캡스톤 18 조 김민송 멘토링 활용 보고서

프로젝트명	O24Sec	팀명	멜러리를 찾아서

팀 멘토링

* 모든 멘토링을 상세하게 적으면 분량이 너무 길어져, 간단히 미팅 주제와 개인별로 느낀점을 적었습니다.

⊙ [01/22] 팀 미팅 - 윤명근 교수님
- 연구에 사용할 데이터 분석 및 도식화에 대한 피드백
=> 일단 목표에 대한 데이터 분석에 대한 필요성이 있었고 데이터가 어떤 필드로 구성되어 있는지 확인을 먼저 해보아야 한다.

⊙ [01/26] 팀 미팅 - 윤명근 교수님
- 1 차 주제 확립: 보안관제 이벤트 처리에서, 비지도학습 기반으로 기존기술의 한계와 알람피로, 데이터 오분류 등을 해결하고, 군집화를 통한 여러 머신러닝등의 데이터셋으로 활용할 수 있는 데이터 군집화 기술 개발
- 주제에 관한 토론 및 피드백
=> 군집화에 대한 스터디 필요, 사용할 데이터 범위 정의, 오분류 된 데이터들의 통계를 통해 어떤 데이터로 했을 때 가장 효율적으로 실험할 수 있을지 통계내어 측정
=> 추가 목표사항으로 오분류가 생긴 데이터에 대해 왜 오분류가 나게 되었는지 조사

⊙ [01/28] 윈스 미팅 - 최병환 팀장님, 윤명근 교수님, 박하명 교수님
- 주식회사 윈스에 방문하여, 기존 연구에 대한 피드백과 제안 기술에 대한 토론 진행.
- 데이터 셋에 대한 요구사항 및 개선사항 접수, 윈스와의 제안 기술 관련 협의
=> 산학협력지정주제에 대한 목표 설정을 했고, 목표에 대해 한계가 생기는 부분이 있을 수 있기 때문에 어느정도 감안하면서 최대한 목표에 비슷하게 가기로 정함, 데이터셋에서도 써야한 특징과 일단 사용하지 않고 진행해볼 특징을 분리하여 진행

⊙ [02/02] 팀 미팅 - 윤명근 교수님
- 목표사항인 비지도학습 군집화를 위한 단계별 계획 수립
- 단계의 목적 및 시나리오 설립, 알고리즘에 관한 토의
- 1 차 목표사항: 보안관제 이벤트의 객체 분류(IPS 장치 기준 내외부 구별, 서버와 클라이언트 구별)
- 2 차 목표사항: 보안관제 이벤트의 페이로드 암호화 판별(암호화하지 않은 데이터셋을 구성)
- 3 차 목표사항: 기존의 주제인 데이터 군집화 기술 개발
=> 목표에 대한 바로 접근이 아닌 단계별 접근을 통해 중간 중간 모듈을 산출해 내기로 결정
먼저 내외부를 구별하기 위해 내외부에 대해 이미 알고있는 특징인 direction 을 가지고 라벨을 먼저 생산해서 나중에 만든 결과물로 정확도를 계산해 볼 수 있게 함
=> 암호화 식별에 대해서는 페이로드를 디코딩 했을 때 읽을 수 있는 문자가 있는지 영어 단어를 체크하는 방식으로 코딩을 해서 분리해보도록 함

⊙ [02/17] Ceeya 미팅 - 문성익 멘토님
- 문제 접근 방식, 문제 해결 방식, 프로젝트 진행 방식에 대한 피드백
- 멘토님과의 통신 체계 마련
- 1 차 목표사항과 관련하여 여러 아이디어 토의:
멘토님의 네트워크 지식을 활용하여 IPS 장비 기준 내외부와 서버 클라이언트를 구별할 수 있는 아이디어를 같이 토론
- 3 차 목표사항과 관련하여 여러 아이디어 토의:

