



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2010-0125076
(43) 공개일자 2010년11월30일

(51) Int. Cl.

H04L 12/56 (2006.01) H04L 12/24 (2006.01)

(21) 출원번호 10-2009-0044093

(22) 출원일자 2009년05월20일

심사청구일자 2009년05월20일

(71) 출원인

고려대학교 산학협력단

서울 성북구 안암동5가 1

(72) 발명자

박대희

서울특별시 강남구 도곡동 타워팰리스 A동 1601호

유재학

충청북도 옥천군 옥천읍 죽향리 옥향아파트 103동 803호

(뒷면에 계속)

(74) 대리인

박준석

전체 청구항 수 : 총 15 항

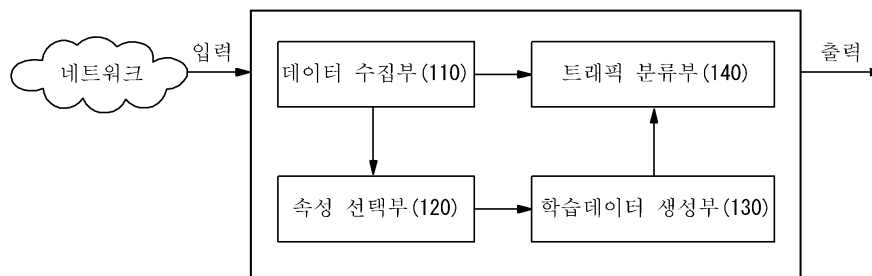
(54) 트래픽 분류 장치 및 방법

(57) 요약

네트워크의 트래픽 데이터를 수집하는 데이터 수집부, 상기 수집된 트래픽 데이터가 갖는 복수의 속성들 중 각 속성들의 상관 관계에 기초하여 중복된 성질을 갖는 속성을 제외하고, 나머지 속성을 기준 속성으로 선택하는 속성 선택부, 상기 기준 속성에 기초하여 상기 수집된 트래픽 데이터로부터 P2P 트래픽 데이터와 비 P2P (non-P2P) 트래픽 데이터를 분류하는 SVM(Support Vector Machine)을 위한 학습 데이터와, 상기 P2P 트래픽 데이터를 애플리케이션 별 트래픽 데이터로 분류하는 SVDD(Support Vector Data Description)를 위한 학습 데이터를 생성하는 학습 데이터 생성부 및 상기 SVM 및 SVDD에 기초하여 상기 네트워크의 트래픽 데이터를 애플리케이션 별로 분류하는 트래픽 분류부를 포함하는 트래픽 분류 장치를 제공한다.

대표도 - 도1

100



(72) 발명자

이한성

경기도 용인시 기흥구 동백동 백현마을서해그랑블
아파트 2607동 2501호

임영희

경기도 수원시 영통구 매탄3동 우남퍼스트빌아파트
208동 702호

김명섭

충청남도 연기군 조치원읍 신흥리 푸르지오1차 10
7동 401호

특허청구의 범위

청구항 1

트래픽 분류 장치에 있어서,

네트워크의 트래픽 데이터를 수집하는 데이터 수집부,

상기 수집된 트래픽 데이터가 갖는 복수의 속성들 중 각 속성들의 상관 관계에 기초하여 중복된 성질을 갖는 속성을 제외하고, 나머지 속성을 기준 속성으로 선택하는 속성 선택부,

상기 기준 속성에 기초하여 상기 수집된 트래픽 데이터로부터 P2P 트래픽 데이터와 비 P2P (non-P2P) 트래픽 데이터를 분류하는 SVM(Support Vector Machine)을 위한 학습 데이터와, 상기 P2P 트래픽 데이터를 애플리케이션 별 트래픽 데이터로 분류하는 SVDD(Support Vector Data Description)를 위한 학습 데이터를 생성하는 학습 데이터 생성부 및

상기 SVM 및 SVDD에 기초하여 상기 네트워크의 트래픽 데이터를 애플리케이션 별로 분류하는 트래픽 분류부를 포함하는 트래픽 분류 장치.

청구항 2

제 1 항에 있어서,

상기 데이터 수집부는 상기 트래픽 데이터를 플로우(flow) 단위로 수집하는 트래픽 분류 장치.

청구항 3

제 1 항에 있어서,

상기 속성 선택부는 제 1 속성 및 제 2 속성에 대한 엔트로피를 산출하고, 제 1 속성이 발생할 때 제 2 속성이 발생하는 조건부 확률을 산출하고, 상기 제 1 속성 및 제 2 속성의 엔트로피 및 상기 조건부 확률에 기초하여 정보 이익을 산출하며, 상기 제 1 속성 및 제 2 속성의 엔트로피 및 상기 정보 이익에 기초하여 제 1 속성과 제 2 속성의 상관 관계를 계산하여 중복된 성질을 갖는 속성을 제외하는 트래픽 분류 장치.

청구항 4

제 1 항에 있어서,

상기 속성 선택부는 전체 속성 집합에 포함되는 속성들의 부분 집합 중 메리트 함수(merit function)의 값이 가장 큰 부분 집합을 상기 속성 집합으로 선택하는 트래픽 분류 장치.

청구항 5

제 1 항에 있어서,

상기 학습 데이터 생성부는 상기 P2P 트래픽 데이터를 세부 클래스로 분류하는 SVDD를 위한 학습 데이터를 생성하고,

상기 세부 클래스에 해당하는 트래픽 데이터를 애플리케이션 별로 분류하는 SVDD를 위한 학습 데이터를 생성하는 트래픽 분류 장치.

청구항 6

제 1 항에 있어서,

상기 학습 데이터 생성부는 상기 비 P2P 트래픽 데이터를 애플리케이션 별로 분류하는 SVDD를 위한 학습 데이터를 생성하는 트래픽 분류 장치.

청구항 7

제 1 항 또는 제 5 항에 있어서,

상기 트래픽 분류부는 상기 SVM에 기초하여 상기 트래픽 데이터를 P2P 트래픽 데이터와 비 P2P 트래픽 데이터로 분류하고,

상기 SVDD에 기초하여 상기 P2P 트래픽 데이터를 애플리케이션 별 트래픽 데이터로 분류하는 트래픽 분류 장치.

청구항 8

제 6 항에 있어서,

상기 트래픽 분류부는 상기 SVDD에 기초하여 상기 비 P2P 트래픽 데이터를 애플리케이션 별 트래픽 데이터로 분류하는 트래픽 분류 장치.

청구항 9

제 7 항에 있어서,

상기 트래픽 분류부는 상기 SVDD에 기초하여 상기 P2P 트래픽 데이터를 파일 공유 트래픽 데이터, 메신저 트래픽 데이터 및 TV 트래픽 데이터로 분류하고,

상기 파일 공유 트래픽 데이터, 메신저 트래픽 데이터 및 TV 트래픽 데이터 각각을 적어도 하나 이상의 애플리케이션 별 트래픽 데이터로 분류하는 트래픽 분류 장치.

청구항 10

트래픽 분류 방법에 있어서,

네트워크의 트래픽 데이터를 수집하는 단계,

상기 수집된 트래픽 데이터가 갖는 복수의 속성들 중 각 속성들의 상관 관계에 기초하여 중복된 성질을 갖는 속성을 제외하고, 나머지 속성을 기준 속성으로 선택하는 단계,

상기 기준 속성에 기초하여 상기 수집된 트래픽 데이터로부터 P2P 트래픽 데이터와 비 P2P (non-P2P) 트래픽 데이터를 분류하는 SVM(Support Vector Machine)을 위한 학습 데이터를 생성하는 단계,

상기 선택된 속성에 기초하여 상기 수집된 트래픽 데이터로부터 상기 P2P 트래픽 데이터를 애플리케이션 별 트래픽 데이터로 분류하는 SVDD(Support Vector Data Description)를 위한 학습 데이터를 생성하는 단계,

상기 SVM에 기초하여 상기 네트워크의 트래픽 데이터를 P2P 트래픽 데이터와 비 P2P (non-P2P) 트래픽 데이터로 분류하는 단계 및

상기 SVDD에 기초하여 상기 P2P 트래픽 데이터를 애플리케이션 별 트래픽 데이터로 분류하는 단계를 포함하는 트래픽 분류 방법.

