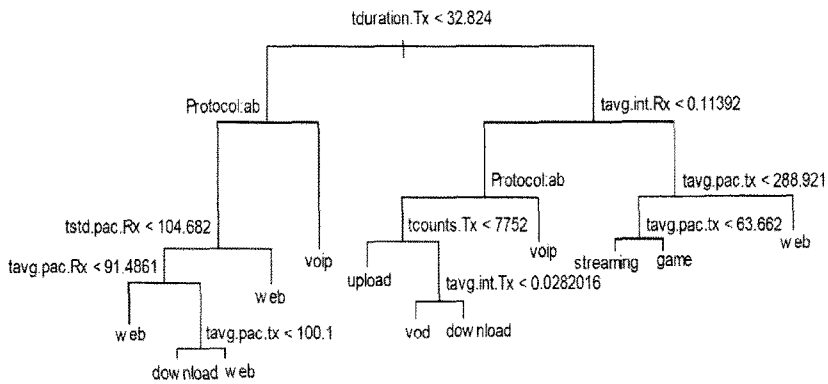

그림 3 Classification with CART

커널 함수의 목적이다. 이 커널 함수는 서로 다른 특징을 갖는 데이터 값들에서 최대로 멀리 떨어져 있게 되는 지점을 알려준다. 데이터 D가 두 클라스로 구성이 되어 있고 $D=\{(x_1,y_1),...,(x_n,y_n)\}$, $x\in R_n$, $y\in\{-1,1\}$ 일 때 이 데이터들은 초평면 $\langle w,x_i\rangle+b=0$에 존재한다. 변수 w, x가 $\min_i|\langle w,x_i\rangle+b|=1$의 제한 조건이 있을 때 두 클라스를 나누는 초평면은 $y_i[\langle w,x_i\rangle+b]\geq 1$, $i=1,...n$의 조건을 만족해야 한다.

5. 성능 평가

표 5와 표 7은 전체 데이터로 학습을 한 후에 동일한 자료로 검증을 한 것을 나타낸 것이다. 그리고, 표 6과 표 8은 검사의 신뢰성을 높이기 위하여 스플릿 검증 또는 K-Fold 교차 검증이라 불리는 방법을 사용하였다[41]. 이 방법은 데이터를 K 등분을 한 후 K-1 등분의 데이터는 학습에 사용하고, 나머지 하나는 검증을 위하여 사용을 하였다. 총 K번 K-Fold 교차 검증을 반복 시행하여 정확도를 높였다. 모든 테이블의 가로축의 내용은 서비스의 종류를 나타내고 세로 축은 각 서비스를 정확히 구분을 했는가를 확인을 위한 서비스 이름 표기이다. 각 칸에 있는 수가 의미하는 바는 세로축의 서비스를 판별기가 해당 서비스로 올바르게 판별하였는가를 나타낸다. 예를 들어 표 6에서 Download의 경우 총 68개의 세션들의 중에 1개를 제외하고는 모두 Download (S1) 서비스로 분류 해낸 것을 확인할 수가 있다. 1개의 세션은 Web(S6) 세션으로 오판하였다. 모든 표에서는 대각 행렬에 있는 수가 높을수록 판별함수의 성능이 좋게 나타난 것을 알려준다. 모든 표에서 마지막 열은 비교를 위하여 해당 서비스의 총 세션 개수를 표현한다.

표 5를 살펴보면 Download(S1), VoIP(S5), 그리고 Web(S6)은 판별율이 96%, 97%, 그리고 100%로 오판율이 매우 낮은 것을 확인할 수 있다. Online Game (S2, 83%)와 Upload(S3, 70%)의 경우 비교적 높은 판별율을 나타내고 있는데, Online Game(S2) 트래픽의 경우 Upload(S3), VoIP(S5)와 Web(S6)로 오판하는 경우가 소폭 있었다. S3의 경우 Download(S1) 트래픽으로 오판하는 경우가 30% 있었다. 오판율이 가장 높았던 것은 VoD(S4) 트래픽으로 Web(S6) 트래픽으로 오판하는 경우가 34% 있었다. SVM의 성능 평가를 나타내는 표 7의 경우 CART에 비해 판별율이 떨어지는 것을 확인할 수 있다. VoIP(S5)와 Web(S6)의 경우 97% 이상의 높은 판별율을 나타내었고 S1에서 S4까지는 60%에서 90% 사이의 판별율을 보였다.

