머신러닝을 사용한 프로토콜 구문 분석 기법 제안

https://youtu.be/kp-6OaNz6xo

한성대학교 HANSUNG UNIVERSITY

CryptoCraft LAB
https://crypto.modoo.at

프로토콜 리버스 엔지니어링

- Protocol Reverse Engineering(PRE)의 주요 과제 중 하나는 주로 수동으로 수행되는 점
  - 수동 PRE는 매우 지루하고 시간이 많이 걸림
  - 특정 프로토콜 사양이 완전히 밝혀지는 데 몇 년이 걸릴 수 있음
  - 사람이 읽을 수 없는 바이너리 프로토콜과 같은 수많은 장애물로 인해 효율적인 PRE 힘듬

- APRE (Automatic Protocol Reverse Engineering)
  - 설명에 의존하지 않고 자동으로 네트워크 프로토콜의 구조를 추출하는 방법

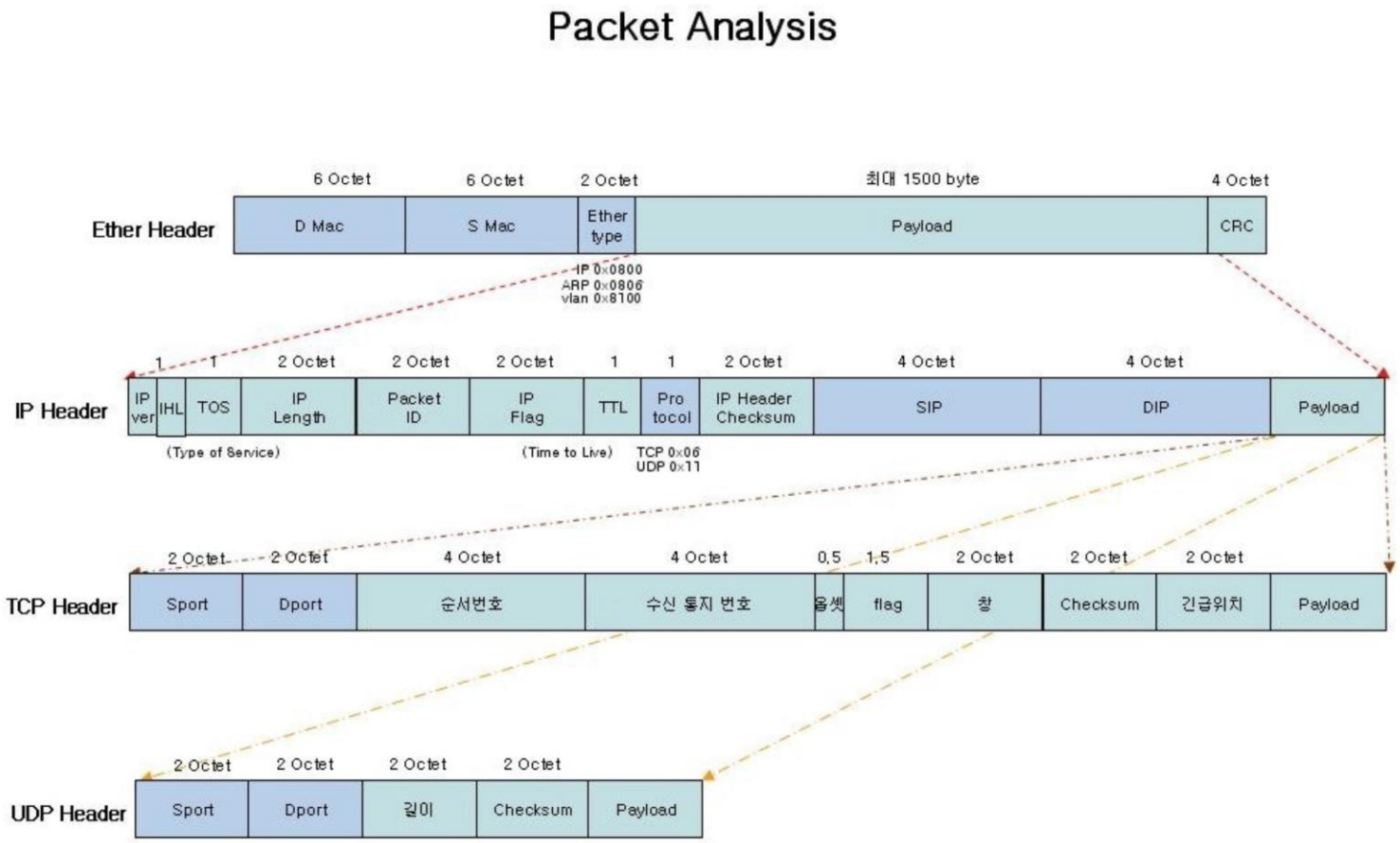
프로토콜 분석

- 구문 추론
  - 프로토콜 필드 경계, 오프셋 위치 및 엔디안을 추론
    두 호스트 간의 통신에 관련된 메시지를 형식화하는 데 사용 된 프로토콜 규칙을 식별

- 시맨틱 추론
  - 구문 추론 후 분명히 다음 단계
    두 통신기간에 교환 된 데이터 내용을 의미와 함께 추론하는 과정입니다.
  - EX)HTTP (Hypertext Transfer Protocol)에서 유추 된 의미 정보는 웹 페이지 콘텐츠

