(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2020년01월21일
(11) 등록번호 10-2068507
(24) 등록일자 2020년01월15일

(51) 국제특허분류(Int. Cl.)
G06N 20/00 (2019.01)
(52) CPC특허분류
G06N 20/00 (2019.01)
(21) 출원번호 10-2019-0084107
(22) 출원일자 2019년07월11일
심사청구일자 2019년07월11일
(56) 선행기술조사문헌
JP2002157262 A
(뒷면에 계속)

(73) 특허권자
(주)시큐레이어
서울시 성동구 성수이로10길 14 , 1202호(성수동2가, 에이스하이엔드성수타워)
(72) 발명자
강필상
대전광역시 동구 홍도로46번길 20, 102동 202호 (용전동, 새피앙아파트)
신강식
대전광역시 서구 만년남로3번길 86-10, 103호 (만년동, 목화빌라)
(74) 대리인
특허법인 수

전체 청구항 수 : 총 28 항 심사관 : 서광훈

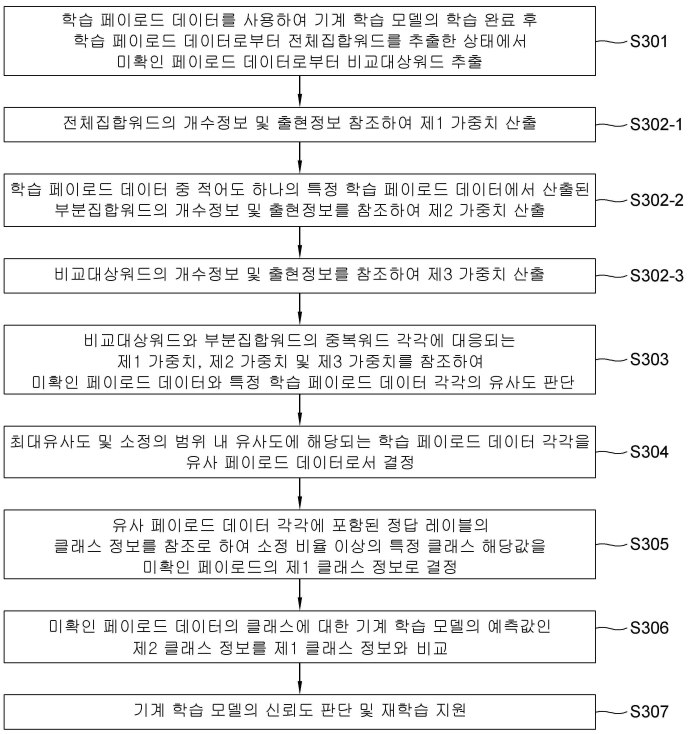
(54) 발명의 명칭 **기계 학습 모델의 신뢰도를 판단하기 위한 후처리 방법 및 이를 사용한 후처리 장치**

(57) 요 약

본 발명에 따르면, 기계 학습 모델의 신뢰도를 판단하기 위한 방법으로서, (a) 복수의 학습 페이로드 데이터 - 상기 학습 페이로드 데이터 각각은, 해당되는 소정의 클래스에 대한 정보인 정답 레이블이 부여됨 - 를 사용하여 상기 기계 학습 모델의 학습이 완료된 후, 미확인 페이로드 데이터가 획득되면, 후처리 장치가, 상기 학습 페이

(뒷면에 계속)

대 표 도 - 도3



로드 데이터 중 적어도 일부로부터 전체집합워드 - 상기 전체집합워드 각각은 상기 학습 페이로드 데이터에 포함된 숫자, 특수문자 및 문자열 중 하나 또는 둘 이상의 조합임 - 를 추출한 상태에서, 상기 미확인 페이로드 데이터로부터 비교대상워드 - 상기 비교대상워드 각각은 상기 미확인 페이로드 데이터에 포함된 숫자, 특수문자 및 문자열 중 하나 또는 둘 이상의 조합임 - 를 추출하는 단계; (b) 상기 후처리 장치가, (i) 상기 학습 페이로드 데이터에서 추출된 상기 전체집합워드의 개수에 대한 정보 및 상기 학습 페이로드 데이터에서 상기 전체집합워드 각각이 출현하는 횟수에 대한 정보를 참조로 하여, 상기 전체집합워드 각각에 대응되는 제1 가중치를 산출하는 프로세스, (ii) 상기 학습 페이로드 데이터 중 적어도 하나의 특정 학습 페이로드 데이터에서 추출된 부분집합워드의 개수에 대한 정보 및 상기 특정 학습 페이로드 데이터에서 상기 부분집합워드 각각이 출현하는 횟수에 대한 정보를 참조로 하여, 상기 부분집합워드 각각에 대응되는 제2 가중치를 산출하는 프로세스, 및 (iii) 상기 미확인 페이로드 데이터에서 추출된 상기 비교대상워드의 개수에 대한 정보 및 상기 미확인 페이로드 데이터에서 상기 비교대상워드 각각이 출현하는 출현 횟수에 대한 정보를 참조로 하여, 상기 비교대상워드 각각에 대응되는 제3 가중치를 산출하는 프로세스를 수행하는 단계; (c) 상기 후처리 장치가, 상기 미확인 페이로드 데이터에서 추출된 상기 비교대상워드와 상기 특정 학습 페이로드 데이터 각각에서 추출된 각각의 상기 부분집합워드를 비교하여, 중복되는 중복워드 각각에 대응되는 제1 가중치, 제2 가중치 및 제3 가중치를 참조로 하여 상기 미확인 페이로드 데이터와 상기 특정 학습 페이로드 데이터 각각의 유사도를 판단하는 단계; (d) 상기 후처리 장치가, (i) 상기 학습 페이로드 데이터 각각에 대응되는 상기 유사도 중 가장 큰 값을 가지는 최대유사도 및 이를 기준으로 하여 소정의 범위 이내에 포함되는 유사도에 해당되는 학습 페이로드 데이터 각각을 유사 페이로드 데이터로서 결정하는 프로세스, 및 (ii) 상기 유사 페이로드 각각에 부여된 정답 레이블 각각의 클래스 정보를 참조로 하여, 소정의 비율 이상의 특정 클래스에 해당되는 값을 상기 미확인 페이로드 데이터의 제1 클래스 정보로 결정하는 프로세스를 수행하는 단계; 및 (e) 상기 후처리 장치가, 상기 기계 학습 모델에 의하여 도출된 상기 미확인 페이로드 데이터의 클래스에 대한 모델 예측값이 제2 클래스 정보로서 획득된 상태에서, 상기 제1 클래스 정보와 상기 제2 클래스 정보를 비교하여 상기 기계 학습 모델의 신뢰도를 판단하거나 판단할 수 있도록 지원하는 단계; 를 포함하는 방법이 제공된다.

(56) 선행기술조사문헌
JP2005302041 A
KR101681109 B1
US20090300765 A1
EP02182458 A1

이 발명을 지원한 국가연구개발사업
과제고유번호 2017-0-00200
부처명 과학기술정보통신부
연구관리전문기관 (IITP)정보통신기획평가원
연구사업명 정보보호핵심원천기술개발(R&D)
연구과제명 (자가방어-3세부) 진화형 사이버방어 가시화 기술 개발
기 여 율 1/1
주관기관 (주)시큐레이어
연구기간 2017.04.01 ~ 2020.12.31

명 세 서

청구범위

청구항 1

기계 학습 모델의 신뢰도를 판단하기 위한 방법으로서,

(a) 복수의 학습 페이로드 데이터 - 상기 학습 페이로드 데이터 각각은, 해당되는 소정의 클래스에 대한 정보인 정답 레이블이 부여됨 - 를 사용하여 상기 기계 학습 모델의 학습이 완료된 후, 미확인 페이로드 데이터가 획득되면, 후처리 장치가, 상기 학습 페이로드 데이터 중 적어도 일부로부터 전체집합워드 - 상기 전체집합워드 각각은 상기 학습 페이로드 데이터에 포함된 숫자, 특수문자 및 문자열 중 하나 또는 둘 이상의 조합임 - 를 추출한 상태에서, 상기 미확인 페이로드 데이터로부터 비교대상워드 - 상기 비교대상워드 각각은 상기 미확인 페이로드 데이터에 포함된 숫자, 특수문자 및 문자열 중 하나 또는 둘 이상의 조합임 - 를 추출하는 단계;

(b) 상기 후처리 장치가, (i) 상기 학습 페이로드 데이터에서 추출된 상기 전체집합워드의 개수에 대한 정보 및 상기 학습 페이로드 데이터에서 상기 전체집합워드 각각이 출현하는 횟수에 대한 정보를 참조로 하여, 상기 전체집합워드 각각에 대응되는 제1 가중치를 산출하는 프로세스, (ii) 상기 학습 페이로드 데이터 중 적어도 하나의 특정 학습 페이로드 데이터에서 추출된 부분집합워드의 개수에 대한 정보 및 상기 특정 학습 페이로드 데이터에서 상기 부분집합워드 각각이 출현하는 횟수에 대한 정보를 참조로 하여, 상기 부분집합워드 각각에 대응되는 제2 가중치를 산출하는 프로세스, 및 (iii) 상기 미확인 페이로드 데이터에서 추출된 상기 비교대상워드의 개수에 대한 정보 및 상기 미확인 페이로드 데이터에서 상기 비교대상워드 각각이 출현하는 출현 횟수에 대한 정보를 참조로 하여, 상기 비교대상워드 각각에 대응되는 제3 가중치를 산출하는 프로세스를 수행하는 단계;

(c) 상기 후처리 장치가, 상기 미확인 페이로드 데이터에서 추출된 상기 비교대상워드와 상기 특정 학습 페이로드 데이터 각각에서 추출된 각각의 상기 부분집합워드를 비교하여, 중복되는 중복워드 각각에 대응되는 제1 가중치, 제2 가중치 및 제3 가중치를 참조로 하여 상기 미확인 페이로드 데이터와 상기 특정 학습 페이로드 데이터 각각의 유사도를 판단하는 단계;

(d) 상기 후처리 장치가, (i) 상기 학습 페이로드 데이터 각각에 대응되는 상기 유사도 중 가장 큰 값을 가지는 최대유사도 및 이를 기준으로 하여 소정의 범위 이내에 포함되는 유사도에 해당되는 학습 페이로드 데이터 각각을 유사 페이로드 데이터로서 결정하는 프로세스, 및 (ii) 상기 유사 페이로드 각각에 부여된 정답 레이블 각각의 클래스 정보를 참조로 하여, 소정의 비율 이상의 특정 클래스에 해당되는 값을 상기 미확인 페이로드 데이터의 제1 클래스 정보로 결정하는 프로세스를 수행하는 단계; 및

(e) 상기 후처리 장치가, 상기 기계 학습 모델에 의하여 도출된 상기 미확인 페이로드 데이터의 클래스에 대한 모델 예측값이 제2 클래스 정보로서 획득된 상태에서, 상기 제1 클래스 정보와 상기 제2 클래스 정보를 비교하여 상기 기계 학습 모델의 신뢰도를 판단하거나 판단할 수 있도록 지원하는 단계;

를 포함하는, 방법.