CART와 유난히 다른 특성을 나타내고 있는 것은 모

표 5 CART를 이용한 판별분석 결과(모든 자료 이용)

Appl. Pred.	S1	S2	S3	S4	S5	S6
S1	96%	0%	24%	5%	0%	0%
S2	0%	83%	0%	0%	1%	0%
S3	2%	2%	70%	2%	0%	0%
S4	0%	0%	4%	57%	0%	0%
S5	0%	5%	0%	1%	97%	0%
S6	2%	9%	2%	34%	3%	100%
Ratio	96%	83%	70%	57%	97%	100%

표 6 CART를 이용한 판별분석 결과 (K-Fold Cross Validation)

Appl. Pred.	S1	S2	S3	S4	S5	S6
S1	96%	0%	21%	7%	1%	0%
S2	0%	82%	0%	0%	0%	0%
S3	2%	3%	74%	1%	0%	0%
S4	2%	0%	3%	52%	0%	0%
S5	0%	7%	0%	1%	96%	0%
S6	0%	9%	2%	38%	3%	100%
Ratio	96%	82%	74%	52%	96%	100%

표 7 SVM을 이용한 판별분석 결과(모든 자료 이용)

Appl. Pred.	S1	S2	S3	S4	S5	S6
S1	70%	0%	0%	0%	0%	0%
S2	0%	87%	0%	0%	0%	0%
S3	0%	0%	74%	1%	0%	0%
S4	0%	0%	0%	64%	0%	0%
S5	0%	4%	0%	0%	97%	0%
S6	30%	9%	26%	34%	3%	100%
Ratio	70%	87%	74%	64%	97%	100%

표 8 SVM을 이용한 판별분석 결과 (K-Fold Cross Validation)

Appl. Pred.	S1	S2	S3	S4	S5	S6
S1	77%	0%	0%	0%	0%	0%
S2	0%	81%	0%	0%	1%	0%
S3	0%	0%	76%	1%	0%	0%
S4	0%	0%	0%	58%	0%	0%
S5	0%	5%	0%	0%	92%	0%
S6	23%	14%	24%	41%	7%	100%
Ratio	77%	81%	76%	58%	92%	100%

든 서비스에서 Web(S6) 트래픽으로 오판하는 경우가 비중 있게 나타나고 있다는 것이다. 판별을 위해 K-fold 교차 검증으로 얻은 결과(표 6)의 경우 표 5와 마찬가지로 S4 VoD의 오판율이 높게 나타나고 있다. SVM의 경우인 표 8에서도 데이터를 나누지 않은 표 7와 비슷한 비율을 보이고 있다.

표 5, 표 6, 표 7, 그리고 표 8에서 전반적으로 알 수 있는 것은 Web 트래픽을 다른 트래픽에서의 구분하는 것과 VoD 트래픽을 Web 트래픽에서 구분해 내는 것이 간단한 문제가 아니라는 것을 확인 할 수가 있다. VoD 서비스의 경우 YouTube와 Daum UCC에 접속을 하였다. 웹 브라우저를 사용하지 다른 서비스들과 달리 VoD의 경우 Web 서비스와 동일한 웹 브라우저를 사용하고 있다. 두 서비스 모두 80 포트를 통해 서비스를 제공받았다. 웹 서비스의 경우 하나의 페이지에 여러 종류의 컨텐츠가 혼재 할 수가 있는데, 택스트가 주를 이루겠지만 용량이 큰 플래쉬 광고나 유저가 올린 그림 등도 있을 수 있다. 용량이 큰 파일들을 전송을 할 때 MTU에 의해서 결정된 1500 바이트가 전송이 된다. 이와 똑 같은 경우가 VoD 서비스의 경우에서도 재현이 된다. 다만 차이가 있다면 미디어의 파일 크기이다. 또한 Web 서비스와 유사한 패턴을 보이는 부분은 TCP의 AIMD를 사용하는 플로우 콘트롤 현상이다. VoD 서비스가 처음 컨넥션을 맺어서 서버에서 패킷을 받을 때 오고 간 패킷들의 크기가 Web 서비스에 큰 파일들을 서버에서 받을 때와 다르지 않았다.