청구항 11

제 10 항에 있어서,

상기 속성을 선택하는 단계는

제 1 속성 및 제 2 속성에 대한 엔트로피를 산출하는 단계

상기 제 1 속성이 발생할 때 상기 제 2 속성이 발생하는 조건부 확률을 산출하는 단계,

상기 제 1 속성 및 제 2 속성의 엔트로피 및 상기 조건부 확률에 기초하여 정보 이익을 산출하는 단계,

상기 제 1 속성 및 제 2 속성의 엔트로피 및 상기 정보 이익에 기초하여 상기 제 1 속성과 제 2 속성의 상관 관계를 계산하여 중복된 성질을 갖는 속성을 제외하는 단계

를 포함하는 트래픽 분류 방법.

청구항 12

제 10 항에 있어서,

상기 속성을 선택하는 단계는

전체 속성 집합에 포함되는 속성들의 부분 집합 중 메리트 함수(merit function)의 값이 가장 큰 부분 집합을 상기 속성 집합으로 선택하는 단계

를 포함하는 트래픽 분류 방법.

청구항 13

제 10 항에 있어서,

상기 학습 데이터를 생성하는 단계는 상기 P2P 트래픽 데이터를 세부 클래스로 분류하는 SVDD를 위한 학습 데이터를 생성하는 단계 및

상기 세부 클래스에 해당하는 트래픽 데이터를 애플리케이션 별로 분류하는 SVDD를 위한 학습 데이터를 생성하는 단계

를 포함하는 트래픽 분류 방법.

청구항 14

제 10 항에 있어서,

상기 학습 데이터를 생성하는 단계는 상기 비 P2P 트래픽 데이터를 애플리케이션 별로 분류하는 SVDD를 위한 학습 데이터를 생성하는 단계

를 포함하는 트래픽 분류 방법.

청구항 15

제 14 항에 있어서,

상기 SVDD에 기초하여 상기 비 P2P 트래픽 데이터를 애플리케이션 별 트래픽 데이터로 분류하는 단계

를 더 포함하는 트래픽 분류 방법.

명세서

발명의 상세한 설명

기술 분야

[0001] 본 발명은 트래픽 분류 장치 및 방법에 관한 것이다.

배경 기술

[0002] 인터넷의 급속한 발전과 이를 이용하는 사용자의 수가 급증함에 따라 보다 다양하고 새로운 네트워크 서비스들이 개발되어 상용화되고 있다. 따라서 다양한 네트워크 서비스에 보다 적합한 QoS(Quality of Service) 및 안전한 네트워크 환경의 제공을 목적으로 하는 네트워크 망 관리시스템에서, 인터넷 애플리케이션 트래픽 분류는 가장 중요한 기본 기능 중 하나이다. 특히 네트워크 서비스 중 P2P(Peer-to-Peer) 관련 트래픽은 2004년 CacheLogic사의 조사에 의하면 전체 인터넷 트래픽에서의 비중이 60%에 이르고 있으며 현재는 그 비중이 보다 커지고 있다. 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 재분할 등의 추가적 비용을 요구하고 있다. 따라서 P2P를 포함하는 인터넷 애플리케이션 트래픽의 보다 빠르고 정확한 분류가 필요한 실정이다.

[0003] 인터넷 애플리케이션 트래픽을 분류하는 전통적인 방법으로 포트(port) 번호를 이용하는 방법과 페이로드(payload) 정보를 이용하는 방법을 들 수 있다. 포트 번호 기반의 분류방법은 IANA에서 할당한 잘 알려진 포트(well-known port) 번호를 분석하는 비교적 단순하면서도 실용적인 방법이지만, 동적으로 포트 번호를 할당하여 패킷(packet)을 발생하거나 기존에 사용한 잘 알려진 포트 번호를 다른 목적으로 이용하는 최근의 애플리케이션들이 증가함에 따라 정확한 트래픽 분류가 어렵다. 페이로드 정보에 기반 한 분류방법은 페이로드의 특징을 추출하고 이를 패킷과 비교하는 방법으로써, 패킷을 발생할 때 페이로드를 암호화할 경우 이를 사용할 수 없다. 또한 최근에 불법침입의 차단 및 고비용의 접근 등을 이유로 페이로드의 접근을 금지하고 있는 실정에 따라 그 사용이 더욱 어렵다. 따라서 기존의 전통적인 방법에서 벗어난 새로운 방법이 요구된다.

[0004] 최근의 연구문헌 조사에 의하면, 애플리케이션의 변화에 대처할 수 있는 새로운 해결책으로써 데이터 마이닝 및 기계학습 기법을 트래픽 분류에 적용하려는 시도가 진행 중이다. 기계학습기법은 동적으로 변하는 포트 번호와 암호화된 페이로드에 독립적인 데이터의 특성 벡터로부터 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 데이터임을 고려할 때, 대용량의 데이터 처리를 위한 데이터 마이닝 기법의 적용은 매우 적절하다.

[0005] 이러한 연구 방법들 중 패턴분석을 위한 자동화 알고리즘으로서 가장 강력하다고 이미 검증된 서포트 벡터 머신(Support Vector Machine, SVM)을 인터넷 애플리케이션 트래픽 분류에 적용하려는 연구가 주목을 받고 있다. 현재까지의 SVM을 이용한 인터넷 애플리케이션 트래픽 분류는 패턴 분류 및 함수 근사(function approximation) 등의 문제에서 매우 우수한 성능을 보이는 SVM을 사용하는 것으로, 다음의 두 가지 방법을 취하고 있다. 첫 번째 방법에서는 이진 분류기(binary classifier)인 SVM을 이용하여 P2P 트래픽과 비 P2P(non-P2P) 트래픽을 분류하는 방법을 취하고 있으며, 두 번째 방법에서는 다중 클래스 SVM(multi-class SVM)을 구축함으로써 P2P 트래픽의 식별 및 트래픽의 분류를 수행한다. 이와 관련하여, 최근 발명자는 “계층적 다중 클래스 SVM을 이용한 인터넷 애플리케이션 트래픽 분류”(저자 유재학, 김성운, 이한성, 김명섭, 박대희, 2008년 한국컴퓨터종합학술대회 논문집)를 통해 다중 클래스 SVM을 구축하는 대신 단일 클래스를 위한 SVDD를 구축하는 방법을 제시하였다.

발명의 내용

해결 하고자하는 과제

[0006] 본 발명의 일부 실시예는 트래픽 데이터의 여러 속성들 중 최적으로 선택된 일부 속성을 기초로 하여 학습 데이터를 생성하고, 그를 기초로 트래픽 데이터를 분류할 수 있는 트래픽 분류 장치 및 방법을 제공한다.

[0007] 본 발명의 또 다른 실시예는 SVM과 SVDD를 사용하여 트래픽 데이터를 애플리케이션 별로 분류할 수 있는 트래픽 분류 장치 및 방법을 제공한다.