청구항 2

제1항에 있어서,

상기 (c) 단계는,

(c1) 상기 후처리 장치가, (i) 상기 중복워드 각각에 대응되는 제1 가중치로 상기 중복워드 각각에 대응되는 제2 가중치를 나눈 값인 학습데이터가중치를 산출하는 프로세스, (ii) 상기 중복워드 각각에 대응되는 제1 가중치로 상기 중복워드 각각에 대응되는 제3 가중치를 나눈 값인 미확인데이터가중치를 산출하는 프로세스, 및 (iii) 상기 학습데이터가중치와 상기 미확인데이터가중치를 참조로 하여 소정의 제1 연산을 수행하고, 그 결과로서 상기 중복워드 각각의 최종가중치를 산출하는 프로세스를 수행하는 단계; 및

(c2) 상기 후처리 장치가, 상기 중복워드 각각에 대응되는 상기 최종가중치 각각에 대하여 소정의 제2 연산을 수행한 결과값을 상기 유사도로서 획득하는 단계;

를 포함하는, 방법.

청구항 3

제1항에 있어서,

상기 (e) 단계에서,

상기 신뢰도의 판단은, 상기 후처리 장치가, 상기 제1 클래스 정보와 상기 제2 클래스 정보를 비교하여, (i) 상기 제1 클래스 정보와 상기 제2 클래스 정보가 일치하는 경우 상기 모델 예측값을 상기 미확인 페이로드 데이터에 대응되는 클래스 값으로 판단하는 프로세스, 및 (ii) (1) 상기 제1 클래스 정보와 상기 제2 클래스 정보가 일치하지 않거나 (2) 상기 제1 클래스 정보가 도출되지 않는 경우에는 상기 모델 예측값을 별도의 검사 대상으로 분류하는 프로세스 중 적어도 하나를 수행함으로써 이루어지는 것을 특징으로 하는, 방법.

청구항 4

제3항에 있어서,

상기 모델 예측값이 별도의 검사 대상으로 분류되는 경우, 상기 후처리 장치가, (i) 상기 제1 클래스 정보 및 상기 제2 클래스 정보를 제공함으로써 상기 기계 학습 모델의 신뢰도의 판단이 가능하도록 지원하는 프로세스, 및 (ii) 상기 후처리 장치에 연결된 별도의 사용자 단말로 하여금 상기 제1 클래스 정보 및 상기 제2 클래스 정보를 제공하도록 함으로써 상기 기계 학습 모델의 신뢰도의 판단이 가능하도록 지원하는 프로세스 중 적어도 하나를 수행하는 것을 특징으로 하는, 방법.

청구항 5

제1항에 있어서,

상기 (a) 단계에서,

상기 전체집합워드를 추출하면, 상기 후처리 장치가, 상기 전체집합워드에 대한 정보를 참조로 하여 상기 전체집합워드의 데이터를 포함하는 제1 딕셔너리를 생성하는 프로세스를 추가로 수행하고,

상기 (b) 단계에서,

상기 부분집합워드를 추출하면, 상기 후처리 장치가, 상기 부분집합워드에 대한 정보를 참조로 하여 상기 특정 학습 페이로드 데이터 각각에 대응되는 부분집합워드의 데이터를 포함하는 제2 딕셔너리 각각을 생성하는 프로세스를 추가로 수행하여,

상기 후처리 장치가, 상기 제1 딕셔너리 및 상기 제2 딕셔너리를 참조로 하여 상기 제1 가중치 및 상기 제2 가중치를 산출하는 것을 특징으로 하는, 방법.

청구항 6

제1항에 있어서,

상기 (b) 단계에서,

상기 후처리 장치가, 사전에 결정되어 있는 복수의 사전공격워드에 대한 정보를 추가로 참조하여, 상기 미확인 페이로드 데이터에서 추출된 상기 비교대상워드 중 상기 사전공격워드에 해당되는 비교대상워드에 대해서만 상기 제3 가중치를 산출하는 것을 특징으로 하는, 방법.

청구항 7

제1항에 있어서,

상기 학습 페이로드 데이터 각각은 그 각각에 복수개의 정답 레이블이 부여되되, 그 중 일부의 정답 레이블은 상기 학습 페이로드 데이터 각각에 대한 오탐 또는 정탐에 대한 정답 레이블 및 상기 학습 페이로드 데이터 각각이 해당되는 공격 유형 정보에 대한 정답 레이블 중 적어도 일부를 포함하는 상태에서, 상기 학습 페이로드 데이터 각각이 상기 기계 학습 모델의 학습에 제공되는 것을 특징으로 하는, 방법.

청구항 8

제1항에 있어서,

상기 학습 페이로드 데이터는, 소정의 보안 위협 탐지 시스템에 의하여 탐지된 복수의 탐지 로그 데이터 각각에 대응되는 페이로드 데이터로서, 상기 학습 페이로드 데이터 각각에, 이에 해당되는 소정의 클래스에 대한 상기 정답 레이블이 부여되어 연동되도록 지원되는 것을 특징으로 하는, 방법.

청구항 9

제8항에 있어서,

상기 학습 페이로드 데이터 각각은 그 각각에 복수개의 정답 레이블이 부여되되, 그 중 일부의 정답 레이블은 상기 학습 페이로드 데이터 각각에 대한 오탐 또는 정탐에 대한 정답 레이블 및 상기 학습 페이로드 데이터 각각이 해당되는 공격 유형에 대응되는 정답 레이블 중 적어도 일부를 포함하는 상태에서, 상기 학습 페이로드 데이터 각각이 상기 기계 학습 모델의 학습에 제공되는 것을 특징으로 하는, 방법.

청구항 10

제9항에 있어서,

상기 학습 페이로드 데이터 각각은, 상기 학습 페이로드 데이터에 포함된 숫자, 특수문자 및 문자열 중 하나 또는 둘 이상의 조합인 워드 중 상기 학습 페이로드 데이터 각각이 해당되는 공격 유형에 대응되는 워드를 공격 키워드로서 별도로 분류한 상태에서 상기 기계 학습 모델의 학습에 사용되는 것을 특징으로 하는, 방법.

청구항 11

제1항에 있어서,

상기 미확인 페이로드 데이터는, 상기 학습 페이로드 데이터를 사용하여 상기 기계 학습 모델의 학습이 완료된 상태에서 신규로 소정의 보안 위협 탐지 시스템에 입력되어 탐지되는 특정 탐지 로그 데이터에 대응되는 페이로드 데이터인 것을 특징으로 하는, 방법.

청구항 12

제11항에 있어서,

상기 미확인 페이로드 데이터는, 별도의 정답 레이블이 부여되지 않은 페이로드 데이터인 상태로 상기 기계 학습 모델 및 상기 후처리 장치 각각에 제공됨으로써, 상기 미확인 페이로드 데이터에 대응되는 상기 제1 클래스 정보 및 상기 제2 클래스 정보가 각각 획득되는 것을 특징으로 하는, 방법.

청구항 13

제12항에 있어서,

상기 (b) 단계에서,

상기 후처리 장치가, 사전에 결정되어 있는 복수의 사전공격워드에 대한 정보를 추가로 참조하여, 상기 미확인 페이로드 데이터에서 추출된 상기 비교대상워드 중 상기 사전공격워드에 해당되는 비교대상워드에 대해서만 상기 제3 가중치를 산출하는 것을 특징으로 하는, 방법.

청구항 14

제1항에 있어서,

상기 (e) 단계 이후에,

(f) 상기 제1 클래스 정보 및 상기 제2 클래스 정보가 일치하지 않거나, 상기 제1 클래스 정보가 도출되지 않는 경우, 상기 후처리 장치가, 상기 기계 학습 모델의 재학습이 가능하도록 지원하는 단계를 추가로 포함하는, 방법.

청구항 15

기계 학습 모델의 신뢰도를 판단하기 위한 후처리 장치로서,

인스트럭션들을 저장하는 적어도 하나의 메모리; 및

상기 인스트럭션들을 실행하기 위해 구성된 적어도 하나의 프로세서; 를 포함하고,

상기 프로세서가,

(I) 복수의 학습 페이로드 데이터 - 상기 학습 페이로드 데이터 각각은, 해당되는 소정의 클래스에 대한 정보인 정답 레이블이 부여됨 - 를 사용하여 상기 기계 학습 모델의 학습이 완료된 후, 미확인 페이로드 데이터가 획득되면, 상기 학습 페이로드 데이터 중 적어도 일부로부터 전체집합워드 - 상기 전체집합워드 각각은 상기 학습 페이로드 데이터에 포함된 숫자, 특수문자 및 문자열 중 하나 또는 둘 이상의 조합임 - 를 추출한 상태에서, 상기 미확인 페이로드 데이터로부터 비교대상워드 - 상기 비교대상워드 각각은 상기 미확인 페이로드 데이터에 포함된 숫자, 특수문자 및 문자열 중 하나 또는 둘 이상의 조합임 - 를 추출하는 프로세스; (II) (i) 상기 학습 페이로드 데이터에서 추출된 상기 전체집합워드의 개수에 대한 정보 및 상기 학습 페이로드 데이터에서 상기 전체집합워드 각각이 출현하는 횟수에 대한 정보를 참조로 하여, 상기 전체집합워드 각각에 대응되는 제1 가중치를 산출하는 서브프로세스, (ii) 상기 학습 페이로드 데이터 중 적어도 하나의 특정 학습 페이로드 데이터에서 추출된 부분집합워드의 개수에 대한 정보 및 상기 특정 학습 페이로드 데이터에서 상기 부분집합워드 각각이 출현하는 횟수에 대한 정보를 참조로 하여, 상기 부분집합워드 각각에 대응되는 제2 가중치를 산출하는 서브프로세스, 및 (iii) 상기 미확인 페이로드 데이터에서 추출된 상기 비교대상워드의 개수에 대한 정보 및 상기 미확인 페이로드 데이터에서 상기 비교대상워드 각각이 출현하는 출현 횟수에 대한 정보를 참조로 하여, 상기 비교대상워드 각각에 대응되는 제3 가중치를 산출하는 서브프로세스를 수행하는 프로세스; (III) 상기 미확인 페이로드 데이터에서 추출된 상기 비교대상워드와 상기 특정 학습 페이로드 데이터 각각에서 추출된 각각의 상기 부분집합워드를 비교하여, 중복되는 중복워드 각각에 대응되는 제1 가중치, 제2 가중치 및 제3 가중치를 참조로 하여 상기 미확인 페이로드 데이터와 상기 특정 학습 페이로드 데이터 각각의 유사도를 판단하는 프로세스; (IV) (i) 상기 학습 페이로드 데이터 각각에 대응되는 상기 유사도 중 가장 큰 값을 가지는 최대유사도 및 이를 기준으로 하여 소정의 범위 이내에 포함되는 유사도에 해당되는 학습 페이로드 데이터 각각을 유사 페이로드 데이터로서 결정하는 서브프로세스, 및 (ii) 상기 유사 페이로드 각각에 부여된 정답 레이블 각각의 클래스 정보를 참조로 하여, 소정의 비율 이상의 특정 클래스에 해당되는 값을 상기 미확인 페이로드 데이터의 제1 클래스 정보로 결정하는 서브프로세스를 수행하는 프로세스; 및 (V) 상기 기계 학습 모델에 의하여 도출된 상기 미확인 페이로드 데이터의 클래스에 대한 모델 예측값이 제2 클래스 정보로서 획득된 상태에서, 상기 제1 클래스 정보와 상기 제2 클래스 정보를 비교하여 상기 기계 학습 모델의 신뢰도를 판단하거나 판단할 수 있도록 지원하는 프로세스; 를 수행하는, 후처리 장치.

