http://dx.doi.org/10.14400/JDC.2015.13.9.269

빅데이터 환경에서 기계학습 알고리즘 응용을 통한 보안 성향 분석 기법

최도현*, 박중오**
숭실대학교 컴퓨터학과*, 동양미래대학 정보통신공학과**

Security tendency analysis techniques through machine learning algorithms applications in big data environments

Do-Hyeon Choi*, Jung-Oh Park**
Computer Science, Soongsil University*
Information & Communications, DongYang Mirae University**

요 약 최근 빅데이터 관련 산업 활성화에 따라 글로벌 보안 업체들은 지능적인 보안 위협 모니터링과 예방을 위해 분석 데이터의 범위를 정형/비정형 데이터로 확대하고, 보안 예방을 목적으로 사용자의 성향 분석 기법을 활용하려는 추세이다. 이는 기존 정형 데이터(기존 수치화 가능한 자료)의 분석 결과에서 추론할 수 있는 정보의 범위가 한정적이기 때문이다. 본 논문은 빅데이터 환경에서 기계학습 알고리즘(Naïve Bayes, Decision Tree, K-nearest neighbor, Apriori)을 효율적으로 응용하여 보안 성향(목적 별 항목 분류, 긍정·부정 판단, 핵심 키워드 연관성 분석)을 분석하는데 활용한다. 성능 분석 결과 보안 성향 판단을 위한 보안항목 및 특정 지표를 정형/비정형 데이터에서 추출할 수 있음을 확인하였다.

주제어 : 빅데이터, 머신러닝, 성향분석, 데이터마이닝, 기계학습 알고리즘

Abstract Recently, with the activation of the industry related to the big data, the global security companies have expanded their scopes from structured to unstructured data for the intelligent security threat monitoring and prevention, and they show the trend to utilize the technique of user's tendency analysis for security prevention. This is because the information scope that can be deducted from the existing structured data(Quantify existing available data) analysis is limited. This study is to utilize the analysis of security tendency(Items classified purpose distinction, positive, negative judgment, key analysis of keyword relevance) applying the machine learning algorithm(Naïve Bayes, Decision Tree, K-nearest neighbor, Apriori) in the big data environment. Upon the capability analysis, it was confirmed that the security items and specific indexes for the decision of security tendency could be extracted from structured and unstructured data.

Key Words : Big data, Machine Learning, Sentiment Analysis, Data Mining, Machine Learning Algorithm

Received 20 July 2015, Revised 27 August 2015
Accepted 20 September 2015
Corresponding Author: Jung-Oh Park
(DongYang Mirae University)
Email: jopark13@dongyang.ac.kr

ISSN: 1738-1916

1. 서론

글로벌 보안 업체(IBM, Symantec, Kaspersky, Trend Micro 등)는 최근 신기술의 패러다임 변화에 따라 보안 시장 선점을 위한 경쟁 및 제휴가 진행 중이며 핵심 기술로 보안 인텔리전스(Intelligence)를 앞 다퉈 내세우고 있다[1]. 보안 인텔리전스는 하트블리드(HeartBleed), 셸쇼크(Shellshock) 등 전 세계적으로 영향을 미치는 보안 위협이나 알려지지 않은 취약점에 대응하는데 주요 핵심 보안 키워드로 전체 보안프레임 워크를 의미한다[2,3].

보안 인텔리전스 핵심은 클라우드·빅데이터·모바일을 화두로 지능형 위협(APT)에 대한 대응과 보안위협을 미리 파악하고 선제 대응할 수 있는 것을 목표로 하고 있다[3]. 이와 관련이 높은 분야가 빅데이터를 활용한 데이터 마이닝 분야의 기계학습 연구이다.

보안 분야에서 기계학습 연구는 주로 악성코드를 탐지하는 휴리스틱 탐지 엔진에 활용되어 왔다. 기존의 기계학습을 이용한 악성코드 탐지는 성능 저하와 오탐율(False Positive)에 있어 한계가 존재한다[4]. 특히 보안 분야에서 기존의 기계학습 연구들은 연산 량이 많아 실시간에 적합하지 않고, 학습패턴의 양에 따라 탐지율(Detection rates)이 변하기 때문에 비효율적인 문제가 있다[5]. 최근 기계학습 연구가 많은 발전을 이루었고, 데이터 처리의 속도가 한계를 넘어서면서 보안 분야에서도 기계학습 연구에 대한 관심이 높아지고 있다[6].