과제 해결수단

[0008] 상술한 기술적 과제를 달성하기 위한 기술적 수단으로서, 본 발명의 제 1 측면은 네트워크의 트래픽 데이터를 수집하는 데이터 수집부, 상기 수집된 트래픽 데이터가 갖는 복수의 속성들 중 각 속성들의 상관 관계에 기초하여 중복된 성질을 갖는 속성을 제외하고, 나머지 속성을 기준 속성으로 선택하는 속성 선택부, 상기 기준 속성에 기초하여 상기 수집된 트래픽 데이터로부터 P2P 트래픽 데이터와 비 P2P (non-P2P) 트래픽 데이터를 분류하는 SVM(Support Vector Machine)을 위한 학습 데이터와, 상기 P2P 트래픽 데이터를 애플리케이션 별 트래픽 데이터로 분류하는 SVDD(Support Vector Data Description)를 위한 학습 데이터를 생성하는 학습 데이터 생성부 및 상기 SVM 및 SVDD에 기초하여 상기 네트워크의 트래픽 데이터를 애플리케이션 별로 분류하는 트래픽 분류부를 포함하는 트래픽 분류 장치를 제공한다.

[0009] 또한, 본 발명의 제 2 측면은 네트워크의 트래픽 데이터를 수집하는 단계, 상기 수집된 트래픽 데이터가 갖는 복수의 속성들 중 각 속성들의 상관 관계에 기초하여 중복된 성질을 갖는 속성을 제외하고, 나머지 속성을 기준 속성으로 선택하는 단계, 상기 기준 속성에 기초하여 상기 수집된 트래픽 데이터로부터 P2P 트래픽 데이터와 비 P2P (non-P2P) 트래픽 데이터를 분류하는 SVM(Support Vector Machine)을 위한 학습 데이터를 생성하는 단계, 상기 선택된 속성에 기초하여 상기 수집된 트래픽 데이터로부터 상기 P2P 트래픽 데이터를 애플리케이션 별 트래픽 데이터로 분류하는 SVDD(Support Vector Data Description)를 위한 학습 데이터를 생성하는 단계, 상기 SVM에 기초하여 상기 네트워크의 트래픽 데이터를 P2P 트래픽 데이터와 비 P2P (non-P2P) 트래픽 데이터로 분류하는 단계 및 상기 SVDD에 기초하여 상기 P2P 트래픽 데이터를 애플리케이션 별 트래픽 데이터로 분류하는 단계를 포함하는 트래픽 분류 방법을 제공한다.

효과

[0010] 진술한 본 발명의 과제 해결 수단에 의하면, 실시간으로 P2P와 non-P2P 트래픽을 분류하기 때문에 보다 안정적인 네트워크 환경을 지원할 수 있다. 또한, P2P 트래픽의 대표적인 3가지 유형별 분류가 가능함으로써 보다 적합한 QoS의 보장이 가능하다. 또한, 트래픽의 세분화된 분류로 보다 원활한 대역폭의 사용 및 효율적 시스템 자원관리의 지원도 가능하다. 또한, 새로운 애플리케이션 트래픽이 추가될 때, 전체 시스템의 재학습이 아닌 해당 애플리케이션 트래픽 클래스에 해당하는 모듈만을 추가 학습하는 점증적 갱신 (incremental updating)이 가능하다. 또한, 최적 속성의 선택을 통해 데이터의 차원을 줄여 패턴 인식기의 수행 시간을 단축시키고 분류 성능을 향상시키는 효과가 있다. 이와 같이 전체 트래픽 데이터를 그 속성에 따라 분류할 수 있으므로, 분류된 트래픽 데이터에 대하여 각각 서로 다른 우선 순위를 부여하는 방식으로 트래픽 데이터를 관리할 수 있고, 그에 따라 네트워크 관리에 소요되는 자원을 최적화할 수 있다.

발명의 실시를 위한 구체적인 내용

[0011] 아래에서는 첨부한 도면을 참조하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 본 발명의 실시예를 상세히 설명한다. 그러나 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시예에 한정되지 않는다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.

[0012] 명세서 전체에서, 어떤 부분이 다른 부분과 "연결"되어 있다고 할 때, 이는 "직접적으로 연결"되어 있는 경우뿐 아니라, 그 중간에 다른 소자를 사이에 두고 "전기적으로 연결"되어 있는 경우도 포함한다. 또한 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있는 것을 의미한다.

[0013] 도 1은 본 발명의 일 실시예에 따른 트래픽 분류 장치를 도시한 도면이다.

[0014] 트래픽 분류 장치(100)는 데이터 수집부(110), 속성 선택부(120), 학습 데이터 생성부(130), 트래픽 분류부

(140)를 포함한다.

- [0015] 참고로, 본 발명의 실시예에 따른 도 1에 도시된 구성 요소들은 소프트웨어 또는 FPGA(Field Programmable Gate Array) 또는 ASIC(Application Specific Integrated Circuit)와 같은 하드웨어 구성 요소를 의미하며, 소정의 역할들을 수행한다.
- [0016] 그렇지만 '구성 요소들'은 소프트웨어 또는 하드웨어에 한정되는 의미는 아니며, 각 구성 요소는 어드레싱할 수 있는 저장 매체에 있도록 구성될 수도 있고 하나 또는 그 이상의 프로세서들을 재생시키도록 구성될 수도 있다.
- [0017] 따라서, 일 예로서 구성 요소는 소프트웨어 구성 요소들, 객체지향 소프트웨어 구성 요소들, 클래스 구성 요소들 및 태스크 구성 요소들과 같은 구성 요소들과, 프로세스들, 함수들, 속성들, 프로시저들, 서브루틴들, 프로그램 코드의 세그먼트들, 드라이버들, 펌웨어, 마이크로 코드, 회로, 데이터, 데이터베이스, 데이터 구조들, 테이블들, 어레이들 및 변수들을 포함한다.
- [0018] 구성 요소들과 해당 구성 요소들 안에서 제공되는 기능은 더 작은 수의 구성 요소들로 결합되거나 추가적인 구성 요소들로 더 분리될 수 있다.
- [0019] 데이터 수집부(110)는 네트워크를 통해 수신된 트래픽 데이터를 수집한다. 이때, 네트워크는 근거리 통신망(Local Area Network; LAN), 광역 통신망(Wide Area Network; WAN) 또는 부가가치 통신망(Value Added Network; VAN) 등과 같은 유선 네트워크나 이동 통신망(mobile radio communication network) 또는 위성 통신망 등과 같은 모든 종류의 무선 네트워크를 포함한다.
- [0020] 수집된 트래픽 데이터는 속성 선택부(120) 및 트래픽 분류부(140)로 전달된다. 속성 선택부(120)로 전달된 트래픽 데이터는 최적의 속성을 선택하는 과정에서 표본 역할을 수행하게 된다. 그리고 이는 다시 학습 데이터 생성부(130)로 전달되어 실제 트래픽 데이터의 분류 기준이 될 학습 데이터를 생성하는 과정에서 표본 역할을 수행하게 된다.
- [0021] 학습 데이터의 생성이 완료되면, 데이터 수집부(110)는 트래픽 데이터를 수집하여 트래픽 분류부(140)로만 전달하고, 속성 선택부(120), 학습 데이터 생성부(130)로는 트래픽 데이터를 전달하지 않을 수 있다.
- [0022] 학습 데이터에 기초하여 트래픽 데이터를 분류하는 과정은 온라인 상태로 반복적으로 수행된다. 그러나 최적의 속성을 선택하는 과정과 학습 데이터를 생성하는 과정은 오프라인 상태로 한정된 시간 동안 수행될 수 있다.
- [0023] 한편, 추후 설명할 트래픽 데이터의 분류 클래스가 추가되는 경우, 해당 클래스만을 위한 학습 데이터 생성 과정을 더 수행할 수 있다.
- [0024] 본 발명에서는 트래픽 데이터의 분류를 위해서 두 가지 사항을 고려한다.
- [0025] 첫 번째, 애플리케이션 트래픽 분류의 기준을 규정하는 것에 대한 사항이다. 응용레벨 프로토콜을 기준으로 분류할 것인지, 스트리밍이나 파일 다운로드와 같이 트래픽의 타입을 기준으로 할 것인지, 아니면 애플리케이션으로 할 것인지에 대한 선택이다.
- [0026] 애플리케이션 트래픽 분류 규정은 분석 목적에 따라 달라질 수 있지만 본 발명에서는 다음과 같은 원칙으로 애플리케이션 트래픽을 분류한다. 기본적으로 애플리케이션의 관점에서 해당 애플리케이션이 발생시킨 모든 트래픽을 그 애플리케이션의 트래픽으로 규정한다. 여러 애플리케이션에서 공유하여 사용하는 공개된 응용레벨 프로토콜의 경우는 응용레벨 프로토콜의 단위로 분리하여 분석한다. 이렇게 분류하는 것은 QoS, IPS, IDS와 같은 트래픽 제어 시스템의 제어 기준과도 일치하므로 향후 분석결과를 그대로 또는 최소한의 수정으로 활용할 수 있다.
- [0027] 두 번째, 애플리케이션 트래픽 분류를 위한 데이터의 처리 단위를 결정하는 것에 대한 사항이다. IP 네트워크에서 트래픽 데이터의 최소 단위인 패킷 단위로 학습 또는 분류할 수도 있고, 패킷의 집합인 플로우(flow)를 처리 단위로 할 수도 있다. 본 발명에서는 플로우를 처리 단위로 한다.
- [0028] 이때, 플로우 단위의 트래픽 데이터는 패킷의 5-tuple 정보(source IP, source port, destination IP, destination port, protocol number)를 공유하는 양방향 패킷들의 집합으로 정의한다. 양방향 패킷들의 집합으로 정의되는 플로는 하나의 접속(connection)에서 입력 플로우(in-flow)와 출력 플로우(out-flow)의 특징이 서로 다르게 나타난다. 도면을 참조하여 설명하기로 한다.
- [0029] 도 2는 본 발명에 따른 플로우 단위의 트래픽 데이터를 설명하기 위한 도면이다.