청구항 16

제15항에 있어서,

상기 (III) 프로세스는, 상기 프로세서가,

(III-1) (i) 상기 중복워드 각각에 대응되는 제1 가중치로 상기 중복워드 각각에 대응되는 제2 가중치를 나눈 값인 학습데이터가중치를 산출하는 서브프로세스, (ii) 상기 중복워드 각각에 대응되는 제1 가중치로 상기 중복워드 각각에 대응되는 제3 가중치를 나눈 값인 미확인데이터가중치를 산출하는 서브프로세스, 및 (iii) 상기 학습데이터가중치와 상기 미확인데이터가중치를 참조로 하여 소정의 제1 연산을 수행하고, 그 결과로서 상기 중복워드 각각의 최종가중치를 산출하는 서브프로세스를 수행하는 프로세스; 및 (III-2) 상기 중복워드 각각에 대응되는 상기 최종가중치 각각에 대하여 소정의 제2 연산을 수행한 결과값을 상기 유사도로서 획득하는 프로세스; 를 수행하는, 후처리 장치.

청구항 17

제15항에 있어서,

상기 (V) 프로세스에서,

상기 신뢰도의 판단은, 상기 프로세서가, 상기 제1 클래스 정보와 상기 제2 클래스 정보를 비교하여, (i) 상기

제1 클래스 정보와 상기 제2 클래스 정보가 일치하는 경우 상기 모델 예측값을 상기 미확인 페이로드 데이터에 대응되는 클래스 값으로 판단하는 프로세스, 및 (ii) (1) 상기 제1 클래스 정보와 상기 제2 클래스 정보가 일치하지 않거나 (2) 상기 제1 클래스 정보가 도출되지 않는 경우에는 상기 모델 예측값을 별도의 검사 대상으로 분류하는 프로세스 중 적어도 하나를 수행함으로써 이루어지는 것을 특징으로 하는, 후처리 장치.

청구항 18

제17항에 있어서,

상기 모델 예측값이 별도의 검사 대상으로 분류되는 경우, 상기 프로세서가, (i) 상기 제1 클래스 정보 및 상기 제2 클래스 정보를 제공함으로써 상기 기계 학습 모델의 신뢰도의 판단이 가능하도록 지원하는 프로세스, 및 (ii) 상기 후처리 장치에 연결된 별도의 사용자 단말로 하여금 상기 제1 클래스 정보 및 상기 제2 클래스 정보를 제공하도록 함으로써 상기 기계 학습 모델의 신뢰도의 판단이 가능하도록 지원하는 프로세스 중 적어도 하나를 수행하는 것을 특징으로 하는, 후처리 장치.

청구항 19

제15항에 있어서,

상기 (I) 프로세스에서,

상기 전체집합워드를 추출하면, 상기 프로세서가, 상기 전체집합워드에 대한 정보를 참조로 하여 상기 전체집합워드의 데이터를 포함하는 제1 딕셔너리를 생성하는 프로세스를 추가로 수행하고,

상기 (II) 프로세스에서,

상기 부분집합워드를 추출하면, 상기 프로세서가, 상기 부분집합워드에 대한 정보를 참조로 하여 상기 특정 학습 페이로드 데이터 각각에 대응되는 부분집합워드의 데이터를 포함하는 제2 딕셔너리 각각을 생성하는 프로세스를 추가로 수행하여,

상기 프로세서가, 상기 제1 딕셔너리 및 상기 제2 딕셔너리를 참조로 하여 상기 제1 가중치 및 상기 제2 가중치를 산출하는 것을 특징으로 하는, 후처리 장치.

청구항 20

제15항에 있어서,

상기 (II) 프로세스에서,

상기 프로세서가, 사전에 결정되어 있는 복수의 사전공격워드에 대한 정보를 추가로 참조하여, 상기 미확인 페이로드 데이터에서 추출된 상기 비교대상워드 중 상기 사전공격워드에 해당되는 비교대상워드에 대해서만 상기 제3 가중치를 산출하는 것을 특징으로 하는, 후처리 장치.

청구항 21

제15항에 있어서,

상기 학습 페이로드 데이터 각각은 그 각각에 복수개의 정답 레이블이 부여되되, 그 중 일부의 정답 레이블은 상기 학습 페이로드 데이터 각각에 대한 오탐 또는 정탐에 대한 정답 레이블 및 상기 학습 페이로드 데이터 각각이 해당되는 공격 유형 정보에 대한 정답 레이블 중 적어도 일부를 포함하는 상태에서, 상기 학습 페이로드 데이터 각각이 상기 기계 학습 모델의 학습에 제공되는 것을 특징으로 하는, 후처리 장치.

청구항 22

제15항에 있어서,

상기 학습 페이로드 데이터는, 소정의 보안 위협 탐지 시스템에 의하여 탐지된 복수의 탐지 로그 데이터 각각에 대응되는 페이로드 데이터로서, 상기 학습 페이로드 데이터 각각에, 이에 해당되는 소정의 클래스에 대한 상기 정답 레이블이 부여되어 연동되도록 지원되는 것을 특징으로 하는, 후처리 장치.

청구항 23

제22항에 있어서,

상기 학습 페이로드 데이터 각각은 그 각각에 복수개의 정답 레이블이 부여되되, 그 중 일부의 정답 레이블은 상기 학습 페이로드 데이터 각각에 대한 오탐 또는 정탐에 대한 정답 레이블 및 상기 학습 페이로드 데이터 각각이 해당되는 공격 유형에 대응되는 정답 레이블 중 적어도 일부를 포함하는 상태에서, 상기 학습 페이로드 데이터 각각이 상기 기계 학습 모델의 학습에 제공되는 것을 특징으로 하는, 후처리 장치.

청구항 24

제23항에 있어서,

상기 학습 페이로드 데이터 각각은, 상기 학습 페이로드 데이터에 포함된 숫자, 특수문자 및 문자열 중 하나 또는 둘 이상의 조합인 워드 중 상기 학습 페이로드 데이터 각각이 해당되는 공격 유형에 대응되는 워드를 공격 키워드로서 별도로 분류한 상태에서 상기 기계 학습 모델의 학습에 사용되는 것을 특징으로 하는, 후처리 장치.

청구항 25

제15항에 있어서,

상기 미확인 페이로드 데이터는, 상기 학습 페이로드 데이터를 사용하여 상기 기계 학습 모델의 학습이 완료된 상태에서 신규로 소정의 보안 위협 탐지 시스템에 입력되어 탐지되는 특정 탐지 로그 데이터에 대응되는 페이로드 데이터인 것을 특징으로 하는, 후처리 장치.

청구항 26

제25항에 있어서,

상기 미확인 페이로드 데이터는, 별도의 정답 레이블이 부여되지 않은 페이로드 데이터인 상태로 상기 기계 학습 모델 및 상기 후처리 장치 각각에 제공됨으로써, 상기 미확인 페이로드 데이터에 대응되는 상기 제1 클래스 정보 및 상기 제2 클래스 정보가 각각 획득되는 것을 특징으로 하는, 후처리 장치.

청구항 27

제26항에 있어서,

상기 (II) 프로세스에서,

상기 프로세서가, 사전에 결정되어 있는 복수의 사전공격워드에 대한 정보를 추가로 참조하여, 상기 미확인 페이로드 데이터에서 추출된 상기 비교대상워드 중 상기 사전공격워드에 해당되는 비교대상워드에 대해서만 상기 제3 가중치를 산출하는 것을 특징으로 하는, 후처리 장치.

청구항 28

제15항에 있어서,

상기 (V) 프로세스 이후에,

(VI) 상기 제1 클래스 정보 및 상기 제2 클래스 정보가 일치하지 않거나, 상기 제1 클래스 정보가 도출되지 않는 경우, 상기 프로세서가, 상기 기계 학습 모델의 재학습이 가능하도록 지원하는 프로세스를 추가로 수행하는, 후처리 장치.

발명의 설명

기 술 분 야

[0001] 본 발명은 기계 학습 모델의 신뢰도를 판단하기 위한 후처리 방법 및 이를 사용한 후처리 장치에 대한 것이다.

배 경 기 술

[0002] 최근 기술의 발전과 더불어 방대한 데이터의 통제 및 활용이 가능해지게 되면서, 수집된 데이터를 활용하여 기계 학습을 수행하여 인공지능을 포함한 기계 학습 모델을 개선하고 발전시키기 위한 많은 연구가 이루어지고 있

다. 하지만, 기계 학습 알고리즘으로 학습된 기계 학습 모델이 어떻게 작동하는지에 대해서는 블랙박스와도 같은 그 특성상 명확한 설명이 어려운 한계가 있고, 이는 기계 학습 모델이 도출한 결과값을 신뢰할 수 있는지에 대한 문제로 귀결되고 있다.

[0003] 한 인터넷 매체의 기사(한수연, "인간이 이해 못하는 인공지능, 믿어도 되나", https://www.bloter.net/archives/277243)에서 이러한 문제에 대한 내용을 확인할 수 있다.

[0004] 따라서, 기계 학습 알고리즘으로 학습된 기계 학습 모델을 신뢰할 수 있는지에 대한 근거를 제시하고, 이를 바탕으로 기계 학습 모델의 결과를 검증할 수 있는 방법이 필요한 실정이다.

발명의 내용

해결하려는 과제

[0005] 따라서, 상술한 문제점을 모두 해결하는 것을 그 목적으로 한다.

[0006] 또한, 본 발명은, 기계 학습 모델의 신뢰도를 판단하기 위한 후처리 방법을 제공함으로써, 기계 학습 모델이 도출한 결과값에 대한 검증을 효율적으로 수행할 수 있도록 하는 것을 다른 목적으로 한다.

[0007] 또한, 본 발명은, 기계 학습 모델의 신뢰도를 판단하기 위한 후처리 방법을 제공함으로써, 시의적절한 기계 학습 모델의 재학습의 필요성 판단이 가능하도록 지원하는 것을 그 목적으로 한다.

과제의 해결 수단

[0008] 상기한 바와 같은 본 발명의 목적을 달성하고, 후술하는 본 발명의 특징적인 효과를 실현하기 위한, 본 발명의 특징적인 구성은 하기와 같다.