CART 분류기의 성능 도표를 그림 4에 Box-and-Whisker-Plot을 이용하여 나타내었다. 이 도표는 5개 통계값에 대한 효율적 표현 방법으로서 중간값과 제 1 분위수 와 3분위수 그리고 최소와 최대 값에 대한 분포를 보여주고, 또한 도표에서 "+"으로 표기 된 이상 값에 대한 이해도 할 수가 있다. 도표의 X 축에서는 분류하고자 하는 서비스들의 목록이 나와 있다. 그리고 Y 축에는 CART를 사용하여 분류한 서비스들을 나타내고 있다. Download의 경우를 보면 5개의 통계 값인 최소값, 제 1분위수, 중간값, 제 3분위수, 그리고 최대값이 모두 Download(S1)를 가리키고 있다. 이상 값을 보면 통계 값에 영향을 주지는 못한 비중을 갖고 있지만 Upload(S3)와 VoD(S4)로 판별한 경우가 있었다는 것을 보여 준다.

그림 4에서 다른 서비스들에 비해서 VoD(S4)는 다른 특성을 보이고 있다. VoD서비스가 Download로 분류된 경우가 있지만 최소값으로 표현되었고 Web으로 분류된 경우가 의미 있는 분포를 갖고 있음을 제 3분위수의 값을 보아 알 수가 있다. 실제 분포를 보면 Download 7%, VoD 52%, 그리고 Web이 38%의 분포를 갖고 있다.

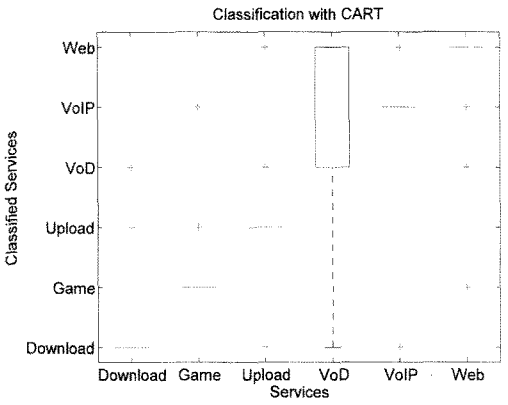

그림 4 CART 분류기의 성능 Box Plot

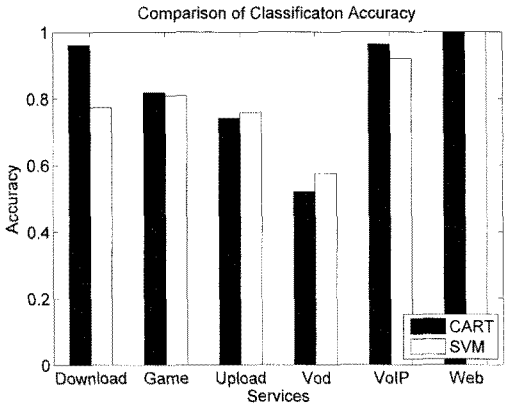

그림 5 CART와 SVM의 성능 비교

그림 5에서는 CART와 SVM이 해당 서비스를 얼마나 잘 분류해내는지 비교하고 있다. Download, Game과 VoIP를 CART가 더 잘 분류하고 있음을 알 수 있다.