- [0030] 클라이언트(210)와 서버(220)가 네트워크를 통해 접속된 시스템이 도시되어 있다. TCP 방식의 경우 클라이언트(210)에서 서버(220)로의 패킷(pkt)을 입력 플로우로, 서버(220)에서 클라이언트(210)로의 패킷을 출력 플로우로 결정하여 SYN 패킷에 의한 연결시작에서부터 FIN 패킷이 발생하는 연결 끝 사이에 발생한 모든 양방향 패킷들을 하나의 플로우로 정의한다.
- [0031] UDP 방식의 경우에는 클라이언트와 서버의 관계를 정의할 수 없기 때문에 최초 패킷 발생 시점에서 마지막 패킷이 발생한 시점 사이의 모든 양방향 패킷들을 하나의 플로우로 정의한다.
- [0032] 다시 도 1을 참조하면, 속성 선택부(120)는 트래픽 데이터가 갖는 전체 속성에 의한 확률 분포도를 가능한 가깝게 표현할 수 있는 최소 개수의 속성을 선택한다. 즉, 최초의 속성 집합 D로부터 거의 사용되지 않거나 중복된 성질을 갖는 속성을 제외한 속성 집합 d를 찾는다. 이때, 선택된 속성을 기준 속성이라 한다. 이를 위해, 트래픽 데이터가 갖는 복수의 속성들 중 각 속성들의 상관 관계에 기초하여 중복된 성질을 갖는 속성을 제외하고, 나머지 속성을 기준 속성으로서 선택하게 된다. 이는 가능한 최소한의 성능 저하를 유지하면서, 정확한 분류기를 제공하기 위함이다. 본 발명에서는 최적우선탐색(best first search) 방법과 속성 혹은 특징(attribute or feature) 값 Y에 대한 엔트로피(entropy), 목표 클래스(target class)와 속성들 간의 피어슨 상관 계수(Pearson's correlation coefficient)를 이용한 조건부 확률(conditional probability)을 계산하여 전체 속성들의 확률 분포도를 가능한 가깝게 표현할 수 있는 최소 개수의 속성집합을 찾도록 한다.
- [0033] 도 3은 트래픽 데이터가 갖는 속성의 예를 설명하기 위한 도면이다.
- [0034] 입력 플로우와 출력 플로우 별로 각각, 일정 기간 동안 전송된 패킷량 또는 바이트량을 나타내는 속성, 시간 간격을 나타내는 속성, SYN 패킷, ACK 패킷, RST 패킷 또는 FIN 패킷의 번호 등을 나타내는 속성이 있을 수 있다.
- [0035] 또한, 패킷의 최저값, 최고값, 평균값 또는 표준 편차값 등을 나타내는 속성, 윈도우(window)의 최저값, 최고값, 평균값 또는 표준 편차값 등을 나타내는 속성, 지터(jitter)의 최저값, 최고값, 평균값 또는 표준 편차값 등을 나타내는 속성이 있을 수 있다. 이와 같은 속성들 중에서 트래픽 데이터를 분류할 최적의 속성을 선택하게 된다. 이때, 도시된 속성 외에 다른 속성을 기초로 트래픽 데이터를 분류할 수 있으며, 본 명세서에 제시된 속성의 종류가 본 발명을 제한하는 것은 아니다.
- [0036] 먼저, 각 속성들에 대한 정보 이익(information gain)을 얻기 위해 임의의 속성 Y에 대한 엔트로피를 수학적 1을 통해 산출한다.

수학적 1

$$H(Y) = - \sum_{y \in Y} p(y) \log_2(p(y))$$

[0037]

- [0038] 속성 X와 Y사이의 관계는 X가 주어졌을 때 Y가 발생하는 조건부 확률로써 수학적 2와 같이 계산된다.

수학적 2

$$H(Y|X) = - \sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log_2(p(y|x))$$

[0039]

- [0040] 각 특징에 대한 정보 이익은 수학적 1과 수학적 2를 이용하여 수학적 3으로 정의된다.

수학적 3

$$Gain = H(Y) + H(X) - H(X, Y)$$

[0041]

- [0042] 수학적 3에서 얻은 정보 이익을 기반으로 수학적 4에서와 같이 대칭적 불확실성(symmetrical uncertainty)을 이용하여 임의의 두 속성 X와 Y의 분포와 상관관계를 계산한다. 이때 속성 X를 기준으로 Y가 높은 분포와 상관관

계를 보이면 전체 속성들을 효율적으로 표현할 수 있는 부분집합에 속성 X는 포함되지만 Y는 포함되지 않는다. 즉, 서로 중복되는 성향을 가진 두 가지 속성이 있는 경우 하나의 속성만을 포함시키도록 한다. 마찬가지로 목표 클래스와 속성들 간의 분포와 상관관계를 계산하여 부분집합을 구성한다.

수학식 4

$$\text{symmetrical uncertainty coefficient} = 2.0 \times \left[\frac{\text{Gain}}{H(Y) + H(X)} \right]$$

각각의 부분집합 $F_S \subset F$ 이 전체 속성들을 얼마나 효율적으로 표현하는지를 평가하기 위하여 메리트 함수 (merit function)인 수학식 5를 사용한다. 메리트 함수의 값이 가장 큰 부분집합이 전체 속성들을 최적으로 표현할 수 있는 부분집합으로 결정된다.