[0009] 본 발명의 일 태양에 따르면, 기계 학습 모델의 신뢰도를 판단하기 위한 방법으로서, (a) 복수의 학습 페이로드 데이터 - 상기 학습 페이로드 데이터 각각은, 해당되는 소정의 클래스에 대한 정보인 정답 레이블이 부여됨 - 를 사용하여 상기 기계 학습 모델의 학습이 완료된 후, 미확인 페이로드 데이터가 획득되면, 후처리 장치가, 상기 학습 페이로드 데이터 중 적어도 일부로부터 전체집합워드 - 상기 전체집합워드 각각은 상기 학습 페이로드 데이터에 포함된 숫자, 특수문자 및 문자열 중 하나 또는 둘 이상의 조합임 - 를 추출한 상태에서, 상기 미확인 페이로드 데이터로부터 비교대상워드 - 상기 비교대상워드 각각은 상기 미확인 페이로드 데이터에 포함된 숫자, 특수문자 및 문자열 중 하나 또는 둘 이상의 조합임 - 를 추출하는 단계; (b) 상기 후처리 장치가, (i) 상기 학습 페이로드 데이터에서 추출된 상기 전체집합워드의 개수에 대한 정보 및 상기 학습 페이로드 데이터에서 상기 전체집합워드 각각이 출현하는 횟수에 대한 정보를 참조로 하여, 상기 전체집합워드 각각에 대응되는 제1 가중치를 산출하는 프로세스, (ii) 상기 학습 페이로드 데이터 중 적어도 하나의 특정 학습 페이로드 데이터에서 추출된 부분집합워드의 개수에 대한 정보 및 상기 특정 학습 페이로드 데이터에서 상기 부분집합워드 각각이 출현하는 횟수에 대한 정보를 참조로 하여, 상기 부분집합워드 각각에 대응되는 제2 가중치를 산출하는 프로세스, 및 (iii) 상기 미확인 페이로드 데이터에서 추출된 상기 비교대상워드의 개수에 대한 정보 및 상기 미확인 페이로드 데이터에서 상기 비교대상워드 각각이 출현하는 출현 횟수에 대한 정보를 참조로 하여, 상기 비교대상워드 각각에 대응되는 제3 가중치를 산출하는 프로세스를 수행하는 단계; (c) 상기 후처리 장치가, 상기 미확인 페이로드 데이터에서 추출된 상기 비교대상워드와 상기 특정 학습 페이로드 데이터 각각에서 추출된 각각의 상기 부분집합워드를 비교하여, 중복되는 중복워드 각각에 대응되는 제1 가중치, 제2 가중치 및 제3 가중치를 참조로 하여 상기 미확인 페이로드 데이터와 상기 특정 학습 페이로드 데이터 각각의 유사도를 판단하는 단계; (d) 상기 후처리 장치가, (i) 상기 학습 페이로드 데이터 각각에 대응되는 상기 유사도 중 가장 큰 값을 가지는 최대 유사도 및 이를 기준으로 하여 소정의 범위 이내에 포함되는 유사도에 해당되는 학습 페이로드 데이터 각각을 유사 페이로드 데이터로서 결정하는 프로세스, 및 (ii) 상기 유사 페이로드 각각에 부여된 정답 레이블 각각의 클래스 정보를 참조로 하여, 소정의 비율 이상의 특정 클래스에 해당되는 값을 상기 미확인 페이로드 데이터의 제1 클래스 정보로 결정하는 프로세스를 수행하는 단계; 및 (e) 상기 후처리 장치가, 상기 기계 학습 모델에 의하여 도출된 상기 미확인 페이로드 데이터의 클래스에 대한 모델 예측값이 제2 클래스 정보로서 획득된 상태에서, 상기 제1 클래스 정보와 상기 제2 클래스 정보를 비교하여 상기 기계 학습 모델의 신뢰도를 판단하거나 판단할 수 있도록 지원하는 단계; 를 포함하는 방법이 제공된다.

[0010] 일례로서, 상기 (c) 단계는, (c1) 상기 후처리 장치가, (i) 상기 중복워드 각각에 대응되는 제1 가중치로 상기 중복워드 각각에 대응되는 제2 가중치를 나눈 값인 학습데이터가중치를 산출하는 프로세스, (ii) 상기 중복워드

각각에 대응되는 제1 가중치로 상기 중복워드 각각에 대응되는 제3 가중치를 나눈 값인 미확인데이터가중치를 산출하는 프로세스, 및 (iii) 상기 학습데이터가중치와 상기 미확인데이터가중치를 참조로 하여 소정의 제1 연산을 수행하고, 그 결과로서 상기 중복워드 각각의 최종가중치를 산출하는 프로세스를 수행하는 단계; 및 (c2) 상기 후처리 장치가, 상기 중복워드 각각에 대응되는 상기 최종가중치 각각에 대하여 소정의 제2 연산을 수행한 결과값을 상기 유사도로서 획득하는 단계; 를 포함하는 방법이 제공된다.

[0011] 일례로서, 상기 (e) 단계에서, 상기 신뢰도의 판단은, 상기 후처리 장치가, 상기 제1 클래스 정보와 상기 제2 클래스 정보를 비교하여, (i) 상기 제1 클래스 정보와 상기 제2 클래스 정보가 일치하는 경우 상기 모델 예측값을 상기 미확인 페이로드 데이터에 대응되는 클래스 값으로 판단하는 프로세스, 및 (ii) (1) 상기 제1 클래스 정보와 상기 제2 클래스 정보가 일치하지 않거나 (2) 상기 제1 클래스 정보가 도출되지 않는 경우에는 상기 모델 예측값을 별도의 검사 대상으로 분류하는 프로세스 중 적어도 하나를 수행함으로써 이루어지는 것을 특징으로 하는 방법이 제공된다.

[0012] 일례로서, 상기 모델 예측값이 별도의 검사 대상으로 분류되는 경우, 상기 후처리 장치가, (i) 상기 제1 클래스 정보 및 상기 제2 클래스 정보를 제공함으로써 상기 기계 학습 모델의 신뢰도의 판단이 가능하도록 지원하는 프로세스, 및 (ii) 상기 후처리 장치에 연결된 별도의 사용자 단말로 하여금 상기 제1 클래스 정보 및 상기 제2 클래스 정보를 제공하도록 함으로써 상기 기계 학습 모델의 신뢰도의 판단이 가능하도록 지원하는 프로세스 중 적어도 하나를 수행하는 것을 특징으로 하는 방법이 제공된다.

[0013] 일례로서, 상기 (a) 단계에서, 상기 전체집합워드를 추출하면, 상기 후처리 장치가, 상기 전체집합워드에 대한 정보를 참조로 하여 상기 전체집합워드의 데이터를 포함하는 제1 딕셔너리를 생성하는 프로세스를 추가로 수행하고, 상기 (b) 단계에서, 상기 부분집합워드를 추출하면, 상기 후처리 장치가, 상기 부분집합워드에 대한 정보를 참조로 하여 상기 특정 학습 페이로드 데이터 각각에 대응되는 부분집합워드의 데이터를 포함하는 제2 딕셔너리 각각을 생성하는 프로세스를 추가로 수행하여, 상기 후처리 장치가, 상기 제1 딕셔너리 및 상기 제2 딕셔너리를 참조로 하여 상기 제1 가중치 및 상기 제2 가중치를 산출하는 것을 특징으로 하는 방법이 제공된다.

[0014] 일례로서, 상기 (b) 단계에서, 상기 후처리 장치가, 사전에 결정되어 있는 복수의 사전공격워드에 대한 정보를 추가로 참조하여, 상기 미확인 페이로드 데이터에서 추출된 상기 비교대상워드 중 상기 사전공격워드에 해당되는 비교대상워드에 대해서만 상기 제3 가중치를 산출하는 것을 특징으로 하는 방법이 제공된다.

[0015] 일례로서, 상기 학습 페이로드 데이터 각각은 그 각각에 복수개의 정답 레이블이 부여되되, 그 중 일부의 정답 레이블은 상기 학습 페이로드 데이터 각각에 대한 오탐 또는 정탐에 대한 정답 레이블 및 상기 학습 페이로드 데이터 각각이 해당되는 공격 유형 정보에 대한 정답 레이블 중 적어도 일부를 포함하는 상태에서, 상기 학습 페이로드 데이터 각각이 상기 기계 학습 모델의 학습에 제공되는 것을 특징으로 하는 방법이 제공된다.

[0016] 일례로서, 상기 학습 페이로드 데이터는, 소정의 보안 위협 탐지 시스템에 의하여 탐지된 복수의 탐지 로그 데이터 각각에 대응되는 페이로드 데이터로서, 상기 학습 페이로드 데이터 각각에, 이에 해당되는 소정의 클래스에 대한 상기 정답 레이블이 부여되어 연동되도록 지원되는 것을 특징으로 하는 방법이 제공된다.

[0017] 일례로서, 상기 학습 페이로드 데이터 각각은 그 각각에 복수개의 정답 레이블이 부여되되, 그 중 일부의 정답 레이블은 상기 학습 페이로드 데이터 각각에 대한 오탐 또는 정탐에 대한 정답 레이블 및 상기 학습 페이로드 데이터 각각이 해당되는 공격 유형에 대응되는 정답 레이블 중 적어도 일부를 포함하는 상태에서, 상기 학습 페이로드 데이터 각각이 상기 기계 학습 모델의 학습에 제공되는 것을 특징으로 하는 방법이 제공된다.

[0018] 일례로서, 상기 학습 페이로드 데이터 각각은, 상기 학습 페이로드 데이터에 포함된 숫자, 특수문자 및 문자열 중 하나 또는 둘 이상의 조합인 워드 중 상기 학습 페이로드 데이터 각각이 해당되는 공격 유형에 대응되는 워드를 공격 키워드로서 별도로 분류한 상태에서 상기 기계 학습 모델의 학습에 사용되는 것을 특징으로 하는 방법이 제공된다.

[0019] 일례로서, 상기 미확인 페이로드 데이터는, 상기 학습 페이로드 데이터를 사용하여 상기 기계 학습 모델의 학습이 완료된 상태에서 신규로 소정의 보안 위협 탐지 시스템에 입력되어 탐지되는 특정 탐지 로그 데이터에 대응되는 페이로드 데이터인 것을 특징으로 하는 방법이 제공된다.

[0020] 일례로서, 상기 미확인 페이로드 데이터는, 별도의 정답 레이블이 부여되지 않은 페이로드 데이터인 상태로 상기 기계 학습 모델 및 상기 후처리 장치 각각에 제공됨으로써, 상기 미확인 페이로드 데이터에 대응되는 상기 제1 클래스 정보 및 상기 제2 클래스 정보가 각각 획득되는 것을 특징으로 하는 방법이 제공된다.

[0021] 일례로서, 상기 (b) 단계에서, 상기 후처리 장치가, 사전에 결정되어 있는 복수의 사전공격워드에 대한 정보를 추가로 참조하여, 상기 미확인 페이로드 데이터에서 추출된 상기 비교대상워드 중 상기 사전공격워드에 해당되는 비교대상워드에 대해서만 상기 제3 가중치를 산출하는 것을 특징으로 하는 방법이 제공된다.

[0022] 일례로서, 상기 (e) 단계 이후에, (f) 상기 제1 클래스 정보 및 상기 제2 클래스 정보가 일치하지 않거나, 상기 제1 클래스 정보가 도출되지 않는 경우, 상기 후처리 장치가, 상기 기계 학습 모델의 재학습이 가능하도록 지원하는 단계를 추가로 포함하는 방법이 제공된다.