판별 분석의 오차는 전체 트래픽을 이용해 판별을 한 것에서는 CART가 0.83% 그리고 SVM이 0.81%의 오차율을 나타내었지만 학습과 판별을 나누어 진행했을 때에는 그 오판율이 CART가 더 작음을 알 수 있다. CART의 경우 0.85%이고 SVM의 경우 0.94%이다. 수치상으로는 CART 기법은 학습과 판별을 위한 데이터를 구분하지 않아도 성능의 차이가 크지 않았던 반면 SVM의 경우 성능이 크게 차이가 나고 있음을 알 수 있다. 그렇지만 분류를 하는데 있어서 오차율이 1% 미만으로 두 경우에 나타났기 때문에 구현 단계에서 발생할 수 있는 문제들을 다루지 않을 수가 없다. 특별히 고려를 해야 하는 것이 있다면 먼저는 간단한 결과 해석의 가능 여부와 빠른 연산 처리 속도, 그리고 신규 서비스가 추가 되었을 때 모델의 적응력을 들 수가 있다.

먼저 간단한 결과 해석 가능 여부에 대해서 보면

CART는 분류 항목의 크고 작음에 대하여 이원 분배를 통한 파티션을 한다. 그리고 최종적으로 파티션을 마치게 되면 정확하게 분류된 결과를 얻을 수 있을 뿐 만 아니라 도식적으로 그 분류를 확인할 수가 있다. 이것은 매우 큰 장점으로 인지적으로 이해력을 도와 쉬운 해석을 가능하게 하기 때문이다. 두 번째로 빠른 연산처리 속도를 보면 SVM은 정교한 판별식으로 높은 복잡도의 연산을 통해야만 정확한 판별을 할 수가 있다. 또한 서비스들의 관계와 특성이 복잡하게 중첩이 될수록 연산의 복잡도는 증가를 하게 된다. 그에 반해 CART는 패킷 흐름 데이터와 세션 데이터에서 미리 계산된 기초 통계자료와 정보들을 통해서 서로의 차이가 가장 큰 크기 비교를 통해서 간단히 분류를 하기 때문에 낮은 복잡도의 연산을 통해서도 성공적인 분류 분석을 가능하게 한다. 마지막으로 신규 서비스가 추가 되었을 때에 복잡도 문제를 들 수가 있다. SVM은 새로운 서비스를 처리하기 위해서 분류기의 변수들을 수정을 하여야 한다. 그리고 이 과정에서 분류기의 복잡도는 높아지게 된다. 하지만 CART의 경우 새로운 서비스를 추가되면 기존의 트리 기반의 분류기에서 가장 유사한 노드로 분기하게 된다.

6. 판별 시스템의 구축

본 연구에서의 분석 결과를 바탕으로 서비스의 판별을 위한 시스템에서는 CART를 이용한 판별함수를 사용한다. 판별 시스템의 구조는 그림 6에서 보인다. 패킷 덤프에서 파싱 과정을 거치고 데이터베이스에 올리고 판별 함수에 적용한 분류를 한다.

그림 1의 WiCO 장치를 통해서 추출한 WiBro 망의 패킷 흐름 자료는 서비스 제공자의 필요에 맞게 정보를 파싱을 한다. 파싱되는 정보들은 플로우를 구성하는 다섯

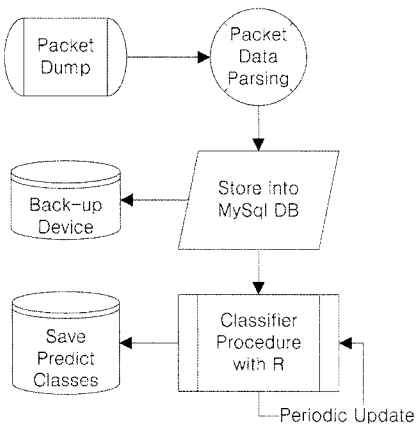

그림 6 Classifier System Structure

표 9 CART 분류기법에서 사용된 변수

변수 명	의미(세션을 기준으로 함)
tduration	세션의 지속 시간
protocol:ab	프로토콜의 비교 및 구분
tstd.pac	패킷 크기의 표준편차
tavg.pac	패킷 크기의 평균
tavg.int	패킷의 평균도착간격 시간
tcounts	패킷의 수
Rx, Tx	플로우들의 방향, 사용자측에서 서버측으로 전송되는 플로우를 Tx로 하고 그 반대를 Rx로 함