	멘토님의 클러스터링 기술 사용 경험을 통한 조언. 클러스터링 기술 별 성능의 차이, 데이터 전처리와 피쳐 선정, 가중치관련 조언 네트워크 데이터의 특수성을 활용한 클러스터 방안 조언. 페이로드의 특성을 활용하여 클러스터링 => 일단 클러스터링에 대해서는 어떤 모델을 사용할지 고민하고 있었는데 모델 간의 큰 차이가 없다 하셨기 때문에 그에 대해서는 고민을 접고 한가지 모델 예를 들어 KNN 으로 진행을 먼저 해보고 결과를 만들어 낸 다음에 마지막으로 모델을 다른 걸로 해보기로 하였다. => 암호화 패킷 분리에 대해서는 생각하고 있던 내용인 지정된 크기별로 청킹해서 Entropy 를 측정해가지고 대표 값을 나타내는 것이 어떻겠냐 하는 의견에 멘토님도 그게 좋을 것 같다 하셔서 그대로 진행해 보기로 하였다. => 클러스터링은 일단 여러 시도를 통해 결과값이 좋은 걸로 하는게 좋을 것 같다 하셔서 특징도 바꿔보고 전처리를 다르게 해보면서 일단 시도를 계속 해보면서 결과 기반으로 해보는게 좋을 것 같다. ◎ [02/26] 팀 미팅 - 윤명근 교수님 - 1 차, 2 차 목표사항의 현재 진행 상황에 대한 PT 및 그에 대한 피드백: 1 차 목표사항에 있어 간결한 룰을 유지해야 할 필요성이 있으며, 강건한 규칙을 순차적으로 적용해야 한다는 피드백 2 차 목표사항에 있어 엔트로피를 구하는 방식과 데이터 형 변환 관련 피드백 => 엔트로피 관련해서 페이로드의 길이가 짧아지면 최대 엔트로피의 크기도 급하게 작아지는 특징을 보였기 때문에 이는 8 비트로 표현되어 있는 바이트 페이로드의 한계이기 때문에 4 비트로 엔트로피를 측정해보자고 조언해 주셨다 일단 8bit 엔트로피 측정에서 4bit 엔트로피 측정으로 방법을 바꿔보고 결과를 만들어 봐야할 것 같다. ◎ [03/12] 윈스담당 팀 미팅 - 윤명근 교수님 - 윈스 프로젝트와 관련한 사람 들과의 팀미팅 진행: 현재까지 진행된 1 차, 2 차 목표사항의 구현에 대한 피드백 및 앞으로의 진행상황에 대한 계획 수립 1 차 목표사항에 있어 겪는 여러 예외사항에 대한 토론 => 현재 1 차 목표에서 서버/클라이언트는 어느정도 구분이 되지만 내부/외부에 대해서 예외사항이 너무 많아 일관적으로 만드는 룰에 대한 정의가 좀 부족했다 미팅을 통해 사설 아이피 같은 것은 확정 내부로 진행하도록 고정시키고 나머지에 대해서 다시 시각화 시켜서 관계를 다시 알아보기로 했다. 일단 시각화를 진행을 하고 클러스터링에 대해서는 실험에 시간이 꽤 많이 걸리는 작업이다 보니까 먼저 간단한 코드로 라도 만들어봐서 계속 조금씩 바꿔가면서 진행해 보는 것이 좋을 것 같다.
개인 멘토링	▸ [02/03] 1 대 1 미팅 - 윤명근 교수님 먼저 정답라벨을 만들기 위한 규칙으로 서버/클라이언트 구분은 1024 미만의 포트를 가진 것을 서버로 하여 반대편을 클라이언트로 구분 짓고 이렇게 안될 경우 양쪽에서 작은 크기의 포트를 가지는 것을 서버로 하여 분리하고 내부/외부의 경우 일단 정답데이터를 만드는 것이기 때문에 IPS 장비의 로그를 보고 Direction 타입을 봐서 내부 객체, 외부객체로 정답데이터를 선행적으로 만들기로 하였다. 이것으로 (내부/외부 – 서버/클라이언트) 로 구분되는 4 가지 객체를 정의할 수 있었고 이제 direction 타입을 보지 않고 객체를 구분하는 알고리즘을 한번 생각해보자고 하셔서 IP 의 빈도부터 시작해서 새로운 알고리즘을 찾기 시작했다. ▸ [02/09] 1 대 1 미팅 - 윤명근 교수님

암호화된 데이터를 식별하기 위해서 일단 entropy 를 통해 식별을 해보라고 하셨다. Entropy 를 통해서 먼저 적용을 해보았을 때 내가 본 페이로드에서는 암호화된 것 같은데 엔트로피가 낮은 것 도 있었고 암호화 되지 않은 것 같은데 엔트로피가 높은 것도 있었다. 이 때문에 엔트로피를 사용하지 않고 페이로드를 디코딩 해 보았을 때 파이썬 라이브러리에 저장되어 있는 영어 사전과 비교해서 영어 사전에 등록된 단어가 3 개 이상 나오게 된다면 암호화가 안된 것으로 해보려고 했다 하지만 이는 시간이 너무 오래 걸리게 되었고 이에 대해 이야기해보고자 교수님과 미팅을 가지게 되었다. 얻은 답변으로는 일단 암호화가 안되었더라도 압축된 파일이나 이미지, 동영상 등 인코딩 되어서 보내지는 파일의 경우에는 엔트로피가 높을 수 밖에 없다는 이야기 였다. 그렇기 때문에 먼저 포트 기반으로 한번 해보자 80(HTTP)의 경우에는 평문위주이고 443(SSL)의 경우에는 암호문 위주이니 이를 토대로 엔트로피로 한번 조사해보라고 하셨다.