[0023] 또한, 본 발명의 다른 태양에 따르면, 기계 학습 모델의 신뢰도를 판단하기 위한 후처리 장치로서, 인스트럭션들을 저장하는 적어도 하나의 메모리; 및 상기 인스트럭션들을 실행하기 위해 구성된 적어도 하나의 프로세서; 를 포함하고, 상기 프로세서가, (I) 복수의 학습 페이로드 데이터 - 상기 학습 페이로드 데이터 각각은, 해당되는 소정의 클래스에 대한 정보인 정답 레이블이 부여됨 - 를 사용하여 상기 기계 학습 모델의 학습이 완료된 후, 미확인 페이로드 데이터가 획득되면, 상기 학습 페이로드 데이터 중 적어도 일부로부터 전체집합워드 - 상기 전체집합워드 각각은 상기 학습 페이로드 데이터에 포함된 숫자, 특수문자 및 문자열 중 하나 또는 둘 이상의 조합임 - 를 추출한 상태에서, 상기 미확인 페이로드 데이터로부터 비교대상워드 - 상기 비교대상워드 각각은 상기 미확인 페이로드 데이터에 포함된 숫자, 특수문자 및 문자열 중 하나 또는 둘 이상의 조합임 - 를 추출하는 프로세스; (II) (i) 상기 학습 페이로드 데이터에서 추출된 상기 전체집합워드의 개수에 대한 정보 및 상기 학습 페이로드 데이터에서 상기 전체집합워드 각각이 출현하는 횟수에 대한 정보를 참조로 하여, 상기 전체집합워드 각각에 대응되는 제1 가중치를 산출하는 서브프로세스, (ii) 상기 학습 페이로드 데이터 중 적어도 하나의 특정 학습 페이로드 데이터에서 추출된 부분집합워드의 개수에 대한 정보 및 상기 특정 학습 페이로드 데이터에서 상기 부분집합워드 각각이 출현하는 횟수에 대한 정보를 참조로 하여, 상기 부분집합워드 각각에 대응되는 제2 가중치를 산출하는 서브프로세스, 및 (iii) 상기 미확인 페이로드 데이터에서 추출된 상기 비교대상워드의 개수에 대한 정보 및 상기 미확인 페이로드 데이터에서 상기 비교대상워드 각각이 출현하는 출현 횟수에 대한 정보를 참조로 하여, 상기 비교대상워드 각각에 대응되는 제3 가중치를 산출하는 서브프로세스를 수행하는 프로세스 ; (III) 상기 미확인 페이로드 데이터에서 추출된 상기 비교대상워드와 상기 특정 학습 페이로드 데이터 각각에서 추출된 각각의 상기 부분집합워드를 비교하여, 중복되는 중복워드 각각에 대응되는 제1 가중치, 제2 가중치 및 제3 가중치를 참조로 하여 상기 미확인 페이로드 데이터와 상기 특정 학습 페이로드 데이터 각각의 유사도를 판단하는 프로세스; (IV) (i) 상기 학습 페이로드 데이터 각각에 대응되는 상기 유사도 중 가장 큰 값을 가지는 최대유사도 및 이를 기준으로 하여 소정의 범위 이내에 포함되는 유사도에 해당되는 학습 페이로드 데이터 각각을 유사 페이로드 데이터로서 결정하는 서브프로세스, 및 (ii) 상기 유사 페이로드 각각에 부여된 정답 레이블 각각의 클래스 정보를 참조로 하여, 소정의 비율 이상의 특정 클래스에 해당되는 값을 상기 미확인 페이로드 데이터의 제1 클래스 정보로 결정하는 서브프로세스를 수행하는 프로세스; 및 (V) 상기 기계 학습 모델에 의하여 도출된 상기 미확인 페이로드 데이터의 클래스에 대한 모델 예측값이 제2 클래스 정보로서 획득된 상태에서, 상기 제1 클래스 정보와 상기 제2 클래스 정보를 비교하여 상기 기계 학습 모델의 신뢰도를 판단하거나 판단할 수 있도록 지원하는 프로세스; 를 수행하는 후처리 장치가 제공된다.

[0024] 일례로서, 상기 (III) 프로세스는, 상기 프로세서가, (III-1) (i) 상기 중복워드 각각에 대응되는 제1 가중치로 상기 중복워드 각각에 대응되는 제2 가중치를 나눈 값인 학습데이터가중치를 산출하는 서브프로세스, (ii) 상기 중복워드 각각에 대응되는 제1 가중치로 상기 중복워드 각각에 대응되는 제3 가중치를 나눈 값인 미확인데이터가중치를 산출하는 서브프로세스, 및 (iii) 상기 학습데이터가중치와 상기 미확인데이터가중치를 참조로 하여 소정의 제1 연산을 수행하고, 그 결과로서 상기 중복워드 각각의 최종가중치를 산출하는 서브프로세스를 수행하는 프로세스; 및 (III-2) 상기 중복워드 각각에 대응되는 상기 최종가중치 각각에 대하여 소정의 제2 연산을 수행한 결과값을 상기 유사도로서 획득하는 프로세스; 를 수행하는 후처리 장치가 제공된다.

[0025] 일례로서, 상기 (V) 프로세스에서, 상기 신뢰도의 판단은, 상기 프로세서가, 상기 제1 클래스 정보와 상기 제2 클래스 정보를 비교하여, (i) 상기 제1 클래스 정보와 상기 제2 클래스 정보가 일치하는 경우 상기 모델 예측값을 상기 미확인 페이로드 데이터에 대응되는 클래스 값으로 판단하는 프로세스, 및 (ii) (1) 상기 제1 클래스 정보와 상기 제2 클래스 정보가 일치하지 않거나 (2) 상기 제1 클래스 정보가 도출되지 않는 경우에는 상기 모델 예측값을 별도의 검사 대상으로 분류하는 프로세스 중 적어도 하나를 수행함으로써 이루어지는 것을 특징으로 하는 후처리 장치가 제공된다.

[0026] 일례로서, 상기 모델 예측값이 별도의 검사 대상으로 분류되는 경우, 상기 프로세서가, (i) 상기 제1 클래스 정보 및 상기 제2 클래스 정보를 제공함으로써 상기 기계 학습 모델의 신뢰도의 판단이 가능하도록 지원하는 프로세스, 및 (ii) 상기 후처리 장치에 연결된 별도의 사용자 단말로 하여금 상기 제1 클래스 정보 및 상기 제2 클

래스 정보를 제공하도록 함으로써 상기 기계 학습 모델의 신뢰도의 판단이 가능하도록 지원하는 프로세스 중 적어도 하나를 수행하는 것을 특징으로 하는 후처리 장치가 제공된다.

[0027] 일례로서, 상기 (I) 프로세스에서, 상기 전체집합워드를 추출하면, 상기 프로세서가, 상기 전체집합워드에 대한 정보를 참조로 하여 상기 전체집합워드의 데이터를 포함하는 제1 딕셔너리를 생성하는 프로세스를 추가로 수행하고, 상기 (II) 프로세스에서, 상기 부분집합워드를 추출하면, 상기 프로세서가, 상기 부분집합워드에 대한 정보를 참조로 하여 상기 특정 학습 페이로드 데이터 각각에 대응되는 부분집합워드의 데이터를 포함하는 제2 딕셔너리 각각을 생성하는 프로세스를 추가로 수행하여, 상기 프로세서가, 상기 제1 딕셔너리 및 상기 제2 딕셔너리를 참조로 하여 상기 제1 가중치 및 상기 제2 가중치를 산출하는 것을 특징으로 하는 후처리 장치가 제공된다.

[0028] 일례로서, 상기 (II) 프로세스에서, 상기 프로세서가, 사전에 결정되어 있는 복수의 사전공격워드에 대한 정보를 추가로 참조하여, 상기 미확인 페이로드 데이터에서 추출된 상기 비교대상워드 중 상기 사전공격워드에 해당되는 비교대상워드에 대해서만 상기 제3 가중치를 산출하는 것을 특징으로 하는 후처리 장치가 제공된다.

[0029] 일례로서, 상기 학습 페이로드 데이터 각각은 그 각각에 복수개의 정답 레이블이 부여되되, 그 중 일부의 정답 레이블은 상기 학습 페이로드 데이터 각각에 대한 오탐 또는 정탐에 대한 정답 레이블 및 상기 학습 페이로드 데이터 각각이 해당되는 공격 유형 정보에 대한 정답 레이블 중 적어도 일부를 포함하는 상태에서, 상기 학습 페이로드 데이터 각각이 상기 기계 학습 모델의 학습에 제공되는 것을 특징으로 하는 후처리 장치가 제공된다.

[0030] 일례로서, 상기 학습 페이로드 데이터는, 소정의 보안 위협 탐지 시스템에 의하여 탐지된 복수의 탐지 로그 데이터 각각에 대응되는 페이로드 데이터로서, 상기 학습 페이로드 데이터 각각에, 이에 해당되는 소정의 클래스에 대한 상기 정답 레이블이 부여되어 연동되도록 지원되는 것을 특징으로 하는 후처리 장치가 제공된다.

[0031] 일례로서, 상기 학습 페이로드 데이터 각각은 그 각각에 복수개의 정답 레이블이 부여되되, 그 중 일부의 정답 레이블은 상기 학습 페이로드 데이터 각각에 대한 오탐 또는 정탐에 대한 정답 레이블 및 상기 학습 페이로드 데이터 각각이 해당되는 공격 유형에 대응되는 정답 레이블 중 적어도 일부를 포함하는 상태에서, 상기 학습 페이로드 데이터 각각이 상기 기계 학습 모델의 학습에 제공되는 것을 특징으로 하는 후처리 장치가 제공된다.

[0032] 일례로서, 상기 학습 페이로드 데이터 각각은, 상기 학습 페이로드 데이터에 포함된 숫자, 특수문자 및 문자열 중 하나 또는 둘 이상의 조합인 워드 중 상기 학습 페이로드 데이터 각각이 해당되는 공격 유형에 대응되는 워드를 공격 키워드로서 별도로 분류한 상태에서 상기 기계 학습 모델의 학습에 사용되는 것을 특징으로 하는 후처리 장치가 제공된다.

[0033] 일례로서, 상기 미확인 페이로드 데이터는, 상기 학습 페이로드 데이터를 사용하여 상기 기계 학습 모델의 학습이 완료된 상태에서 신규로 소정의 보안 위협 탐지 시스템에 입력되어 탐지되는 특정 탐지 로그 데이터에 대응되는 페이로드 데이터인 것을 특징으로 하는 후처리 장치가 제공된다.

[0034] 일례로서, 상기 미확인 페이로드 데이터는, 별도의 정답 레이블이 부여되지 않은 페이로드 데이터인 상태로 상기 기계 학습 모델 및 상기 후처리 장치 각각에 제공됨으로써, 상기 미확인 페이로드 데이터에 대응되는 상기 제1 클래스 정보 및 상기 제2 클래스 정보가 각각 획득되는 것을 특징으로 하는 후처리 장치가 제공된다.

[0035] 일례로서, 상기 (II) 프로세스에서, 상기 프로세서가, 사전에 결정되어 있는 복수의 사전공격워드에 대한 정보를 추가로 참조하여, 상기 미확인 페이로드 데이터에서 추출된 상기 비교대상워드 중 상기 사전공격워드에 해당되는 비교대상워드에 대해서만 상기 제3 가중치를 산출하는 것을 특징으로 하는 후처리 장치가 제공된다.

[0036] 일례로서, 상기 (V) 프로세스 이후에, (VI) 상기 제1 클래스 정보 및 상기 제2 클래스 정보가 일치하지 않거나, 상기 제1 클래스 정보가 도출되지 않는 경우, 상기 프로세서가, 상기 기계 학습 모델의 재학습이 가능하도록 지원하는 프로세스를 추가로 수행하는 후처리 장치가 제공된다.

발명의 효과

[0037] 본 발명에 의하면, 다음과 같은 효과가 있다.

[0038] 본 발명은, 기계 학습 모델의 신뢰도를 판단하기 위한 방법을 제공함으로써, 기계 학습 모델이 도출한 결과를 신뢰할 수 있는지에 대한 근거를 제시할 수 있는 효과가 있다.

[0039] 또한, 본 발명은, 기계 학습 모델의 신뢰도를 판단하기 위한 후처리 방법을 제공함으로써, 기계 학습 모델이 도

출한 결과값에 대한 검증을 효율적으로 수행할 수 있는 효과가 있다.

[0040] 또한, 본 발명은, 기계 학습 모델의 신뢰도를 판단하기 위한 후처리 방법을 제공함으로써, 시의적절한 기계 학습 모델의 재학습 및 추가학습의 필요성 판단이 가능하도록 지원할 수 있는 효과가 있다.

도면의 간단한 설명

[0041] 도 1은 본 발명의 일 실시예에 따른, 기계 학습 모델의 신뢰도를 판단하기 위한 후처리 장치를 개략적으로 나타내는 도면이다.