개의 요소(SrcIP, SrcPort, DstIP, DstPort, Protocol) 그리고 세션을 파악하기 위한 패킷의 방향, 크기, 도착시간 등 패킷의 헤더에서 얻을 수 있는 정보들이다. 이렇게 파싱된 패킷 흐름 자료는 데이터베이스에 저장이 된다. 저장과 동시에 패킷 흐름 자료의 기초 통계 자료들이 같이 연산되어 저장이 된다. 패킷 흐름 자료는 패킷 단위의 정보이기 때문에 연구에서 사용된 한 사용자의 연속적 행동의 집합인 세션 단위의 정보로 변환을 한다.

모델링을 위해서는 R 패키지[42]를 이용하여 구현하였다. 그러나 실제 시스템 설계 때에는 C 언어로 작성이 되었다. CART의 구현은 4.1절에서 설명한 것처럼 네 단계를 거쳐서 구현된다. 먼저는 데이터베이스에 있는 세션 정보와 기타 통계 자료들을 이용하여 구분을 가장 잘할 수 있는 행들을 먼저 이용하여 트리를 만든다. 그렇기 때문에 표 4에서 CART의 알고리즘에서 사용 가능한 통계값들이 나타나 있지만 모든 값들이 사용된 것은 아니다. 트리를 구성하는 주요 노드와 신장하는 기준이 된 통계 값들은 표 9에 표기 되어있다.

이 값의 결정은 학습하는 과정에서 얻은 정보를 사용하여 주기적으로 갱신이 되도록 설계가 되었다. 이 트리가 신장을 하는데 불필요하게 많이 신장하지 않도록 하는 것과 가지를 잘라 내었을 때 문턱 값을 넘지 않도록 하는 최적화 과정이 둘째와 셋째 단계이다. 모델링을 하는 과정에서 임의로 생성한 패킷 흐름 자료를 학습 자료로 하여 CART를 학습을 시킨다. CART를 통해서 얻은 최종 결과 정보를 판별기에 적용한다. 그리고 새로 생성되는 패킷 흐름 자료를 이 판별기를 통과 시켜 사용된 서비스의 종류를 판단해 낸다. 이렇게 구현된 트래픽 판별 시스템은 주기적으로 갱신이 가능하도록 설계하였다.

7. 결 론

본 연구에서는 서비스 분류를 위하여 세션 단위의 트래픽 분류를 시도하였고, CART와 SVM을 이용한 트래픽 판별 함수를 개발하여 성능 검증을 하였다. 두 종류의 다른 판별 함수의 오판별율은 CART가 0.94%와

SVM이 0.89%로 0.1%에 불과하였다. 이를 통해서 두 방법 모두 매우 우수한 성능으로 분류를 할 수 있음을 보였다. 두 방법이 모두 우수하기는 하나 연산량, 판별함수의 결과에 대한 이해력, 그리고 사후 변화 적응력에 대해서 비교를 해본다면 CART를 사용하는 것이 더 용이하다는 것을 보였다. 이러한 연구 결과는 다음의 분야 등에서 활용될 수 있다. 지능적 트래픽 분석 솔루션의 개발과, 서비스 별 트래픽 모델링 및 서비스 사용 패턴 데이터 제공에 사용될 수 있으며 트래픽 발생 패턴 및 발생량 그리고 그 발생된 트래픽으로 인한 네트워크에 미치는 영향도 분석을 가능하게 하고 서비스를 이용하는 사용자 패턴의 이해를 도울 수 있다. 더 나아가 서비스 및 네트워크 개발 시 필요한 네트워크 시뮬레이션과 사용자의 과금 체계를 위한 기초 자료를 제공할 수 있다.