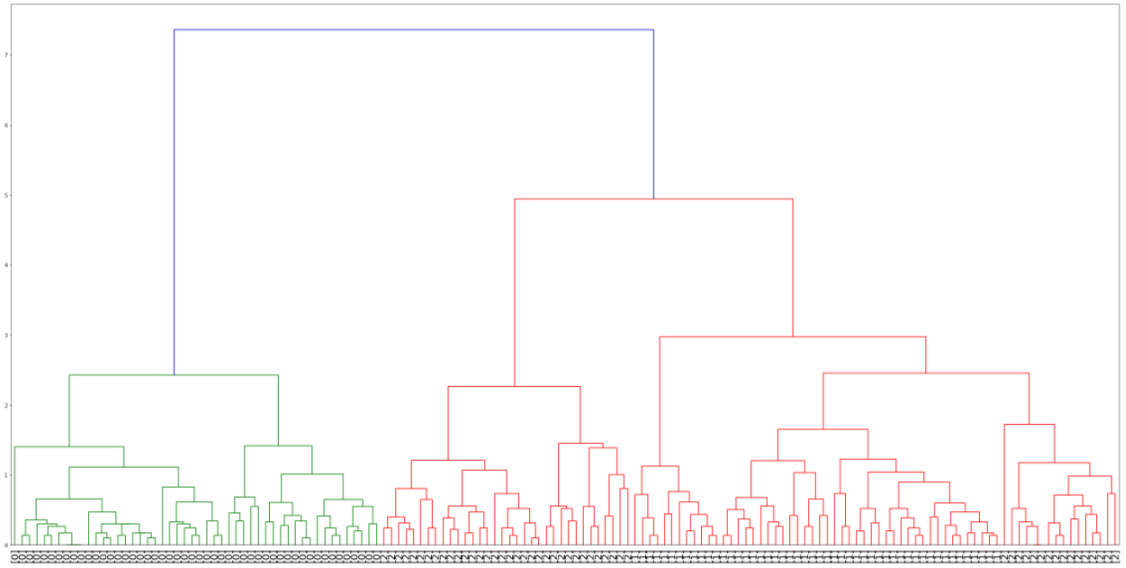
머신러닝 기반 기법 제안

- 계층적 클러스터링
- SVM
- 순차패턴



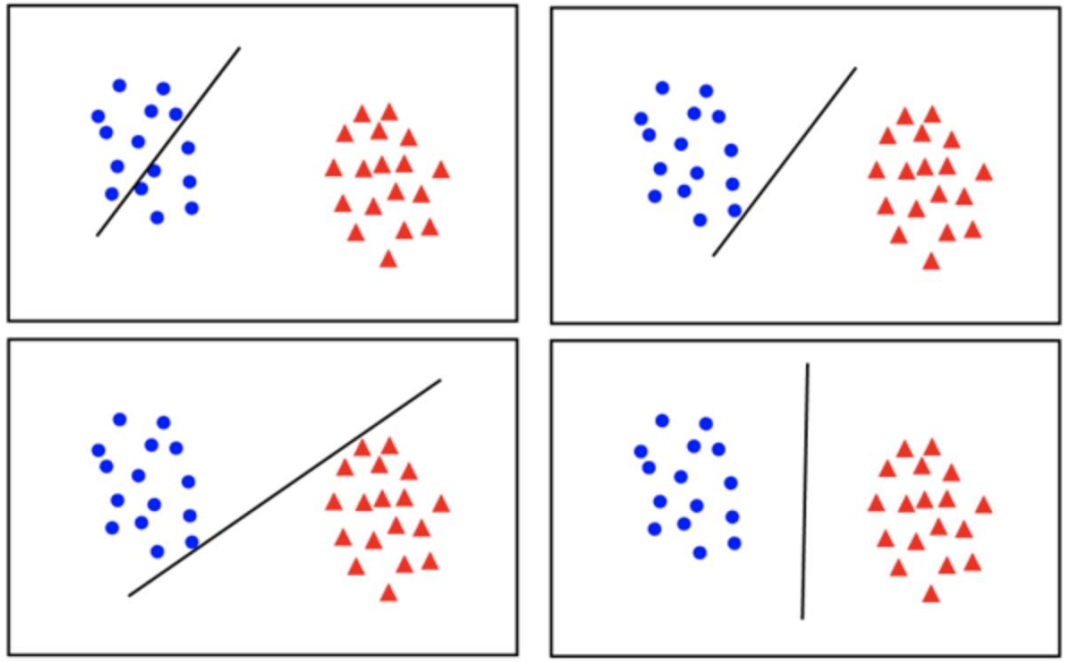
계층적 군집 분석(Hierarchical clustering)