수학식 5

$$\text{Merit}(F_S) = \frac{k \overline{r_{cf}}}{\sqrt{k + k(k-1) \overline{r_{ff}}}}$$

여기서, k는 부분집합 F_S 에서의 속성들의 개수를 의미하며, $\overline{r_{cf}}$ 는 F_S 에 포함된 속성의 평균 분포 (contribution), $\overline{r_{ff}}$ 는 속성의 평균 상관관계 값이다.

이와 같은 과정에 의하여 최적의 속성 집합이 선택된다.

학습 데이터 생성부(130)는 상기 선택된 기준 속성에 기초하여 애플리케이션 별 트래픽 데이터를 분류하는 학습 데이터(training data)를 생성한다. 본 발명에서는 P2P 트래픽 데이터와 비 P2P 트래픽 데이터를 분류하는 SVM 과, P2P 트래픽 데이터 및 비 P2P 트래픽 데이터를 애플리케이션 별로 분류하는 SVDD(Support Vector Data Description)를 이용한다. 따라서 SVM을 위한 학습 데이터와 SVDD를 위한 학습 데이터를 각각 생성한다. 이는 데이터 수집부(110)를 통해 수집된 트래픽 데이터 들 중 속성 선택부(120)에 의하여 선택된 속성들을 기초로 하여 수행된다.

먼저 SVM을 위한 학습 데이터를 생성하는 과정을 설명한다.

도 4는 본 발명에 적용되는 SVM을 구하는 과정을 설명하기 위한 도면이다.

통계적 학습이론 (statistical learning theory)에 기반을 둔 SVM은 주어진 문제를 항상 전역적 최적해가 보장 되는 컨벡스 쿼드라틱(convex quadratic) 문제로 변환하여 해를 구하기 때문에 패턴인식 분야에서 매우 우수한 성능을 보여주고 있다. 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. 두 집합을 분류하기 위한 모델을 정의하기 위하여 도 4와 같은 선형 식별함수인 초평면 (hyperplane)을 정의할 수 있다. 최적 경계 초평면이 학습 데이터의 분류 함수로 주어질 때, 초평면에서 학습 데이터와의 거리 r은 수학식 6과 같다.

수학식 6

$$r = \frac{|w^T x + b|}{\|w\|} \geq \frac{1}{\|w\|}$$

[0053] 이 하한 값의 거리에 있는 데이터는 최적 경계 초평면과 가장 가까운 거리에 위치하게 된다. 이 데이터들을 서포트 벡터(support vector)라고 부른다. 따라서 최적 경계 초평면에 의해 분류되는 두 클래스간의 거리는 수학적 식 7과 같게 된다.

수학적 식 7

$$\rho = 2r = \frac{2}{\|w\|}$$

[0054]

[0055] 이때 ρ 를 분류 한도 (margin of separation)라 정의한다. 최적의 초평면을 구하는 문제는 한도(margin)를 최대화하는 문제로 정의될 수 있으며, 이 경우 SVM의 학습은 다음의 최적화 문제로 정의된다.

수학적 식 8

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

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

[0056]

[0057] 수학적 식 8에 관한 쌍대 문제(dual problem)를 구하기 위하여 라그랑제 함수 (Largrange function) α 을 도입한다.

수학적 식 9

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

[0058]

[0059] 학습 종료 후, 입력된 트래픽 데이터가 어떤 클래스인지의 소속 여부를 판정하는 결정함수는 수학적 식 10과 같이 정의된다.

수학적 식 10

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

[0060]

[0061] 수학적 식 10은 트래픽 분류부(140)가 트래픽 데이터를 분류하는 과정에서 사용된다.

[0062] 다음으로, SVDD를 위한 학습 데이터를 생성하는 과정을 설명한다.

[0063] 이진 분류기라는 SVM의 기능적 한계점으로 인하여, 주어진 문제가 현재 우리가 다루고자 하는 트래픽 분류와 같이 다중 분류 문제인 경우에는 SVM을 직접적으로 적용할 수가 없다. 따라서 여러 개의 이진 분류기인 SVM을 유기적으로 결합하여 다중 클래스 SVM을 설계하는 것이 일반적인 연구방법론이다. 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)로써 결정 경계면을 선택하는 것이 다중 클래스 SVM의 설계 시 보다 유리하다. 따라서 본 발명에서는 단일 클래스 SVM의 알고리즘인 SVDD(Support Vector Data Description)를 기반으로 다중 클래스 SVM을 설계하여 인터넷 애플리케이션 트래픽 유형들을 자동 분류하는 새로운 시스템을 제안한다.

[0064] 먼저, 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 분류하

기 위한 분류기는 각 클래스의 학습 데이터를 최대한 많이 포함하면서 동시에 체적을 최소화하는 구체(sphere)를 구하는 문제로 정의되며, 수학식 11의 최적화 문제를 통하여 수식화 된다.

수학식 11

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

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

a_k : k -번째 클래스를 표현하는 구체의 중심

R_k^2 : 구체의 반경의 제곱

ξ_i^k : k -번째의 클래스에 속한 i -번째 학습 데이터 x_i^k 가 구체에서 벗어나는 정도를 나타내는 벌점 항

C : 상대적 중요성을 조정하는 상수(trade-off constant)

[0065]

[0066] 수학식 11에 관한 쌍대 문제를 구하기 위하여 라그랑제 함수 (Largrange function) $\square$ 을 도입한다.

수학식 12

$$L(R_k^2, a_k, \xi_k, \alpha_k, \eta_k)$$

$$= R_k^2 + C \sum_{i=1}^{N_k} \xi_i^k$$

$$+ \sum_{i=1}^{N_k} \alpha_i^k [(x_i^k - a_k)^T (x_i^k - a_k) - R_k^2 - \xi_i^k]$$

$$- \sum_{i=1}^{N_k} \eta_i^k \xi_i^k$$

단, $\alpha_i^k \geq 0, \eta_i^k \geq 0, \forall i.$

[0067]

[0068] 단, 수학식 12는 R_k^2, a_k, ξ_k 변수에 대해서는 최소값을, 변수 α_k, η_k 에 대해서는 최대값을 가져야 하므로, 아래의 수학식 13을 만족해야 한다.

수학식 13

$$\frac{\partial L}{\partial R_k^2} = 0 : \sum_{i=1}^{N_k} \alpha_i^k = 1.$$

$$\frac{\partial L}{\partial \xi_i^k} = 0 : C - \alpha_i^k - \eta_i^k = 0 \therefore \alpha_i^k \in [0, C], \forall i.$$

$$\frac{\partial L}{\partial a_k} = 0 : a_k = \sum_{i=1}^{N_k} \alpha_i^k x_i^k$$

[0069]

[0070] 수학식 13을 라그랑제 함수에 대입하면, 다음의 쌍대 문제를 얻는다.

수학식 14

$$\min_{\alpha_k} \sum_{i=1}^{N_k} \sum_{j=1}^{N_k} \alpha_i^k \alpha_j^k \langle x_i^k, x_j^k \rangle - \sum_{i=1}^{N_k} \alpha_i^k \langle x_i^k, x_i^k \rangle$$

$$s.t. \quad \sum_{i=1}^{N_k} \alpha_i^k = 1, \alpha_i^k \in [0, C], \forall_i$$

[0071]

[0072]

입력 공간 위에서 정의되는 구체는 매우 간단한 형태의 영역만을 나타낼 수 있다. 이러한 한계를 극복하기 위하여 커널 함수(kernel function)를 통하여 정의되는 고차원의 특징 공간(feature space)위에서 정의되는 구체를 사용하는 방향으로 확장될 수 있다. 각각의 클래스는 각자의 특징공간에서 자신의 경계를 보다 정확하게 표현할 수 있으므로, 시스템의 학습은 각각의 클래스들이 매핑되는 특징 공간의 독립성을 고려하여 아래의 컨벡스 쿼드릭(convex quadratic)문제의 해답을 얻음으로써 이루어진다.