본 연구에서는 세션 단위 트래픽의 서비스 별 성공적인 판별 가능성을 보여줬다. 그러나 본 연구에서 사용된 트래픽 자료는 실제 네트워크에서 고객들에 의해 생성된 트래픽이 아닌 기획된 트래픽 임을 간과해서는 안 된다. 따라서 본 연구 결과의 실제 네트워크 상에서 응용 가능성에 대해서는 사후 검증이 필요할 것이다. 실제 트래픽은 시간대 별, 요일 별, 계절 별, 그리고 이벤트에 등 여러 요인에 따라 다양한 패턴을 가질 수 있다는 것을 예상할 수 있으며 이에 대한 연구도 이루어져야 할 것이다. 따라서 향후 지능형 트래픽 분석 솔루션 시스템의 구축을 위해서는 다음의 연구가 필요할 것이다. 먼저는 실제 네트워크에서의 데이터의 추출과 연구결과의 적용 및 성능 분석이 이루어져야 하고 두 번째로 서비스 어플리케이션 별 실제 트래픽의 패턴 분석과 트래픽의 샘플링 방안 연구와 샘플링 자료를 이용한 전체 네트워크의 분석 및 예측 방안 연구가 필요로 하다.

참 고 문 헌

[1] D. Hamza, V. Sandrine, and R. David, "A markovian signature-based approach to IP traffic classification," in *Proceedings of the 3rd annual ACM workshop on Mining network data*, San Diego, California, USA: ACM, 2007.

[2] C. Manuel, D. Maurizio, G. Francesco, and S. Luca, "Traffic classification through simple statistical fingerprinting," *SIGCOMM Comput. Commun. Rev.*, Vol.37, pp. 5-16, 2007.

[3] V. Paxson, "Bro: a system for detecting network intruders in real-time," pp. 3-3, 1998.

[4] M. Roesch, "Snort: Lightweight Intrusion Detection for Networks."

[5] L. Breiman, *Classification and Regression Trees*: Chapman & Hall/CRC, 1998.

[6] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning*: Springer, 2001.

[7] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*: Cambridge University Press, 2000.

[8] E. Jeffrey, A. Martin, and M. Anirban, "Traffic classification using clustering algorithms," in *Proceedings of the 2006 SIGCOMM workshop on Mining network data*, Pisa, Italy: ACM, 2006.

[9] W. M. Andrew and Z. Denis, "Internet traffic classification using bayesian analysis techniques," in *Proceedings of the 2005 ACM SIGMETRICS international conference on Measurement and modeling of computer systems Banff*, Alberta, Canada: ACM, 2005.

[10] W. Nigel, Z. Sebastian, and A. Grenville, "A preliminary performance comparison of five machine learning algorithms for practical IP traffic Flow classification," *SIGCOMM Comput. Commun. Rev.*, Vol.36, pp. 5-16, 2006.

[11] H. Patrick, S. Subhabrata, S. Oliver, and W. Dongmei, "ACAS: automated construction of application signatures," in *Proceeding of the 2005 ACM SIGCOMM workshop on Mining network data*, Philadelphia, Pennsylvania, USA: ACM, 2005.

[12] X. Kuai, Z. Zhi-Li, and B. Supratik, "Profiling internet backbone traffic: behavior models and applications," in *Proceedings of the 2005 conference on Applications, technologies, architectures, and protocols for computer communications* Philadelphia, Pennsylvania, USA: ACM, 2005.

[13] M. Roughan, S. Sen, O. Spatscheck, and N. Duffield, "Class of Service Mapping for QoS: A Statistical Signature based Approach to IP Traffic classification," in *IMC'04* Taormina, Sicily, Italy, 2004.

[14] K. Thomas, P. Konstantina, and F. Michalis, "BLINC: multilevel traffic classification in the dark," *SIGCOMM Comput. Commun. Rev.*, Vol.35, pp. 229-240, 2005.

[15] E. Jeffrey, M. Anirban, and A. Martin, "Byte me: a case for byte accuracy in traffic classification," in *Proceedings of the 3rd annual ACM workshop on Mining network data* San Diego, California, USA: ACM, 2007.