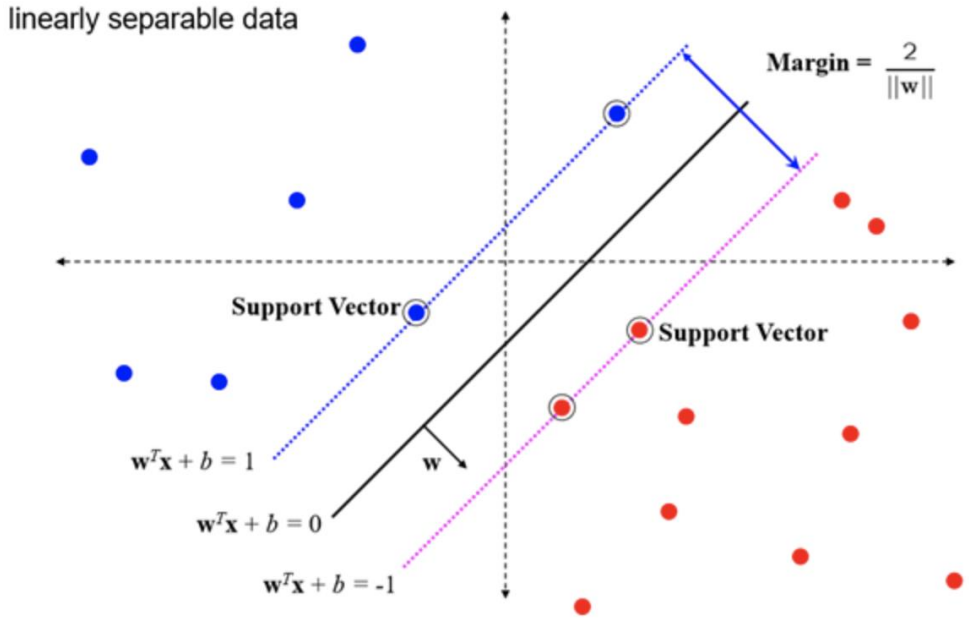
- 비슷한 군집끼리 묶어 가면서 최종적으로는 하나의 케이스가 될 때까지 군집을 묶는 클러스터링 알고리즘
- 군집간의 거리를 기반으로 클러스터링을 하는 알고리즘이며, K-Means와는 다르게 군집의 수를 미리 정해주지 않아도 됨



