Ⅳ. 사이버사고 예측모델 개발

사이버보험 시장 활성화를 위해서는 표준데이터 개발이 필요하다. 이를 위하여 이전 장에서는 빅데이터 기반의 사이버위험 측정체계를 제안하였다. 본 장에서는 제안된 사이버위험 지수를 이용하여 사이버사고(Cyber Incident)를 예측하는 모델을 개발하고자 한다. 이는 향후 정확한 보험금 산출에 기여할 것으로 기대한다.

최근 빅데이터, 인공지능(Artificial Intelligence), 기계학습(Machine Learning)의 급속한 발전으로 미래에 대한 정확한 예측이 가능해졌다. 이러한 발전을 바탕으로, 사이버사고를 예측하는 모델을 제시하고자 한다. 사이버사고에 대한 체계적 데이터가 수집되어있는 미국 자료를 이용하여, 예측모델을 개발하였다. 딥러닝 모델 이용 시 최대 94%의 확률로 기관의 사이버사고 여부를 예측할 수 있음을 확인하였다.

1. 기계학습 개념 및 예측 문제 정의

가. 기계학습의 개념

사이버사고 예측모델 개발을 위하여 기계학습 방법론을 사용하고자 하는데, 독자들의 이해를 돕기 위해서 개념을 먼저 설명하고자 한다. 기계학습에 대한 다양한 정의들이 존재하는데, 공통적으로 “학습데이터(Training Data)로부터 통계적 모델(Model)을 학습하여 새로운 데이터에 대한 결과를 예측(Predict)하는 기술”이라고 정의한다. 기계학습은 예측모델(Predictive Modeling), 데이터마이닝(Data Mining), 예측분석(Predictive Analytics)이라고 불리기도 한다.

기계학습은 다양한 예측 문제에 사용되고 있다. 예를 들어, 수신된 이메일이 스팸인지 여부를 예측하는 스팸필터링(Spam Filtering) 기술, 신용카드 거래의 진위 여부를 예측하는 위조방지(Fraud Detection) 기술, 인터넷쇼핑 사용자가 어떤 광고를 클릭할지 예측하는 타켓광고(Targeted Advertising) 기술 등에 사용되고 있다.

기계학습은 학습데이터 내의 정답 존재 여부에 따라 감독학습(Supervised Learning)과 비감독학습(Unsupervised Learning)으로 구분지을 수 있다.

감독학습에서는 데이터의 속성을 나타내는 특징(Feature)과 정답 레이블(Label)이 필요하다. 예를 들어, 소셜네트워크서비스(SNS) 사진 내의 얼굴인식 문제를 푸는 데에 있어서 SNS 사용자가 직접 입력한 얼굴 위치 및 이름 데이터를 사용하는 경우, 이는 감독학습 문제라고 할 수 있다. 영화 추천시스템(Recommendor System)에서 사용자의 이전 평점을 이용하는 것도 감독학습 문제의 사례이다. 감독학습 모델은, 주어진 레이블 데이터의 속성에 따라 이산값(Discrete Value)의 레이블이 주어지는 분류문제(Classification)와 연속값(Continuous Value)의 레이블로 주어지는 회귀문제(Regression)로 나뉜다. 스팸필터링 문제나 얼굴인식 문제는 분류문제의 사례이다. 반면, 주가예측 문제는 회귀문제라고 할 수 있다.

감독학습 문제를 해결하기 위한 대표적 알고리즘은 K-Nearest Neighbor(KNN), Support Vector Machine(SVM), Decision Tree, Random Forest, Logistic Regression, Linear Regression 등이 있다.

비감독학습에서는 정답 레이블 데이터가 주어지지 않고, 특징들만을 이용하여 유사한 데이터를 클러스터링(Clustering)하거나 전체 데이터의 특성을 요약하는 문제를 해결한다. 비감독학습 알고리즘은 크게 클러스터링과 차원감소(Dimensionality Reduction)로 나뉜다.

비감독학습 문제를 해결하는 대표적 알고리즘은 K-means Clustering, Principal Component Analysis(PCA), Multidimensional Scaling(MDS) 등이 있다.