[16] E. T. David, "Survey and taxonomy of packet classification techniques," *ACM Comput. Surv.*, Vol.37, pp. 238-275, 2005.

[17] http://www.tcpdump.org/, *TCPDump/LIBPCAP Public Repository.*

[18] http://www.wireshark.org/, *WireShark-Network Protocol Analyzer.*

[19] http://www.xtoc.com, *WebHard Service Company.*

[20] http://www.netfolder.co.kr/, *NetFolder.*

[21] http://koongpa.nexon.com/, *3D Online Action Game.*

[22] http://www.maplestory.com/, *MMORPG Game.*

[23] http://ucc.daum.net/, *Daum UCC.*

[24] http://www.youtube.com, *YouTube.*
[25] http://radio.sbs.co.kr, *SBS FM Radio.*
[26] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, "Youtube traffic characterization: a view from the edge," pp. 15-28, 2007,
[27] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system," pp. 1-14, 2007.
[28] http://www.skype.com, *SkyPe.*
[29] http://nateonweb.nate.com/en/, *NateOn Messenger.*
[30] D. Bonfiglio, M. Mellia, M. Meo, D. Rossi, and P. Tofanelli, "Revealing skype traffic: when randomness plays with you," pp. 37-48, 2007.
[31] http://www.naver.com, *Naver.*
[32] http://www.daum.net, *Daum.*
[33] http://www.empas.com, *Empas.*
[34] D. Tang and M. Baker, "Analysis of a local-area wireless network," pp. 1-10, 2000.
[35] H. Kang, M. Kim, and J. Hong, "Streaming Media and Multimedia Conferencing Traffic Analysis Using Payload Examination," *ETRI Journal*, Vol. 26, pp. 203-217, 2004.
[36] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E. Kolaczyk, and N. Taft, "Structural analysis of network traffic Flows," pp. 61-72, 2004.
[37] X. Li, F. Bian, M. Crovella, C. Diot, R. Govindan, G. Iannaccone, and A. Lakhina, "Detection and identification of network anomalies using sketch subspaces," pp. 147-152, 2006.
[38] S. Zander, T. Nguyen, and G. Armitage, "Self-learning IP Traffic Classification based on Statistical Flow Characteristics," 2005.
[39] R. Lewis, "An Introduction to Classification and Regression Tree (CART) Analysis," pp. 1-14, 2000.
[40] C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, Vol.2, pp. 121-167, 1998.
[41] K. Duan, S. Keerthi, and A. Poo, "Evaluation of simple performance measures for tuning SVM hyperparameters," *Neurocomputing*, Vol.51, pp. 41-59, 2003.
[42] http://www.r-project.org, *The R Project for Statistical Computing.*

이 성 진
2006년 한양대학교 전자전기컴퓨터공학과 졸업(학사). 2008년 한양대학교 전자컴퓨터통신공학과 졸업(석사). 2008년~현재 한양대학교 전자컴퓨터통신공학과 박사과정 재학 중. 관심분야는 Network Traffic Modeling and Analysis, Traffic Engineering, Classification

송 종 우
1993년 서울대학교 계산통계학과 졸업(학사). 1995년 서울대학교 통계학과(석사). 2003년 University of Chicago 통계학과 졸업(박사), 2003년~2006년 Purdue University 통계학과 조교수. 2006년~현재 이화여자 대학교 통계학과 조교수. 관심분야는 Data Mining, Statistical Genetics, Microarray Data Analysis, Clustering, Classification, Semiparametric Mixture Model, Dissimilarity Measure in Time-Series Data

안 수 한
정보과학회논문지 : 정보통신
제 36 권 제 2 호 참조

원 유 집
정보과학회논문지 : 정보통신
제 36 권 제 2 호 참조

장 재 성
2003년 2월 서울대학교 산업공학과 학사. 2005년 2월 서울대학교 산업공학과 석사. 2005년 2월~현재 SK Telecom Access Network Center, 연구원. 관심분야는 WCDMA/HSDPA Performance Analysis Network System Modeling Wireless Internet Traffic Pattern