- [02/12] 1 대 1 미팅 - 윤명근 교수님
  이전 미팅에 대한 조사 결과 인코딩 되어 엔트로피가 높은 동영상 이미지 파일의 경우에는 MIME 타입으로 전송된다는 특징이 있었다 이 경우에는 MIME 헤더부분은 엔트로피가 낮기 때문에 이 부분에 대해서 엔트로피를 측정해서 대표 값으로 사용하는 것이 어떻겠냐고 제안을 했다. 즉 전체 엔트로피에서 적절한 크기로 청킹을 하여 각 청크에서 엔트로피를 측정해 그 중 가장 최소값을 대표 값으로 하여 대표 값이 임계치보다 낮을 경우 암호화되지 않은 패킷으로 처리하는 방법에 대해 말씀드리고 괜찮은 것 같다 하시고 해당 방법에 대한 결과 그래프를 만들어 보자고 하셨다.

- [02/19] 1 대 1 미팅 - 윤명근 교수님
  Ceeya 에서 배정해주신 멘토 분과의 회의 결과와 만들어 놓은 청킹된 데이터에 대해 그래프를 만들어 다시 이야기해보았다. 페이로드의 길이가 긴 경우 확실한 탐지가 가능했지만 페이로드의 길이가 짧아질 경우 예를 들어 청킹 사이즈가 256 이라 할 때 256 보다 작은 데이터에 대해서 정확한 판단이 되지 않았다. 256 크기에 맞춰서 랜덤 패딩을 해줄 경우 정상 파일이지만 길이가 짧아 랜덤 패딩이 주를 이루게 되어 엔트로피가 높아지는 경우가 있었고 이에 대해 해결하기 위해 다시 방법을 찾아보기로 하였다.

- [03/02] 1 대 1 미팅 - 윤명근 교수님
  타 논문을 확인해보다가 엔트로피 측정을 해서 어플리케이션 데이터를 확인할 때 어플리케이션 데이터의 헤더의 길이 별 엔트로피를 정해 놓고 해당 엔트로피 보다 높아질 경우 암호화 된 것이라 판단한다는 것에 영감을 받아 페이로드길이가 0 일 때부터 1600 까지의 엔트로피를 각각 모두 구해 놓고 해당 엔트로피 값의 평균, 최소값, 최대값을 잡아 이를 이용해서 암호화 패킷을 구분하는 방법에 대해 말씀을 드렸다. 내가 읽은 논문에 대해서 한번 읽어 보시고 접근 방법이 나쁘지 않은 것 같다 하시면서 일단 만들어 보고 해당 방법으로 안 걸리는 패킷을 찾아보고 또한 평균 최소값 최대값을 어떻게 사용해서 수식을 만들지 정해보자고 하셨다. 이후 여러가지 수식을 통해 적절한 기준을 찾아보기로 생각했다.

- [03/08] 1 대 1 미팅 - 윤명근 교수님
  데이터 중 일부를 샘플로 뽑아 그 데이터들을 기준으로 수식을 여러 개 넣어보면서 기준을 잡았다 또한 짧은 페이로드의 경우 TLS 암호화 통신에서 헤더 때문에 엔트로피가 낮게 나오는 경우가 있었는데 이걸 임의로 바이트 매칭을 시켜서 빼도 될지 의논해 보았다. Alert 메시지가 그 경우였는데 alert 자체가 TLS 에서 경고 메시지가 나올 경우 이기 때문인데 이를 암호문으로 처리해도 될지 생각해 보았다 교수님 의견으로는 alert 데이터들에 대해 기존에 어떤 이벤트로 IPS 가 감지했는지 확인해 보자 해서 확인을 했더니 페이로드의 이유가 아닌 IP 로 인한 이벤트들이 전부였기 때문에 alert 에 대해서 byte 매칭을 시켜 암호문으로 포함하기로 하였다.