나. 사이버사고 예측문제 정의

사이버위험 예측문제는 사이버위험 지수를 사용하여 주어진 기관이 사이버사고를 당할지 여부를 예측하는 문제이다. 본 문제의 경우 사이버위험 지수가 특징(Feature) 변수이고, 사고 여부가 이산값의 레이블(Label)이다. 그러므로 본 문제는 감독학습 중 분류문제로 치환할 수 있다.

2. 데이터 수집

가. 사이버사고 데이터 수집

사이버사고 중에서 가장 대표적인 것이 정보유출(Data Breach) 사고이다. 미국 등 여러 국가들은 법률에 의해 정보유출사고를 당한 기관들이 사고정보를 공시하도록 되어있다. 이 때문에 미국기관의 정보유출사고에 데이터는 비교적 잘 정리되어있다. 관련 데이터는 Privacy Clearing House(PCH),[17] VERIS 커뮤니티 데이터베이스(VCDB)[18]에서 수집하였다. PCH 데이터는 미국 보건복지부(U.S. Department of Health and Human Services), Dataloss DB, 외부미디어 등 다양한 데이터소스에서 제공된 8,068건의 정보유출에 대한 상세한 정보를 제공한다. 버라이즌(Verizon)사가 관리하는 VCDB는 여러 데이터소스와 자발적 자료공유로 수집된 8,451건의 정보유출에 관한 정보를 제공한다. 각 사고기록에는 사고날짜, 사고기관, 유출된 기록 건수 등의 정보가 존재한다.

17) https://www.privacyrights.org/data-breaches
18) https://github.com/vz-risk/VCDB

나. 사이버위험 특징추출

기계학습 알고리즘을 적용하기 위해서는 사이버위험 지수 데이터를 이용하여 각 기관들에 대한 특징변수를 추출(Feature Extraction)해야 한다. 이번 연구에서는 CBL과 PSBL 데이터를 기준으로 특징추출을 수행하였다. 향후 다른 사이버위험 데이터 확보 시, 추가적으로 특징변수로 사용할 수 있다. 먼저 두 데이터상에서 각 기관들의 스팸이메일 발생총량(Volume), 발생 건수(Count), 스팸발생 IP 개수(Host) 등 6개의 특징을 추출하였다. CBL에서는 스팸이메일을 발생시킨 봇넷이 포함되는데, 현재까지 400여 개의 봇넷 정보가 수집되었다. 각 기관에서 발생된 스팸발생량을 400여 개의 봇넷별로 분류한 데이터를 특징으로 추출하였다.

다. 사이버사고와 사이버위험 데이터 통합

다음으로는 정보유출 데이터와 사이버위험 지수 데이터를 통합하는 과정이 필요하다. 하지만 두 가지의 데이터는 통합 ID를 사용하지 않으므로, 이종 데이터를 하나로 통합하는 것은 매우 어려운 과정이다. 이번 연구에서는 양 데이터에서 나온 기관의 이름에 스트링매칭(String Matching) 알고리즘을 적용하여 데이터를 매칭하였다. 그 결과 2,102개의 기관에 대한 데이터를 통합할 수 있었다. 통합데이터에는 243건의 정보유출사고가 포함되었다.

이는 전체 정보유출사고 건수에 비해 적은 비율이다. 많은 매칭이 이루어지지 않은 이유는 정확한 피해기관의 이름이 기재되지 않은 경우, 그리고 한 기관이 다양한 이름으로 기재되는 경우가 많기 때문이다. 수작업으로 데이터를 통합하는 방법을 통하여 더 많은 통합데이터를 수집할 수 있을 것이다. 하지만 이는 확장성(Scalability)이 있어서, 사이버보험 시장에 직접 사용되기는 어려울 것이다. 향후 사이버위험, 사이버사고 등의 데이터 수집 및 통합을 용이하게 하기 위해서는 각 데이터에서 통합 ID를 사용하게 하는 방안이 필요하다.

3. 예측 알고리즘 및 모델 평가

이미 언급된 것처럼, 사이버사고 예측문제는 감독 기계학습 문제 중 분류문제로 생각할 수 있는데, 이미 개발된 전통적 기계학습 알고리즘들을 적용하고, 또한 최근 새롭게 개발된 딥러닝(Deep Learning) 알고리즘을 예측문제에 적용하였다.