도 2a는 본 발명의 일 실시예에 따른, 기계 학습 모델의 학습을 위한 데이터의 흐름을 개략적으로 나타낸 도면이다.

도 2b는 본 발명의 일 실시예에 따른, 기계 학습 모델의 신뢰도를 판단하고 그에 따른 기계 학습 모델의 재학습을 지원하기 위한 후처리 장치의 후처리 과정을 개략적으로 나타낸 도면이다.

도 3은 본 발명의 일 실시예에 따른, 기계 학습 모델의 신뢰도를 판단하는 과정을 개략적으로 나타낸 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0042] 후술하는 본 발명에 대한 상세한 설명은, 본 발명이 실시될 수 있는 특정 실시예를 예시로서 도시하는 첨부 도면을 참조한다. 이들 실시예는 당업자가 본 발명을 실시할 수 있기에 충분하도록 상세히 설명된다. 본 발명의 다양한 실시예는 서로 다르지만 상호 배타적일 필요는 없음이 이해되어야 한다. 예를 들어, 여기에 기재되어 있는 특정 형상, 구조 및 특성은 일 실시예에 관련하여 본 발명의 정신 및 범위를 벗어나지 않으면서 다른 실시예로 구현될 수 있다.

[0043] 또한, 각각의 개시된 실시예 내의 개별 구성요소의 위치 또는 배치는 본 발명의 정신 및 범위를 벗어나지 않으면서 변경될 수 있음이 이해되어야 한다. 따라서, 후술하는 상세한 설명은 한정적인 의미로서 취하려는 것이 아니며, 본 발명의 범위는, 적절하게 설명된다면, 그 청구항들이 주장하는 것과 균등한 모든 범위와 더불어 첨부된 청구항에 의해서만 한정된다. 도면에서 유사한 참조부호는 여러 측면에 걸쳐서 동일하거나 유사한 기능을 지칭한다.

[0044] 이하, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자가 본 발명을 용이하게 실시할 수 있도록 하기 위하여, 본 발명의 바람직한 실시예들에 관하여 첨부된 도면을 참조하여 상세히 설명하기로 한다.

[0045] 도 1은 본 발명의 일 실시예에 따른, 기계 학습 모델의 신뢰도를 판단하기 위한 후처리 장치를 개략적으로 나타내는 도면이다.

[0046] 도 1을 참조하면, 기계 학습 모델의 신뢰도를 판단하기 위한 후처리 장치(100)는 메모리(110) 및 프로세서(120)를 포함할 수 있다. 이 때, 메모리(110)는, 프로세서(120)의 인스트럭션들을 저장할 수 있는데, 구체적으로, 인스트럭션들은 후처리 장치(100)로 하여금 특정의 방식으로 기능하게 하기 위한 목적으로 생성되는 코드로서, 컴퓨터 기타 프로그램 가능한 데이터 프로세싱 장비를 지향할 수 있는 컴퓨터 이용 가능 또는 컴퓨터 판독 가능 메모리에 저장될 수 있다. 인스트럭션들은 본 발명의 명세서에서 설명되는 기능들을 실행하기 위한 프로세스들을 수행할 수 있다.

[0047] 그리고, 프로세서(120)는, MPU(Micro Processing Unit) 또는 CPU(Central Processing Unit), 캐쉬 메모리(Cache Memory), 데이터 버스(Data Bus) 등의 하드웨어 구성을 포함할 수 있다. 또한, 운영체제, 특정 목적을 수행하는 애플리케이션의 소프트웨어 구성을 포함할 수 있다.

[0048] 다음으로, 후처리 장치(100)는 기계 학습 모델의 신뢰도를 판단하는 데 사용되는 정보를 포함하는 데이터베이스(미도시)와 연동될 수 있다. 이 때, 데이터베이스는 플래시 메모리 타입(flash memory type), 하드디스크 타입(hard disk type), 멀티미디어 카드 마이크로 타입(multimedia card micro type), 카드 타입의 메모리(예를 들어 SD 또는 XD 메모리), 램(Random Access Memory, RAM), SRAM(Static Random Access Memory), 롬(ReadOnly Memory, ROM), EEPROM(Electrically Erasable Programmable ReadOnly Memory), PROM(Programmable ReadOnly Memory), 자기 메모리, 자기 디스크, 광디스크 중 적어도 하나의 타입의 저장매체를 포함할 수 있으며, 이에 한정되지 않으며 데이터를 저장할 수 있는 모든 매체를 포함할 수 있다. 또한, 데이터베이스는 후처리 장치(100)의 내부에 설치되어 데이터를 전송하거나 수신되는 데이터를 기록할 수도 있으며, 이는 발명의 실시 조건에 따

라 달라질 수 있다.

[0049] 이와 같은 후처리 장치를 사용하여 기계 학습 모델의 신뢰도를 평가하기 위해서는 기계 학습 모델의 학습이 선행되어야 하는데, 이를 도 2a를 참조하여 설명하면 다음과 같다.

[0050] 도 2a는 본 발명의 일 실시예에 따른, 기계 학습 모델의 학습을 위한 데이터의 흐름을 개략적으로 나타낸 도면이다.

[0051] 도 2a를 참조하면, 기계 학습 시스템(200)의 기계 학습 모델은 복수개의 학습 페이로드 데이터(12)를 사용하여 학습이 이루어 질 수 있는데, 해당 학습 페이로드 데이터(12)는 소정의 보안 위협 탐지 시스템(10)에서 탐지되어 수집된 탐지 로그 데이터(11-1) 각각의 페이로드 데이터에 대하여 그 각각이 해당되는 소정의 클래스에 대한 정보인 정답 레이블이 부여된 데이터일 수 있다. 이 때, 발명의 실시 조건에 따라, 학습 페이로드 데이터(12) 각각에는 그 각각에 해당되는 정오탐 여부, 공격 유형 등에 대한 복수의 정답 레이블이 부여될 수도 있는데, 공격 유형에 대한 학습이 이루어지는 경우에는 도 2a에서 도시된 바와 같이 학습 페이로드 데이터 각각에 대하여 그에 해당되는 공격 유형별 키워드를 분류하는 과정이 추가로 이루어진 공격 유형별 키워드 분류 데이터(13)가 학습 데이터로서 사용될 수 있다.

[0052] 도 2b는 본 발명의 일 실시예에 따른, 기계 학습 모델의 신뢰도를 판단하고 그에 따른 기계 학습 모델의 재학습을 지원하기 위한 후처리 장치의 후처리 과정을 개략적으로 나타낸 도면이다.

[0053] 도 2b를 참조하면, 기계 학습 시스템(200)의 기계 학습 모델에 대한 학습이 완료된 상태에서, 보안 위협 탐지 시스템(10)이 신규로 탐지한 특정 탐지 로그 데이터(11-2)의 페이로드 데이터가 기계 학습 모델의 신뢰도 판단에 사용되는 미확인 페이로드 데이터(15)가 될 수 있다. 그리고, 미확인 페이로드 데이터(15)가 학습이 완료된 기계 학습 모델에 입력되면, 그에 대한 결과값으로서 미확인 페이로드 데이터(15)가 해당되는 클래스를 예측한 모델 예측값이 제2 클래스 정보(201)로서 도출될 수 있다.

[0054] 이와는 별도로, 후처리 장치(100)의 프로세서(120)는, 기계 학습 시스템(200)의 기계 학습 모델을 학습하는 데 사용된 학습 페이로드 데이터(12) 및 미확인 페이로드 데이터(15)를 획득하여, 이를 바탕으로 미확인 페이로드 데이터(15)에 해당되는 특정 클래스에 해당되는 값을 제1 클래스 정보(101)로서 결정할 수 있다. 이 때, 후처리 장치(100)의 프로세서(120)가 제1 클래스 정보(101)를 결정하기 위한 세부적인 내용은, 아래에서 도 3을 참조하여 상세히 설명할 것이다.

[0055] 상술한 과정을 거쳐 제1 클래스 정보 및 제2 클래스 정보가 획득되면, 후처리 장치(100)의 프로세서(120)가 제1 클래스 정보와 제2 클래스 정보를 비교하여, 그 결과에 따라 기계 학습 시스템(200)의 기계 학습 모델에 대한 신뢰도를 판단할 수 있는데, (i) 제1 클래스 정보와 제2 클래스 정보가 일치하는 경우에는 기계 학습 모델을 신뢰할 수 있는 것으로 판단하여 미확인 페이로드 데이터(15)의 클래스 값을 제2 클래스 정보에 대응되는 모델 예측값으로 판단할 수 있고, (ii) (1) 제1 클래스 정보와 제2 클래스 정보가 일치하지 않거나 (2) 제1 클래스 정보가 도출되지 않는 경우에는 기계 학습 모델을 신뢰할 수 없는 것으로 판단하여 제2 클래스 정보에 대응되는 모델 예측값을 별도의 검사 대상으로 분류할 수 있다. 이 때, 기계 학습 모델을 신뢰할 수 없는 것으로 판단되는 경우, 기계 학습 모델의 재학습이 가능하도록 후처리 장치(100)의 프로세서(120)가 지원하는 프로세스가 추가적으로 수행될 수 있다.

[0056] 도 3은 본 발명의 일 실시예에 따른, 기계 학습 모델의 신뢰도를 판단하는 과정을 개략적으로 나타낸 흐름도이다.

[0057] 도 3을 참조하면, 후처리 장치(100)의 프로세서(120)가 기계 학습 모델의 신뢰도를 판단하는 과정은, 학습 페이로드 데이터(12)를 사용하여 기계 학습 시스템(200)의 기계 학습 모델에 대한 학습이 완료된 후, 미확인 페이로드 데이터(15)가 획득되면, 프로세서(120)가 학습 페이로드 데이터(12)로부터 전체집합워드를 추출한 상태에서 미확인 페이로드 데이터(15)로부터 비교대상워드를 추출(S301)하는 것으로부터 시작된다. 이 때, 학습 페이로드 데이터(12)는 소정의 보안 위협 탐지 시스템(10)에 의하여 탐지되어 수집된 탐지 로그 데이터(11-1)의 페이로드 데이터일 수 있으나, 이에 한정되는 것은 아니며, 별도의 과정을 통하여 준비된 데이터일 수도 있다. 또한, 미확인 페이로드 데이터(15)는, 소정의 보안 위협 시스템(10)에서 탐지된 탐지 로그 데이터(11-1)에 대응되는 학습 페이로드 데이터(12)를 사용하여 기계 학습 모델의 학습이 완료된 후, 신규로 보안 위협 시스템(10)에서 탐지된 특정 탐지 로그 데이터(11-2)에 대응되는 페이로드 데이터일 수 있다.

[0058] 그리고, 전체집합워드는 추출된 학습 페이로드 데이터에 포함된 숫자, 특수문자 및 문자열 중 하나 또는 둘 이상의 조합으로서, 학습 페이로드 데이터에 포함된 워드 모두를 의미할 수 있으나 이에 한정되는 것은 아니며,

비교대상워드 역시 그에 대응되는 미확인 페이로드 데이터에 포함된 숫자, 특수문자 및 문자열 중 하나 또는 둘 이상의 조합일 수 있다. 예를 들어, 'select', 'from', 'where', 'join', 'table_name', '=', '1' 등이 페이로드 데이터에 포함되어 있을 경우 페이로드 데이터의 워드로서 추출될 수 있다.