본 논문에서는 최근 이슈화중인 보안 인텔리전스와 같은 통합 보안 관리적인 면에서 기존 특정 이벤트나 네트워크 패킷 분석이 아닌 웹(Web)상의 비정형/정형 데이터를 기계학습 알고리즘을 응용하여 분석한다. 최종 목표는 보안 성향 분석을 통해 활용 가능한 데이터를 추출하는 것이 목적이다. 2장은 관련연구, 3장은 제안 기법, 4장은 성능 평가, 5장 결론으로 마친다.

2. 관련연구

본 관련연구에서는 빅데이터 플랫폼에서 활용 가능한 기계학습의 분류와 각 기계학습 알고리즘에 대하여 설명한다.

2.1 기계학습(Machine Learning)

기계학습이란 수집된 다양한 데이터 분석을 할 수 있는 기준(알고리즘)을 가지고 학습을 통해 주어진 일에 대한 해결책 제시를 자동화하는 것을 의미한다[7]. 기존 기계학습과 관련된 기법들은 최근 등장한 빅데이터 기술 분야에서 점점 실현 가능성이 높아지고 있다[8].

기존의 기계학습은 대용량 데이터 처리에 비효율적이었지만 최근 대용량 처리가 가능한 빅데이터 기반 기술의 등장으로 다시 재조명 되고 있다. 기계학습이 효율적으로 동작하기 위해서는 충분한 학습 데이터와 대용량 처리의 지연을 감소시키기 위해서 알고리즘의 개선, 대용량 처리가 가능한 빅데이터 플랫폼 기반의 기계학습 알고리즘이 적용이 요구된다. <Table 1>은 기계학습 별 알고리즘 분류를 나타낸다[9].

〈Table 1〉 Machine learning algorithms (Classification)

Machine learning algorithms		
Supervised	Categorical : (Classification) Naïve-Bayes, K-Nearest Neighbors Logistic Regression, Support Vector Machine, Trees	
	Continuous : (Regression) Decision tree, Random Forests Boosting Trees, Neural Networks Support Vector Regression	
Unsupervised	Continuous	Clustering : K-means
	Categorical	Gaussian Mixture Model

기계학습은 일반적으로 주어진 입력 데이터 기반으로 출력 데이터를 예측하는 교사학습(Supervised), 주어진 입력 데이터가 아닌 비식별 데이터로 출력 데이터를 예측하는 비교사학습(Unsupervised)으로 구분된다[9].

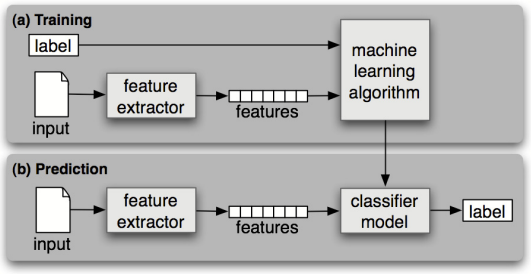
기존 연구들은 교사학습의 분류(Classification), 회귀(Regression)를 중심으로 연구 되었고, 비교사학습의 경우 군집화(Clustering) 등을 중심으로 연구되고 있다.

분류와 회귀의 공통점은 학습에 주어진 데이터에 대한 가정(Best hypothesis)이 요구되며 군집화의 경우 비식별 데이터를 설명해야 할 필요가 있을 때 알고리즘의 최적화에 주로 사용된다.

이외 학습방법으로는 교사학습이나 비교사학습 방법들을 서로 상호 응용된 반지도 학습(Semi-Supervised)이 있다. 반지도 학습은 분류, 회귀, 예측 등 다양한 기법을 함께 사용한다.

2.2 기계학습 알고리즘

교사학습과 비교사학습에서 사용되는 분류, 회귀, 예측 등 각 알고리즘은 보유하고 있는 데이터와 필요한 목적 값의 종류에 따라서 효율적인 알고리즘을 선정해서 사용해야 된다. [Fig. 1]은 기계학습 과정을 나타낸다[10].



[Fig. 1] Machine Learning Process

기계학습 과정은 수집된 데이터(Raw Data)를 이용하여 분류하는 Training 단계와 분류된 데이터를 확인하는 Test 단계로 나뉜다. 데이터의 추출하는 특정 지표(Feature)를 선택하고 추출(Extract)할지 결정해야 하며 핵심 기계학습 알고리즘으로 이를 학습하고 최종결과가 목적의 적정기준에 도달할 때까지 학습과 테스트과정을 반복수행 한다. <Table 2>은 최근 실질적으로 활용되는 기계학습 알고리즘[11,12,13,14,15,16,17,18]의 종류를 나타낸다.