수학식 15

$$\min_{\alpha_k} \sum_{i=1}^{N_k} \sum_{j=1}^{N_k} \alpha_i^k \alpha_j^k k_k(x_i^k, x_j^k) - \sum_{i=1}^{N_k} \alpha_i^k k_k(x_i^k, x_i^k)$$

$$s.t. \quad \sum_{i=1}^{N_k} \alpha_i^k = 1, \alpha_i^k \in [0, C], \forall_i$$

[0073]

[0074]

특히, 가우시안 커널(Gaussian kernel)을 사용할 경우, $k(x, x)=1$ 이 성립하므로, 수학식 15는 수학식 16과 같이 단순화 된다.

수학식 16

$$\min_{\alpha_k} \sum_{i=1}^{N_k} \sum_{j=1}^{N_k} \alpha_i^k \alpha_j^k k_k(x_i^k, x_j^k)$$

$$s.t. \quad \sum_{i=1}^{N_k} \alpha_i^k = 1, \alpha_i^k \in [0, C], \forall_i.$$

[0075]

[0076]

학습 종료 후 트래픽 데이터를 분류하는 동작에 적용하면, 각 클래스의 결정함수는 수학식 17과 같이 정의된다.

수학식 17

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

[0077]

[0078]

서로 다른 특징 공간상에서 정의되는 단일 SVM의 출력 $f_k(x)$ 값은 각 클래스의 특징 공간상의 경계로부터 해당 트래픽 데이터와의 절대 거리를 의미하므로, 서로 다른 특징 공간상의 절대거리를 비교하여 소속 클래스를 결정하는 것은 바람직하지 않다. 따라서 특징 공간상의 절대거리 $f_k(x)$ 를 특징 공간상에서 정의되는 구형체의 반경 R_k 로 나눔으로서 상대적 거리 $\hat{T}_k(x) = f_k(x)/R_k$ 를 계산하고, 상대거리가 가장 큰 클래스를 입력 데이터 x의 소속 클래스로 결정한다.

수학식 18

$$\begin{aligned} \text{Class of } x &\equiv \operatorname{argmax}_{k=1,\dots,K} \hat{f}_k(x) \\ &\equiv \operatorname{argmax}_k \left[\left\{ R_k^2 - \left(1 - 2 \sum_{i=1}^{N_k} \alpha_i^k k_k(x_i^k, x) + \sum_{i=1}^{N_k} \sum_{j=1}^{N_k} \alpha_i^k \alpha_j^k k_k(x_i^k, x_j^k) \right) \right\} / R_k \right] \end{aligned}$$

[0079]

[0080]

수학식 17과 수학식 18은 트래픽 분류부(140)가 트래픽 데이터를 분류하는 과정에서 사용된다. 이와 같이 각 클래스 별로 별도의 SVDD를 이용하여 트래픽 데이터를 분류하는 방법은 새로운 애플리케이션 트래픽이 추가될 때, 전체 시스템의 재학습이 아닌 해당 애플리케이션 트래픽 클래스에 해당하는 모듈만을 추가 학습하는 점증적 갱신(incremental updating)을 가능하게 한다. 예를 들어, 분류하고자 하는 클래스의 개수가 n 개이고, 이에 대하여 n 개의 SVDD를 위한 학습 데이터 생성이 완료되었다고 가정한다. 만약, SVM을 유기적으로 결합한 다중 클래스 SVM을 이용할 경우, 분류하고자 하는 클래스의 개수를 하나 더 추가하려면, 총 n+1 개의 클래스에 대해서 다시 학습 데이터를 생성해야 한다. 그러나 SVDD를 이용한 분류 방식에서는 추가하고자 하는 클래스에 대해서만 학습 데이터를 생성하면 되므로, 자원의 소모를 최소화할 수 있다.

[0081]

트래픽 분류부(140)는 학습 데이터에 기초하여 상기 네트워크의 트래픽 데이터를 애플리케이션 별로 분류한다. 앞서 설명한 바와 같이, 이진 분류를 위한 SVM에 따라 P2P 트래픽 데이터와 비 P2P 트래픽 데이터를 분류하고, SVDD를 기초로 하여 P2P 트래픽 데이터와 비 P2P 트래픽 데이터를 분류한다. 도면을 참조하여 분류 과정을 더욱 상세히 살펴보기로 한다.

[0082]

도 5는 본 발명의 일 실시예에 따른 트래픽 데이터의 분류 방법을 설명하기 위한 도면이다.

[0083]

먼저, SVM을 이용하여 P2P 트래픽 데이터와 비 P2P 트래픽 데이터를 분류한다. 이때, SVM은 데이터 수집부(110)를 통하여 수집된 트래픽 데이터를 기반으로 학습을 통해 생성된 것이다. 또한, 수집된 트래픽 데이터의 각종 속성들 중 속성 선택부(120)를 통해 선택된 최적의 속성만을 기반으로 하여 학습을 수행한다. 즉, P2P 트래픽 데이터와 비 P2P 트래픽 데이터를 분류할 수 있는 여러 속성들 중 대표적인 속성만을 선택하고, 그를 기초로 하여 학습을 수행한다. 이와 같은 SVM의 학습 데이터를 실제 트래픽 데이터의 분류 과정에 이용하여 P2P 트래픽 데이터와 비 P2P 트래픽 데이터를 분류한다.

[0084]

다음으로, SVDD를 이용하여 P2P 트래픽 데이터를 세부 클래스로 분류한다.

[0085]

이때, P2P 트래픽 데이터를 분류하는 클래스는 계층화 구조를 갖도록 할 수 있다. 즉, 제 1 레벨에 해당하는 복수의 클래스로 분류하고, 제 1 레벨 보다 하위 레벨인 제 2 레벨에 해당하는 복수의 클래스로 분류할 수 있다. 도면에서는 파일 공유 클래스, 메신저 클래스, TV 클래스가 상위 레벨에 해당한다. 또한, 하위 레벨에 해당하는 애플리케이션 별 클래스가 구비된다. 이 각각의 클래스 별로 SVDD의 학습 데이터가 생성된다.

[0086]

이때, 파일 공유 클래스는 피어 간의 파일 공유 과정에서 발생하는 트래픽 데이터를 분류하기 위한 클래스이며, 메신저 클래스는 메신저 프로그램의 사용 과정에서 발생하는 트래픽 데이터를 분류하기 위한 클래스이다. 또한, TV 클래스는 IPTV 프로그램의 사용 과정에서 발생하는 트래픽 데이터를 분류하기 위한 클래스이다. 즉, P2P 트래픽 데이터를 그 속성에 따라 몇 가지 클래스로 분류하되, 그 속성 개수는 최소화시키고, 하나의 클래스만을 담당하는 SVDD를 이용하여 P2P 트래픽 데이터를 분류한다. P2P 트래픽 데이터를 세 가지 유형으로 분류하는 경우에는 세 개의 SVDD를 필요로 하게 된다.