[0059] 또한, 학습 페이로드 데이터(12)는 그 각각이 해당되는 소정의 클래스에 대한 정보인 정답 레이블이 부여된 데이터일 수 있다. 발명의 일 예로서, 학습 페이로드 데이터(12) 각각은 정탐인 경우 1, 오탐인 경우 0의 클래스 값이 정답 레이블로서 부여되어 있을 수 있다. 또 다른 발명의 일 예로서, XSS, SQL Injection, File upload 등 학습 페이로드 데이터(12) 각각이 해당되는 공격 유형 정보에 해당되는 정답 레이블이 부여되어 있을 수도 있으며, 복수개의 레이블이 부여되어 있을 수도 있다.

[0060] 그리고, 미확인 페이로드 데이터(15)는, 별도의 정답 레이블이 부여되지 않은 상태에서 기계 학습 장치(200)의 기계 학습 모델 및 후처리 장치(100)에 제공되어, 미확인 페이로드 데이터(15)에 대응되는 클래스 값이 제1 클래스 정보(101) 및 제2 클래스 정보(201)로 각각 도출될 수 있다.

[0061] 그리고 나서, 후처리 장치(100)의 프로세서(120)는 (i) 학습 페이로드 데이터에서 추출된 전체집합워드의 개수에 대한 정보 및 학습 페이로드 데이터에서 전체집합워드 각각이 출현하는 횟수에 대한 정보를 참조로 하여, 전체집합워드 각각에 대응되는 제1 가중치를 산출(S302-1)하는 프로세스, (ii) 학습 페이로드 데이터 중 적어도 하나의 특정 학습 페이로드 데이터에서 추출된 부분집합워드의 개수에 대한 정보 및 특정 학습 페이로드 데이터에서 부분집합워드 각각이 출현하는 횟수에 대한 정보를 참조로 하여, 부분집합워드 각각에 대응되는 제2 가중치를 산출(S302-2)하는 프로세스, 및 (iii) 미확인 페이로드 데이터에서 추출된 비교대상워드의 개수에 대한 정보 및 미확인 페이로드 데이터에서 비교대상워드 각각이 출현하는 출현 횟수에 대한 정보를 참조로 하여, 비교대상워드 각각에 대응되는 제3 가중치를 산출(S302-3)하는 프로세스를 수행할 수 있다. 예를 들어, 전체집합워드의 개수가 100개이고, 그 중 'select' 워드의 출현 횟수가 2번인 경우, 'select' 워드에 대응되는 제1 가중치는 'select' 워드의 출현 빈도인 0.2로 산출될 수 있다.

[0062] 이 때, 제1 가중치, 제2 가중치 및 제3 가중치를 산출하는 상기 프로세스의 수행 순서가 상기한 바와 같이 한정되는 것은 아니며, 발명의 실시 조건에 따라 그 순서가 달라지거나 둘 이상이 동시에 수행될 수도 있다. 또한, 발명의 실시 조건에 따라, 후처리 장치(100)의 프로세서(120)가 사전에 정해진 소정의 사전공격워드에 대한 정보를 참조로 하여, 추출된 비교대상워드 중 사전공격워드에 해당되는 비교대상워드에 대해서만 제3 가중치를 산출할 수도 있다.

[0063] 그리고, 발명의 일 예로서, 후처리 장치(100)의 프로세서(120)는 학습 페이로드 데이터의 적어도 일부로부터 전체집합워드를 추출하면 이를 참조로 하여 전체집합워드의 데이터를 포함하는 제1 딕셔너리를 생성하여 제1 가중치를 산출하는 과정에서 참조될 수 있도록 할 수 있다. 또한, 후처리 장치(100)의 프로세서(120)는 학습 페이로드 데이터 중 적어도 하나의 특정 학습 페이로드 데이터 각각에 대응되는 부분집합워드의 데이터를 포함하는 제2 딕셔너리도 함께 생성하여 이후 제2 가중치를 산출하는 과정에서 참조될 수 있도록 할 수 있으나, 제2 딕셔너리가 제1 딕셔너리와 반드시 함께 생성되어야 하는 것은 아니며, 발명의 실시 조건에 따라 그 생성 시기는 다르게 정해질 수 있다. 미확인 페이로드 데이터에 대응되는 비교대상워드의 데이터를 포함하는 딕셔너리도 발명의 실시 조건에 따라 생성되어 제3 가중치를 산출하는 과정에서 참조될 수도 있으나, 이는 필수적인 것은 아니며, 발명의 실시 조건에 따라 선택적으로 생성될 수 있다.

[0064] 다음으로, 후처리 장치(100)의 프로세서(120)는, 비교대상워드와 부분집합워드를 비교하여, 중복되는 중복워드 각각에 대응되는 제1 가중치, 제2 가중치 및 제3 가중치를 참조로 하여 미확인 페이로드 데이터와 특정 학습 페이로드 데이터 각각의 유사도를 판단(S303)할 수 있다. 이 때, 프로세서(120)는, (i) 중복워드 각각에 대응되는 제1 가중치로 제2 가중치를 나눈 값인 학습데이터가중치를 산출하는 프로세스, (ii) 중복워드 각각에 대응되는 제1 가중치로 제3 가중치를 나눈 값인 미확인데이터가중치를 산출하는 프로세스, 및 (iii) 산출된 학습데이터가중치와 미확인데이터가중치에 대하여 소정의 제1 연산을 수행하여 중복워드 각각의 최종가중치를 산출하는 프로세스를 각각 수행하고, 산출된 최종가중치 각각에 대하여 소정의 제2 연산을 수행하여 그 결과값을 상기 유사도로서 획득할 수 있다. 예를 들어, 'select'워드에 대응되는 제1 가중치가 0.5, 제2 가중치가 0.4, 제3 가중치가 0.3인 경우에, 'select'워드의 학습데이터가중치는 0.4를 0.5로 나눈 0.8이 산출되고, 미확인데이터가중치는 0.3을 0.5로 나눈 0.6이 산출될 수 있다. 그리고, 소정의 제1 연산이 학습데이터가중치와 미확인데이터가중치를 합산하는 것이라면'select'워드의 최종가중치는 1.4가 산출될 수 있다. 또한, 중복워드가 'select', 'from'이고, 'from'워드의 최종가중치가 1.7인 경우, 소정의 제2 연산이 중복워드 각각의 최종가중치를 모두 합산하는 것이라면 미확인 페이로드 데이터와 특정 학습 페이로드 데이터의 유사도는 3.1이 산출될 수 있을 것이

다.

[0065] 그리고 나서, 상술한 바와 같이 미확인 페이로드 데이터와 적어도 하나의 특정 학습 페이로드 데이터 각각의 유사도를 산출하면, 후처리 장치(100)의 프로세서(120)가 산출된 유사도 중에서 최대유사도 및 이를 기준으로 하여 소정의 범위 내에 포함되는 유사도에 해당되는 학습 페이로드 각각을 유사 페이로드 데이터로 결정(S304)하고, 유사 페이로드 데이터 각각에 부여되어 있는 정답 레이블의 클래스 정보를 참조로 하여 소정의 비율 이상의 특정 클래스에 해당되는 값을 미확인 페이로드 데이터의 제1 클래스 정보로 결정(S305)할 수 있다. 예를 들어, 제1 특정 학습 페이로드 데이터의 유사도가 1, 제2 특정 학습 페이로드 데이터의 유사도가 2, 제3 특정 학습 페이로드 데이터의 유사도가 3, 제4 특정 학습 페이로드 데이터의 유사도가 4이고 소정의 범위가 2.5인 경우, 최대유사도인 4에 해당되는 제4 특정 학습 페이로드 데이터와, 4의 유사도를 기준으로 하여 2.5의 범위 이내인 3과 2의 유사도에 해당되는 제3 특정 학습 페이로드 데이터와 제2 특정 학습 페이로드 데이터가 유사 페이로드 데이터로서 결정될 수 있다. 이 때, 제4 특정 학습 페이로드 데이터에 부여된 정답 레이블의 클래스 값이 1, 제3 특정 학습 페이로드 데이터에 부여된 정답 레이블의 클래스 값이 0, 제2 특정 학습 페이로드 데이터에 부여된 정답 레이블의 클래스 값이 1이고, 소정의 비율이 50%라면 1의 클래스 값이 미확인 페이로드 데이터의 제1 클래스 정보로 결정될 수 있을 것이다.

[0066] 다음으로, 후처리 장치(100)의 프로세서(120)가, 기계 학습 장치(200)의 기계 학습 모델에 의하여 도출된 미확인 페이로드 데이터의 클래스에 대한 모델 예측값이 제2 클래스 정보로서 획득된 상태에서 제1 클래스 정보와 제2 클래스 정보를 비교(S306)하여, 그 결과에 따라 기계 학습 모델의 신뢰도를 판단하거나 판단할 수 있도록 지원할 수 있다. 이를 더 자세히 설명하면, 후처리 장치(100)의 프로세서(120)가, 제1 클래스 정보와 제2 클래스 정보를 비교하여, (i) 제1 클래스 정보와 제2 클래스 정보가 일치하는 경우 기계 학습 모델을 신뢰할 수 있는 것으로 판단하여 미확인 페이로드 데이터에 대응되는 모델 예측값을 상기 미확인 페이로드 데이터에 대응되는 클래스 값으로 판단하는 프로세스, 및 (ii) (1) 제1 클래스 정보와 제2 클래스 정보가 일치하지 않거나 (2) 미확인 페이로드 데이터와 학습 페이로드 데이터에 중복워드가 존재하지 않아 제1 클래스 정보가 도출되지 않는 경우에는, 기계 학습 모델을 신뢰할 수 없는 것으로 판단하여 미확인 페이로드 데이터에 대응되는 모델 예측값을 별도의 검사 대상으로 분류하는 프로세스 중 적어도 하나를 수행할 수 있다. 또한, 발명의 일 예로서, 미확인 페이로드 데이터에 대응되는 모델 예측값이 별도의 검사 대상으로 분류되는 경우, 후처리 장치(100)의 프로세서(120)는 제1 클래스 정보 및 제2 클래스 정보를 제공하여 기계 학습 모델의 신뢰도의 판단이 가능하도록 지원하거나, 후처리 장치(100)에 연결된 별도의 사용자 단말로 하여금 제1 클래스 정보 및 제2 클래스 정보를 제공하여 기계 학습 모델의 신뢰도의 판단이 가능하도록 지원하는 프로세스 중 적어도 하나를 수행할 수 있다. 그리고, 발명의 또 다른 일 예로서, 미확인 페이로드 데이터에 대응되는 모델 예측값이 별도의 검사 대상으로 분류되는 경우, 후처리 장치(100)의 프로세서(120)는 기계 학습 모델의 재학습이 이루어질 수 있도록 지원(S307)하는 프로세스를 추가로 수행할 수 있다.

[0067] 상술한 바와 같은 과정을 통하여, 기계 학습 모델의 신뢰도를 판단하고 그 결과를 검증할 수 있으며, 이러한 방법은 기계 학습 모델의 결과값이 어떻게 도출되었는지에 대한 근거를 제시할 수 있는 설명가능한 인공지능(eXplainable AI)과 같은 최근의 인공지능 연구 분야에서도 효과적인 방법으로서 활용될 수 있다.