〈Table 2〉 Machine learning algorithms(Types)

Algorithm	Types
Regression	Ordinary Least Squares, Logistic Regression, Stepwise Regression
Clustering	Connectivity-based Clustering, K-means, Centroid-based Clustering, Gaussian
Decision Tree	Classification and Regression Tree (CART), Iterative Dichotomiser 3(ID3)
Association Rule	Apriori Algorithm, FP-Growth Algorithm
Nenural Networks	Perceptron, Restricted Boltzmann Machine(RBM)
Bayesian	Naïve Bayes, Bayesian Belief Network (BBN)
Regularization	Ridge Regression, Least Absolute Shrinkage and Selection Operator (LASSO)
Dimensionality Resuction	Principal Component Analysis(PCA)
Ensemble	Boosting, Bootstrapped Aggregation (Bagging), Ada Boost, Random Fores

3. 제안하는 보안 성향 분석 기법

본 논문에서 사용한 기계학습 알고리즘은 교사학습과 비교사학습 응용된 반지도 학습 형태를 사용 한다. 데이터 분류에 사용된 기계학습 알고리즘은 Naïve Bayes, Decision Tree, Knn(K-nearest neighbor)을 사용하고, 연관성 분석을 위해 Apriori을 사용한다.

3.1 목표 정의

결과로 추출해야 하는 총 목표는 다음과 같이 정의한다. 시기 별 이슈화되는 데이터의 유형 분류와 특정 핵심 키워드 간에 관계도 측정, 데이터의 성향은 긍정·부정 3가지 항목을 결과 목표로 설정한다.

결과 값을 추출하고 추론하기 위해서는 첫째 대용량의 정형/비정형 데이터의 분류가 선행되고, 이후 분류된 각 유형 별 데이터 긍정·부정 성향과 그룹의 관계성을 측정한다.

3.2 데이터 수집 및 활용

데이터 수집 대상 <Table 3>은 RSS와 웹 크롤러를 이용하여 보안, 금융/증권, 정보통신, 컴퓨팅, SW 등 보안 분야와 관련성이 높은 IT 관련 뉴스(2015년 1월 기준) 7073건을 수집하였다.

〈Table 3〉 Data Collection Source

Source		Entry
dailysecu.com(780), boannews.com(2320), ahnlab.com(648), securityworldmag.co.kr(375), ddaily.co.kr(920), etnews.com(1110), www.cctvnews.co.kr(380), itdaily.kr(540)		Title, Registration Date, Original text
Total	7073	

3.3 기계학습 알고리즘 응용

본 논문에서 사용된 각 기계학습 알고리즘의 역할은 다음과 같다.

① Naïve Bayes : 단어, 어휘 추출과 긍정·부정 분석
② Decision Tree : 데이터 항목 분류
③ K-nearest neighbor : 거리계산(데이터 정렬 및 정규화)
④ Apriori : 연관분석(흥미도 분석)

3.3.1 나이브 베이스(Na ve Bayes) 알고리즘

나이브 베이스 알고리즘은 수집된 데이터를 기반으로 보안 관련 단어인지 아닌지 긍정·부정 단어를 포함하는지 조건부 확률을 통해 분류를 수행한다. 분류된 속성은 이후 의사결정 트리, K-최근접 이웃 알고리즘에서 상호 응용된다. 확률 추출을 위한 나이브 베이스 알고리즘은 다음과 같다.

$$p(c_i|w) = \frac{p(w|c_i)p(c_i)}{p(w)}$$

w는 단어 벡터를 의미하며 각각의 분류항목에 대해서 두 개의 확률을 비교한다. 단어 빈도수를 가지고 $p(c_i)$을 계산할 수 있다. i번째 분류 항목을 확인하여 이를 전체 문서수로 나누고, w를 각각의 속성으로 펼치면 확률은 다음과 같다.

$$p(w_0, w_1, w_2 ... w_N|c_i)$$
$$p(w_0|c_i)p(w_1|c_i)p(w_2|c_i)...p(w_N|c_i)$$

문서전체 개수와 분류 항목 개수만큼 반복하여 개수를 세고 조건부 확률을 구하기 위해 이를 전체개수로 나눈다. 수집된 데이터 7073개 데이터에서 단어의 빈도수와 통계를 추출한다. 첫째 긍정·부정 효과를 내는 기술용어를 두 그룹으로 분류한다. 둘째 긍정·부정을 표현하는 한글 단어와 어휘를 추출하여 분류한다.