[0087]

다음으로, 세부 클래스로 분류된 각각의 P2P 트래픽 데이터를 애플리케이션 별로 분류한다. 즉, P2P 트래픽 데이터를 애플리케이션의 속성에 따라 분류하되, 그 속성의 개수는 최소화시키고, 하나의 클래스만을 담당하는 SVDD를 이용하여 P2P 트래픽 데이터를 분류한다. 도시된 바와 같이, 파일 공유 트래픽 데이터 중에서도 각 애플리케이션(bittorrent, edonky, fileguri, limewire, morpheus)별로 SVDD를 생성하여 트래픽 데이터를 분류한다. 또한, 메신저 트래픽 데이터 중에서도 각 애플리케이션(msnsg, nation, skype) 별로 SVDD를 생성하여 트래픽 데이터를 분류한다. 또한 TV 트래픽 데이터 중에서도 각 애플리케이션(afreeca player, joost, pplive) 별로 SVDD를 생성하여 트래픽 데이터를 분류한다.

[0088]

한편, 비 P2P 트래픽 데이터에 대해서도 각 애플리케이션(gom, gorealra, iexplore, pop3, ftp) 별로 SVDD를 생성하여 트래픽 데이터를 분류한다. 이와 같은 구성에 따라 각 애플리케이션 별로 트래픽 데이터를 분류할 수 있다.

- [0089] 도 6은 본 발명의 일 실시예에 따른 트래픽 데이터 분류 방법을 도시한 순서도이다.
- [0090] 먼저, 네트워크를 통해 수신된 트래픽 데이터를 플로우 단위로 수집한다(S610). 수집된 트래픽 데이터의 일부는 최적 속성 선택, SVM 및 SVDD를 위한 학습 데이터 생성과정에 사용된다. 학습이 완료된 후에는 수집된 트래픽 데이터를 생성된 학습 데이터에 기초하여 분류한다.
- [0091] 다음으로, 수집된 트래픽 데이터가 갖는 전체 속성에 의한 확률 분포도를 가능한 가깝게 표현할 수 있는 최소 개수의 속성을 선택한다(S620). 앞서 설명한 수학적 식 1 내지 5를 수행하여 최적의 속성 집합을 선택할 수 있다.
- [0092] 다음으로, 트래픽 데이터를 P2P 트래픽 데이터와 비 P2P 트래픽 데이터로 분류할 SVM에 사용될 학습 데이터를 생성한다(S630). 앞서 설명한 수학적 식 6 내지 9를 수행하여 학습 데이터를 생성한다.
- [0093] 다음으로, P2P 트래픽 데이터를 복수의 클래스로 분류할 SVDD에 사용될 학습 데이터를 생성한다(S640). 앞서 설명한 수학적 식 11 내지 16를 수행하여 학습 데이터를 생성한다. 도 5에 도시된, 과일 공유 클래스, 메신저 클래스, TV 클래스 및 각 애플리케이션 별 클래스에 해당하는 트래픽 데이터를 분류하기 위해서는 총 17 개의 SVDD를 위한 학습 데이터를 생성하게 된다.
- [0094] 다음으로, SVM을 기초로 하여 트래픽 데이터를 P2P 트래픽 데이터와 비 P2P 트래픽 데이터로 분류한다(S650). 앞선 단계(S620, S630)를 통해 생성된 학습 데이터를 이용하는 SVM에 따라 이진 분류를 수행하며, 수학적 식 10을 이용하여 분류를 수행한다.
- [0095] 다음으로, SVDD를 기초로 하여 P2P 트래픽 데이터 및 비 P2P 트래픽 데이터를 애플리케이션 별로 분류한다(S660). 앞선 단계(S620, S640)를 통해 생성된 학습 데이터를 이용하는 SVDD에 따라 각 클래스 별로 트래픽 데이터를 분류하며, 수학적 식 17 및 18을 이용하여 분류를 수행한다. 이때, 실시예에 따라 애플리케이션 별 클래스보다 상위 레벨의 클래스 단위로 트래픽 데이터 분류를 수행한 후에 애플리케이션 별로 트래픽 데이터를 분류할 수도 있다.
- [0096] 이와 같이 전체 트래픽 데이터를 그 속성에 따라 분류할 수 있으므로, 분류된 트래픽 데이터에 대하여 각각 서로 다른 우선 순위를 부여하는 방식으로 트래픽 데이터를 관리할 수 있고, 그에 따라 네트워크 관리에 소요되는 자원을 최적화할 수 있다.
- [0097] 본 발명의 일 실시예는 컴퓨터에 의해 실행되는 프로그램 모듈과 같은 컴퓨터에 의해 실행가능한 명령어를 포함하는 기록 매체의 형태로도 구현될 수 있다. 컴퓨터 판독 가능 매체는 컴퓨터에 의해 액세스될 수 있는 임의의 가용 매체일 수 있고, 휘발성 및 비휘발성 매체, 분리형 및 비분리형 매체를 모두 포함한다. 또한, 컴퓨터 판독 가능 매체는 컴퓨터 저장 매체 및 통신 매체를 모두 포함할 수 있다. 컴퓨터 저장 매체는 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 또는 기타 데이터와 같은 정보의 저장을 위한 임의의 방법 또는 기술로 구현된 휘발성 및 비휘발성, 분리형 및 비분리형 매체를 모두 포함한다. 통신 매체는 전형적으로 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈, 또는 반송파와 같은 변조된 데이터 신호의 기타 데이터, 또는 기타 전송 매커니즘을 포함하며, 임의의 정보 전달 매체를 포함한다.
- [0098] 본 발명의 방법 및 시스템은 특정 실시예와 관련하여 설명되었지만, 그것들의 구성 요소 또는 동작의 일부 또는 전부는 범용 하드웨어 아키텍처를 갖는 컴퓨터 시스템을 사용하여 구현될 수 있다.
- [0099] 전술한 본 발명의 설명은 예시를 위한 것이며, 본 발명이 속하는 기술분야의 통상의 지식을 가진 자는 본 발명의 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 쉽게 변형이 가능하다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해해야만 한다. 예를 들어, 단일형으로 설명되어 있는 각 구성 요소는 분산되어 실시될 수도 있으며, 마찬가지로 분산된 것으로 설명되어 있는 구성 요소들도 결합된 형태로 실시될 수 있다.
- [0100] 본 발명의 범위는 상기 상세한 설명보다는 후술하는 특허청구범위에 의하여 나타내어지며, 특허청구범위의 의미 및 범위 그리고 그 균등 개념으로부터 도출되는 모든 변경 또는 변형된 형태가 본 발명의 범위에 포함되는 것으로 해석되어야 한다.

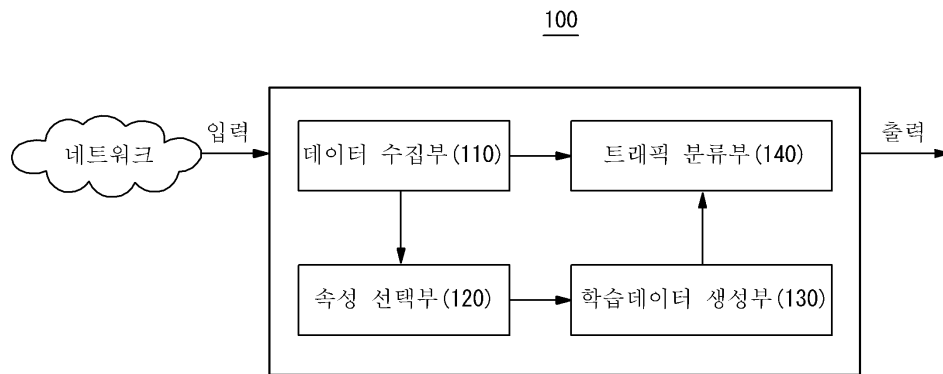
도면의 간단한 설명

- [0101] 도 1은 본 발명의 일 실시예에 따른 트래픽 분류 장치를 도시한 도면이다.