가. 전통적 분류 알고리즘

분류 알고리즘 중 대표적인 K-Nearest Neighbor(KNN),[19] Logistic Regression(Logit),[20] Decision Tree(DT),[21] Random Forest(RF)[22] 등 4가지 알고리즘을 사용하였다. Python 내의 오픈소스 기계학습 라이브러리인 Scikit-Learn을 사용하였다.[23]

Logit은 독립변수의 선형결합을 이용하여 사건의 발생 가능성을 예측하는 데 사용되는 통계 기법인데, 사건의 종류가 2가지인 경우에 주로 사용된다. 독립변수가 수치자료인 경우에 많이 사용된다.

KNN 알고리즘은 각 데이터를 다차원 특징 공간(Feature Space)상에 배열하였을때, 지역적으로 근접한 데이터들이 같은 성질을 가진다고 가정한다. 데이터 간의 거리를 측정하는 방법으로는 유클리어드 거리(Euclidean Distance) 혹은 코사인거리(Cosine Distance)를 사용한다. 이는 인스턴스 기반 학습이고 가장 간단한 기계학습 알고리즘에 속한다. 특징공간이 초고차원일때, 데이터 간의 거리가 기하급수적으로 멀어진다는 단점을 가지고 있다.

Decision Tree 학습법은 어떤 항목에 대한 관측값과 목표값을 연결시켜주는 예측모델로서, 결정 나무를 사용한다. 의사결정 분석에서 DT는 시각적이고 명시적인 방법으로 의사 결정 과정과 결정된 의사를 보여주는 장점이 있다. 반면, 훈련데이터가 제대

19) https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm
20) https://en.wikipedia.org/wiki/Logistic_regression
21) https://en.wikipedia.org/wiki/Decision_tree
22) https://en.wikipedia.org/wiki/Random_forest
23) http://scikit-learn.org/stable/supervised_learning.html

로 일반화하지 못할 경우 너무 복잡한 결정 나무를 만든다는 한계점을 지니고 있다. 그리고 약간의 훈련데이터 차이에 따라 나무의 모양이 크게 달라질 수 있다.

이러한 문제점을 해결하기 위한 방법이 Random Forest이다. 훈련 과정에서 구성한 다수의 Decision Tree로부터 분류치를 종합하는 앙상블(Ensemble) 학습방법이다.

마지막으로, 벤치마킹을 위해서 임의예측(Random) 알고리즘을 사용하였다. 즉, 학습데이터에서 사고발생 확률을 계산한 후, 개별 기관에 대한 개별적 특징을 고려하지 않고, 확률에 따라 사고가 발생할 것으로 예측하는 방법이다.

나. 딥러닝 알고리즘

전통적인 기계학습 방법은 특징추출의 단계를 거치는데, 이때 해당분야에 대한 전문가의 직관이 필요하다. 예를 들어, 사이버보안 문제가 발생하는 데에 영향을 끼치리라 직관적으로 예상되는 변수들을 추출하게 된다. 하지만 이러한 방법은, 특징변수가 전문가의 직관에 제한되는 한계점을 가진다. 이를 극복하기 위한 방법이 인공신경망(Artificial Neural Network)을 이용하는 딥러닝(Deep Learning) 기술이다. 딥러닝 기술에서는 특징추출 과정 없이 원데이터를 직접 인공신경망에 적용하여 원데이터 간의 복잡한 상호작용을 복구할 수 있다. 다만, Decision Tree와 같이 의사결정 과정을 투명하게 보여주기 어려운 특징(Interpretability Issue)을 가지고 있다. 이에 대한 연구를 지속적으로 진행하고 있다.