영어의 경우 일반적으로 SVN(SentiWordNet)와 같은 감정사전을 기반으로 알고리즘을 적용할 수 있지만 한국어 전용 SVN은 초기 연구단계이기 때문에, 한글 단어 및 어휘는 형태소(국립국어원 표준국어대사전)에서 검색결과로 나온 명사, 동사, 관형사·명사, 동사를 참고하여 긍정·부정 형태의 항목을 분류하고 직접 추출하였다. 다음은 데이터의 긍정·부정을 판단하는 방법을 설명한다.

① 혼합 : 긍정·부정 단어가 혼합되어 있는 데이터인 경우 두 유형의 단어 통계치가 현재 데이터에서 60%이상인 것을 선택(사회과학에서는 60%가 옳은 경우 성공적인 것으로 간주)

② 동일 : 긍정·부정 단어 비율이 동일한 경우 제목의 긍정·부정 통계가 비율이 높은 쪽을 판단하여 전체 통계에 가중치 10%(제목은 데이터의 헤드라인을 의미하므로 중요도가 높게 산정)

③ 공통 1 : 제목은 긍정·부정 통계가 비율이 높은 쪽을 판단하여 모든 과정에 통계에 가중치 10%

④ 공통 2 : 긍정·부정 유형의 데이터가 둘 다 존재하지 않는 경우 데이터는 단어추출 대상에서 제외한다. (사실상 없을 가능성이 매우 낮음)

3.3.2 의사결정 트리(Decision Tree) 알고리즘

분류과정으로 의사결정 트리 알고리즘을 사용하여 데이터를 분류 한다. 보안 관련된 키워드(용어)를 포함하고 있는지, 긍정적인지 부정적인지 성향에 따라 특정 결론에 도달하게 된다. [Fig. 2]은 의사결정 트리의 데이터 분류 과정을 나타낸다.

현재 데이터의 제목, 본문에서 보안관련 용어가 존재한다면 사용가능한 데이터로 분류하고, 이후 긍정·부정 용어 발생 횟수를 추출한다. 긍정인 경우 현재 존재하는 보안용어의 유형을 분류하여 보안이슈인지 예방 관련된 데이터인지 사건·사고인지 데이터를 분류한다.

보안관련 용어가 단 하나도 포함되어 있지 않는 경우 관련이 없는 데이터로 정의하고 분석 대상에서 제외한다. 다음은 의사결정 트리의 확률계산과 엔트로피 계산 알고리즘을 나타낸다.

$$L = \log_2 p(x_i) \;\; H = -\sum_{i=1}^{n}(x_i)\log_2 p(x_i)$$

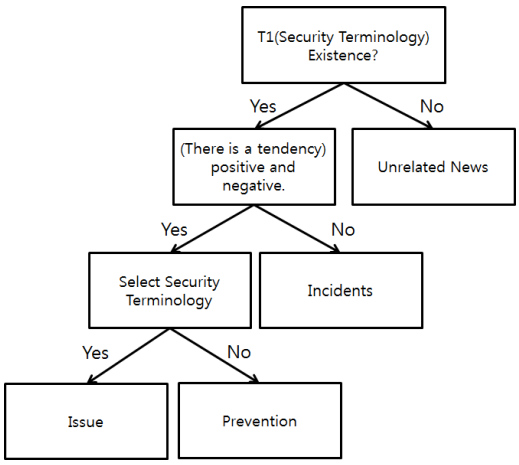
각 분류된 데이터 항목(L)은 발생 빈도에 따라 확률(x_i)을 계산하고, 확률을 이용하여 엔트로피(H)를 계산하여 모두 더한다. 마지막으로 비교하기 전, 후 변화를 비교하여 가장 좋은 속성의 색인을 반환하고, 앞 과정을 반복하여 정보이득(엔트로피)이 가장 높은 속성에 대해서 데이터 분류한다.

3.3.3 K-최근접 이웃(K-nearest neighbor) 알고리즘

의사결정 트리 알고리즘으로 분류된 데이터 항목은 <Table 4> K-최근접 이웃 알고리즘에서 활용된다.