svm

- 분류나 회귀 분석에 사용하는 머신러닝
- 지도학습 알고리즘





순차적 패턴 분석

- 순차 패턴 마이닝(sequential pattern mining)은 대량의 데이터에 숨겨진 “순차적 패턴”을 찾는 분석방법
- 연속하여 일어나는 패턴을 찾는데 유용한 방법으로,
커머스 분야에서 고객이 어떤 순서로 제품을 구매하는지 분석 활용

고객 ID	시퀀스
1	({맥주, 주스}, {기저귀, 맥주}, {우유, 과자})
2	({맥주, 땅콩}, {기저귀}, {우유, 과자}, {사과})
3	({기저귀, 맥주}, {우유}, {맥주, 주스}, {과자})
4	({우유, 주스}, {맥주, 땅콩}, {맥주, 기저귀})

지지도	마이닝 결과
50 %	(맥주, 기저귀, 우유)
75 %	(기저귀, 우유)

제안기법

- 미지의 프로토콜? 감독 학습 알고리즘?

1. 계층적 클러스터링
2. Svm
3. 순차패턴

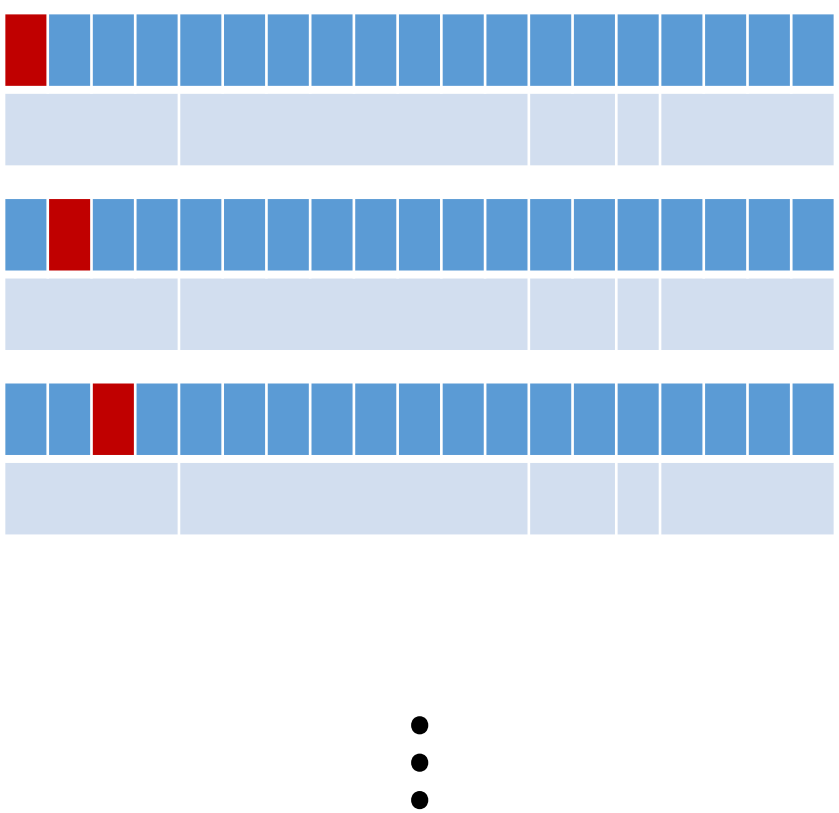
- 과정

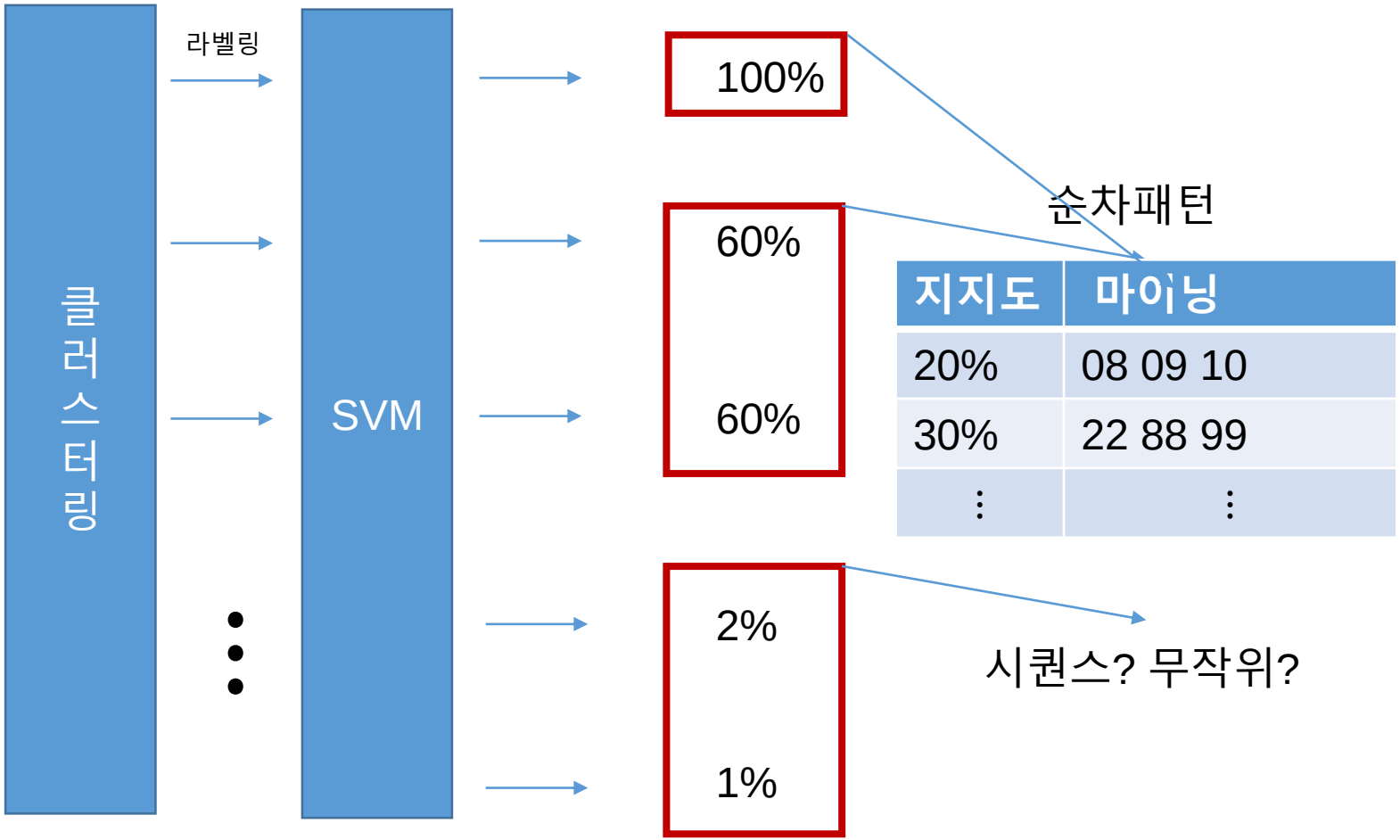
1. 한 바이트씩 클러스터링으로 분류 후 라벨링
2. svm 학습
3. 정확도 확인
   - 정확도 높음 -> 클러스터링 잘됨 -> 규칙성 O
   - 정확도 낮음 -> 클러스터링 안됨 -> 규칙성 X

제안기법

정확도 기반 판별

1. 높은 정확도 낮은 정확도 - > 높은 값 : 고정값 , 낮은값 : 가변적
2. 같은 값의 정확도. - > 같은 구문에 해당





Q & A