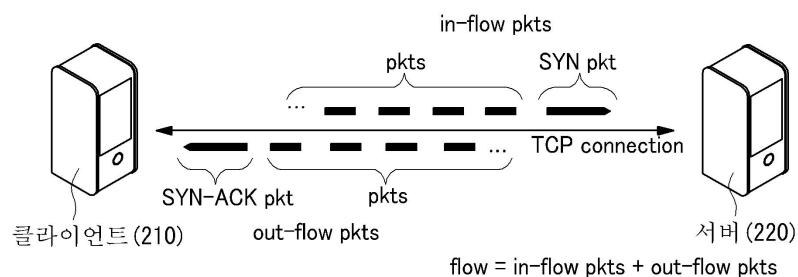
- [0102] 도 2는 본 발명에 따른 플로우 단위의 트래픽 데이터를 설명하기 위한 도면이다.
- [0103] 도 3은 트래픽 데이터가 갖는 속성의 예를 설명하기 위한 도면이다.
- [0104] 도 4는 본 발명에 적용되는 SVM을 구하는 과정을 설명하기 위한 도면이다.
- [0105] 도 5는 본 발명의 일 실시예에 따른 트래픽 데이터의 분류 방법을 설명하기 위한 도면이다.
- [0106] 도 6은 본 발명의 일 실시예에 따른 트래픽 데이터 분류 방법을 도시한 순서도이다.
- [0107] <도면의 주요 부분에 대한 설명>
- [0108] 100: 트래픽 분류 장치 110: 데이터 수집부
- [0109] 120: 속성 선택부 130: 학습 데이터 생성부
- [0110] 140: 트래픽 분류부

도면

도면1



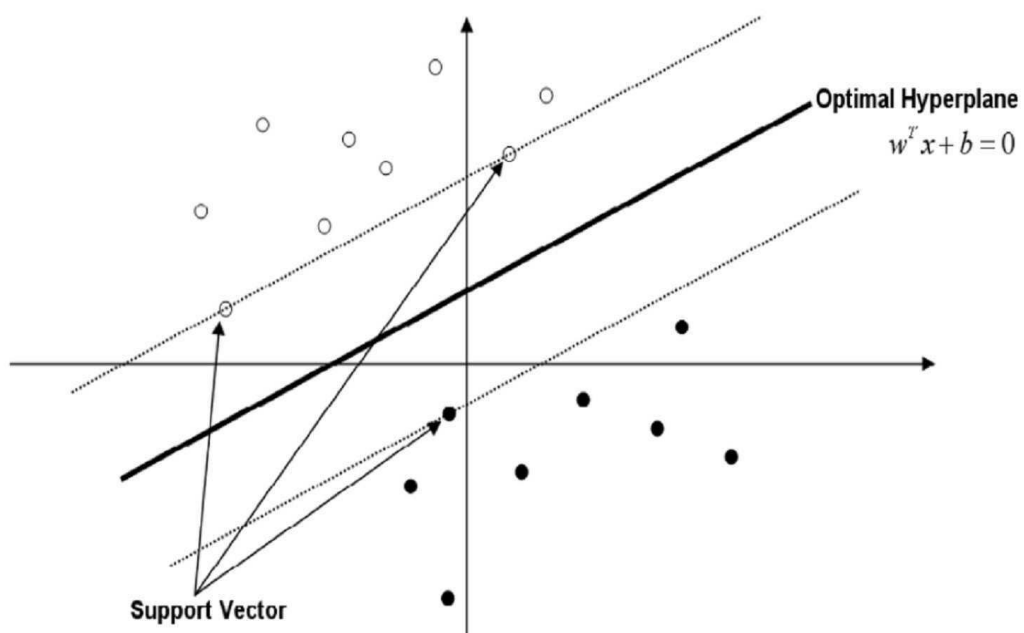
도면2



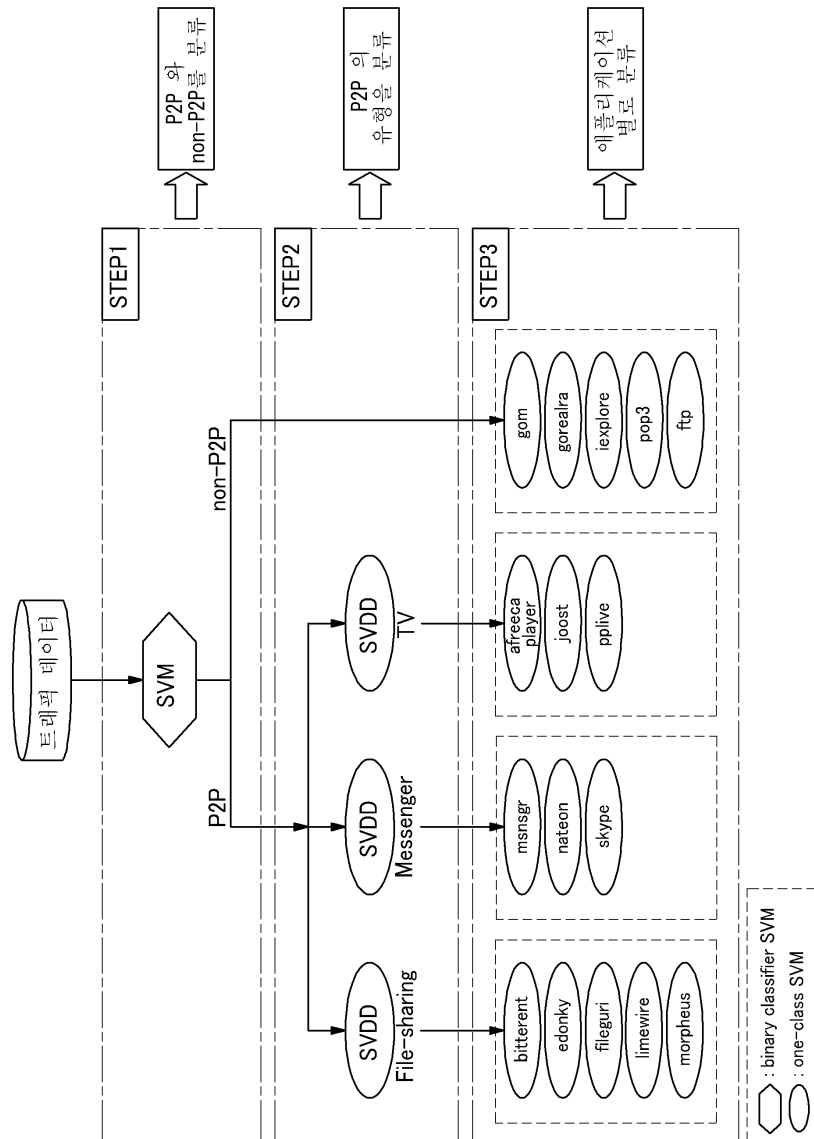
도면3

속성			설명
protocol			IP protocol (TCP, UDP,...)
입력 플로우	dPkts dOctets duration(last - first) syn ack rst fin		unmber of packets unmber of bytes flow duration unmber of SYN packets unmber of ACK packets unmber of RST packets unmber of FIN packets
출력 플로우	"		"
입력 플로우	statistics of packet size	min, max, avg, sdev	minimum, maximum, Value average, standard deviation
	statistics of window size	"	"
	statistics of jitter size	"	"
출력 플로우	statistics of packet size	"	"
	statistics of window size	"	"
	statistics of jitter size	"	"

도면4



도면5



도면6