[0068] 이상 설명된 본 발명에 따른 실시예들은 다양한 컴퓨터 구성요소를 통하여 수행될 수 있는 프로그램 명령어의 형태로 구현되어 컴퓨터 판독 가능한 기록 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능한 기록 매체는 프로그램 명령어, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 컴퓨터 판독 가능한 기록 매체에 기록되는 프로그램 명령어는 본 발명을 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 분야의 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능한 기록 매체의 예에는, 하드디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체, CD-ROM, DVD와 같은 광기록 매체, 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magnetooptical media), 및 ROM, RAM, 플래시 메모리 등과 같은 프로그램 명령어를 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령어의 예에는, 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드도 포함된다. 상기 하드웨어 장치는 본 발명에 따른 처리를 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0069] 이상에서 본 발명이 구체적인 구성요소 등과 같은 특정 사항들과 한정된 실시예 및 도면에 의해 설명되었으나, 이는 본 발명의 보다 전반적인 이해를 돕기 위해서 제공된 것일 뿐, 본 발명이 상기 실시예들에 한정되는 것은 아니며, 본 발명이 속하는 기술분야에서 통상적인 지식을 가진 자라면 이러한 기재로부터 다양한 수정 및 변형

을 꾀할 수 있다.

[0070] 따라서, 본 발명의 사상은 상기 설명된 실시예에 국한되어 정해져서는 아니 되며, 후술하는 특허청구범위뿐만 아니라 이 특허청구범위와 균등하게 또는 등가적으로 변형된 모든 것들은 본 발명의 사상의 범주에 속한다고 할 것이다.

부호의 설명

[0071] 10 : 소정의 보안 위협 탐지 시스템

11-1 : 탐지되어 수집된 탐지 로그 데이터

11-2 : 기계 학습 모델의 학습 완료 후 신규로 탐지된 특정 탐지 로그 데이터

12 : 학습 페이로드 데이터 15 : 미확인 페이로드 데이터

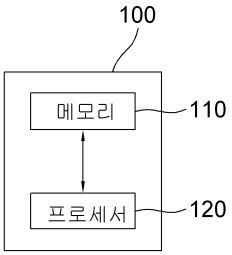
100 : 후처리 장치 110 : 메모리

120 : 프로세서

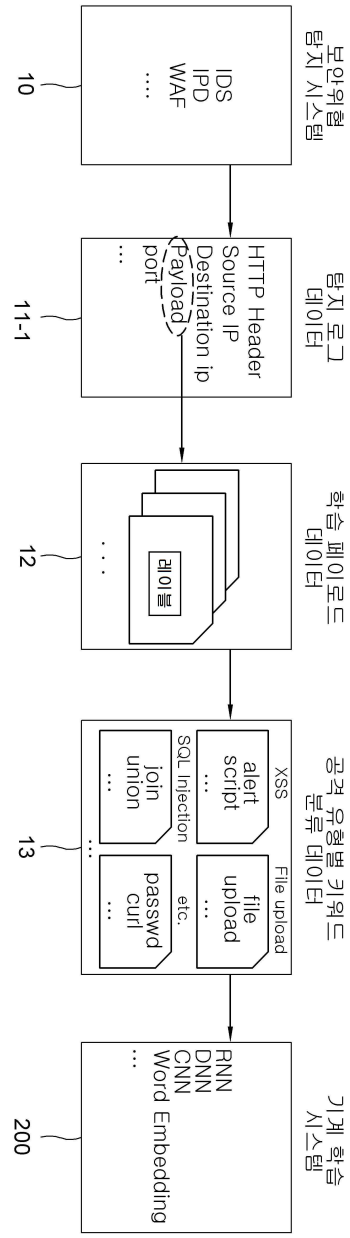
200 : 기계 학습 모델을 포함하는 기계 학습 시스템

도면

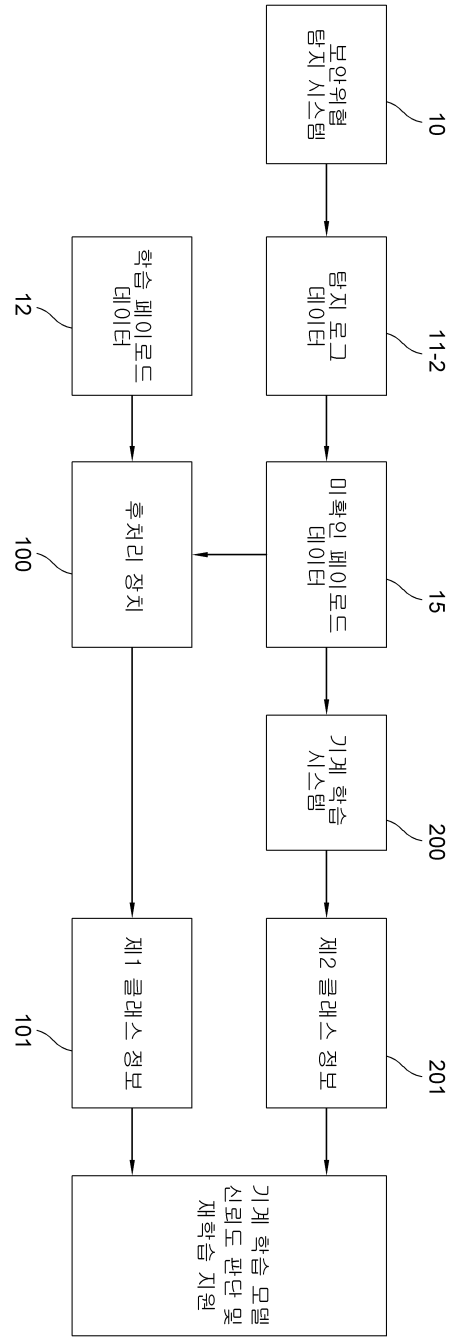
도면1



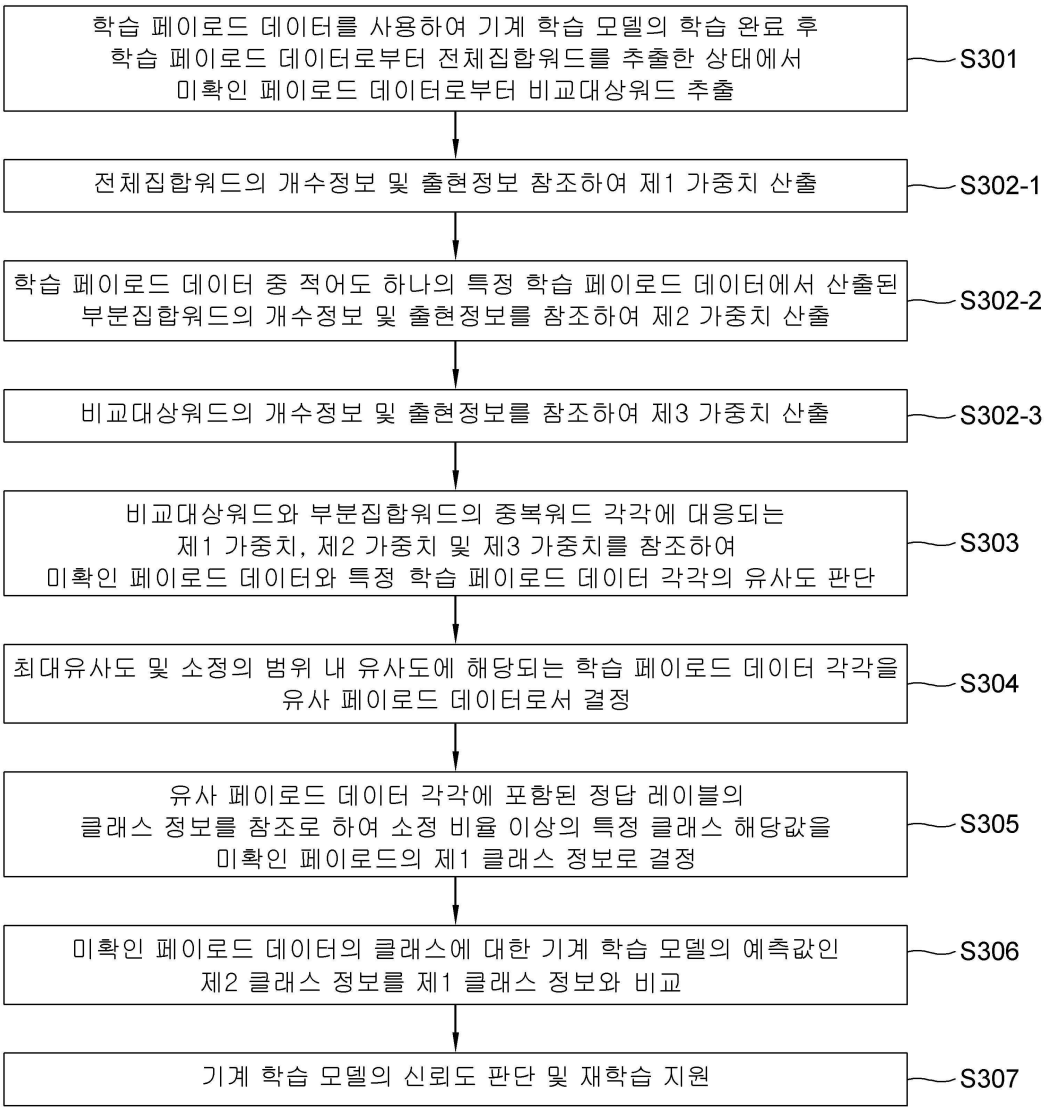
도면2a


보안위협
탐지 시스템
IDS
IPD
WAF
....
10
탐지 로그
데이터
HTTP Header
Source IP
Destination ip
Payload
port
...
11-1
학습 페이로드
데이터
레이블
. . .
12
공격 유형별 키워드
분류 데이터
XSS
alert
script
...
SQL Injection
join
union
...
...
File upload
file
upload
...
etc.
passwd
curl
...
13
기계 학습
시스템
RNN
DNN
CNN
Word Embedding
...
200

도면2b


보안위협
탐지 시스템
10
탐지 로그
데이터
11-2
학습 페이로드
데이터
12
미확인 페이로드
데이터
15
후처리 장치
100
기계 학습
시스템
200
제2 클래스 정보
201
제1 클래스 정보
101
기계 학습 모델
신뢰도 판단 및
재학습 지원

도면3


학습 페이로드 데이터를 사용하여 기계 학습 모델의 학습 완료 후
학습 페이로드 데이터로부터 전체집합워드를 추출한 상태에서
미확인 페이로드 데이터로부터 비교대상워드 추출
S301
전체집합워드의 개수정보 및 출현정보 참조하여 제1 가중치 산출
S302-1
학습 페이로드 데이터 중 적어도 하나의 특정 학습 페이로드 데이터에서 산출된
부분집합워드의 개수정보 및 출현정보를 참조하여 제2 가중치 산출
S302-2
비교대상워드의 개수정보 및 출현정보를 참조하여 제3 가중치 산출
S302-3
비교대상워드와 부분집합워드의 중복워드 각각에 대응되는
제1 가중치, 제2 가중치 및 제3 가중치를 참조하여
미확인 페이로드 데이터와 특정 학습 페이로드 데이터 각각의 유사도 판단
S303
최대유사도 및 소정의 범위 내 유사도에 해당되는 학습 페이로드 데이터 각각을
유사 페이로드 데이터로서 결정
S304
유사 페이로드 데이터 각각에 포함된 정답 레이블의
클래스 정보를 참조로 하여 소정 비율 이상의 특정 클래스 해당값을
미확인 페이로드의 제1 클래스 정보로 결정
S305
미확인 페이로드 데이터의 클래스에 대한 기계 학습 모델의 예측값인
제2 클래스 정보를 제1 클래스 정보와 비교
S306
기계 학습 모델의 신뢰도 판단 및 재학습 지원
S307