필자가 공저한 Shin et al.(2016) 논문 사례에서 볼 수 있듯이, 딥러닝 기술은 사진과 동영상과 같은 비정형 데이터를 분석하는 데 널리 사용되고 있다. 일부 사진분석 문제에 있어서는 인간의 능력보다 뛰어남이 밝혀졌다.[24] 본 연구에서는 딥러닝 알고리즘 중에 하나인 전방전달 신경망(Feed-Forward Neural Network)을 이용하였다.[25]

24) https://www.forbes.com/sites/michaelthomsen/2015/02/19/microsofts-deep-learning-project-outperforms-humans-in-image-recognition/

25) https://en.wikipedia.org/wiki/Feedforward_neural_network

다. 예측모델 평가 방법

머신러닝으로 개발된 예측모델을 평가할 때 교차검증(Cross Validation) 방법이 사용된다. 즉, 수집된 데이터를 학습데이터(Training Data)와 평가데이터(Testing Data)로 나누고, 학습데이터만을 이용하여 예측모델을 학습하고, 학습된 모델을 평가데이터에 적용하여 예측 정확성을 평가하는 방법이다. 본 연구에서는 90%의 데이터를 학습데이터로 10%의 데이터를 평가데이터로 사용하는 10-Fold 교차검증을 사용하였다.

전체 기관들의 데이터를 기준으로 보았을 때, 사이버사고는 비교적 자주 발생하지 않는 사고이다. 그러므로 레이블 데이터에 불균형 문제(Class Imbalance Problem)가 발생한다. 이러한 경우에 공정한 평가를 위해서는 Accuracy를 사용하지 않고, Precision, Recall, F1 score를 사용한다.[26] 아래는 4가지의 예측 정확도 측정 방법이다.

이때, TP는 실제 발생사건을 정확히 예측하는 경우(True Positive)의 건수, FP는 실제 발생하지 않은 사건을 발생한 것으로 잘못 예측하는 경우(False Positive)의 건수, TN은 실제 발생하지 않은 사건을 미발생한 것으로 정확히 예측한 경우(True Negative)의 건수, FN은 실제 발생사건에 대하여 예측하지 못한 경우(False Negative)의 건수를 의미한다.

라. 예측모델 성능비교

〈표 Ⅳ-1〉은 앞서 설명한 6가지 분류 알고리즘에 대한 예측성능을 비교한 결과이다. 딥러닝 방식이 가장 높은 94.73%(F1 기준) 예측률을 기록하였다. 전통적인 기계학습 알고리즘들 중에서는 Decision Tree(91.63%)와 Random Forest(91.92%)가 높은 예측률을 보였다. 반면, KNN과 Logistic Regression은 90% 이하의 예측률을 기록하였다. 벤치마크로 사용한 임의예측(Random) 방식은 59%의 예측률을 보였다.

이 결과를 통해, 딥러닝 방식의 특징 자동추출이 전통적 기계학습 대비 3%의 성능향상을 이룬 것이 확인되었다. 임의예측을 포함한 모든 알고리즘의 90% 이상의

26) https://en.wikipedia.org/wiki/F1_score

Precision 값을 가지지만, Recall 값에서 큰 차이를 보였다. 임의예측은 Recall 값이 45%인 반면, 딥러닝은 Recall 96%를 기록하였다.

〈표 Ⅳ-1〉 예측모델 성능비교

(단위: %)

구분	Precision	Recall	F1 score
Random	91.93	45.97	59.46
KNN	93.66	85.31	88.86
Logistic Regression	95.47	67.3	76.94
Decision Tree	92.88	90.52	91.63
Random Forest	92.94	91	91.92
Deep Learning	96.35	96.21	94.73

4. 소결

본 장에서는 사이버사고의 대표적인 종류인 데이터유출사고 데이터를 이용하여, 사이버사고 예측에 대한 가능성을 타진하였다. 예측모델 개발을 위하여, 본 연구에서 제시된 사이버위험 지수들을 이용하여 기관들의 특징변수들을 추출하였다. 그리고 다양한 기계학습 알고리즘들을 사이버사고/사이버위험 통합데이터에 적용하였다. 특징변수를 연구자가 직접 설계하는 전통적인 기계학습 방법에서는 90% 정도(F1 기준)의 예측률이 달성되었다. 이어 원데이터에서 직접 변수들의 상호작용을 추출하는 딥러닝 알고리즘을 적용하였을때, F1 기준으로 94.73%의 예측률이 달성되었다. 본 결과를 통해, 제시된 사이버위험 지수 데이터가 사이버사고를 예측하는 데 유용하다는 것을 입증한 것이다. 향후 기계학습 기반의 사고예측 모델은 사이버보험 시장에서 보험금 계산에 유용하게 사용될 것으로 기대된다.