〈Table 4〉 K-nearest neighbor algorithms Data Group

Title	Positive	Negative	Terminology	Type
T1	A	B	C	Issue
T2	A	B	C	Prevention
T3	A	B	C	Incidents
?	A	B	C	-



[Fig. 2] Decision Tree Process

각 분류된 뉴스 데이터는 긍정, 부정, 보안 용어 발생 횟수를 이용하여 거리(D)를 측정한다. 유형이 없는 물음표(알 수 없는 데이터)는 다른 모든 데이터의 거리를 계산하여 거리를 측정한다. 거리 측정을 위한 K-최근접 이웃 알고리즘은 다음과 같다.

$$D= \sqrt{(X_A-X_A)^2+(X_B-X_B)^2+(X_C-X_C)^2}$$

추출되는 값은 정규화를 통해 추출된 값을 유클리드 거리를 사용하여 계산하고, 이중 거리가 가장 짧은 개수를 가지는 변수를 연산 파라미터로 사용한다. 이후 가장 가까운 거리부터 긴 순으로 정렬한다.

3.3.4 어프라이어리(Apriori) 알고리즘

관계도출을 위한 연관분석은 앞에서 분류 알고리즘으로 추출된 분류 항목, 단어, 단어의 빈도수 및 확률 통계, 를 기반으로 Apriori 알고리즘과 응용한다.

연관 규칙(association rules)에서 두 아이템 간의 관계의 강도를 측정하기 위해서 지지도와 신뢰도를 추출한다. 지지도는 데이터 그룹에 특정 데이터가 포함된 데이터 집합의 비율(G), 신뢰도(S)는 연관규칙으로 다음과 같이 정의 한다. 특정 아이템 A에서 B에 대한 신뢰도를 지지도를 이용하여 계산한다.

$$G= \frac{DG_i}{DG_{total}} \quad IF\{A\}\rightarrow\{B\} \; S= G(\{A,B\}/G(\{A\})$$

4. 성능 평가

실험환경은 데이터 처리/분석은 하둡(Hadoop)을 기반으로 파이썬(python) 모듈을 이용하여 알고리즘을 수행한다. 각 기계학습 알고리즘 출력결과는 연계할 수 있는 API나 자동화된 연산 프로그램이 없기 때문에 출력결과를 직접 계산하고 데이터 시각화 부분은 D3(Data-Driven Documents)를 이용하였다.

4.1 입력 데이터 정의

다음은 성능평가를 위해 입력되는 데이터 예를 나타낸다.

① Naïve Bayes : 단어 : “빅데이터, 보안, 해킹, 모바일”, 타입 : “명사”(기준 : 한국어 단어 빈도수)

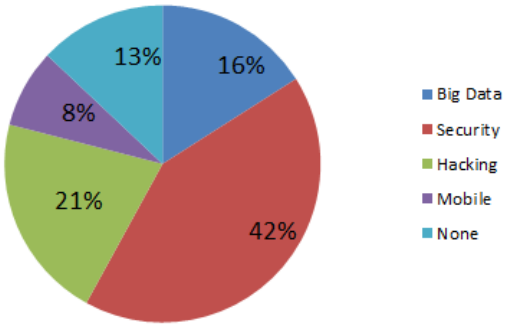
② Decision Tree : 세분화 타입 : 이슈, 예방, 사건·사고, 일반, 관련 없음(기준 : 보안 용어 단어 빈도수)

③ K-nearest neighbor : 비교 대상 : (긍정, 부정, 용어의 비율 및 거리), 정규화 범위 (0.00~1.00)

④ Apriori : 추출된 (한글단어, 보안용어) 그룹의 지지도와 신뢰도

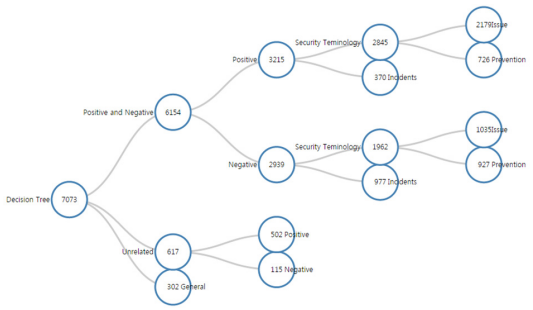
4.2 출력결과

[Fig. 3]은 빅데이터, 보안, 해킹, 모바일 단어의 빈도와 비율을 나타낸다. 단어 별 비율을 보면 보안 키워드 42%로 가장 높게 나타났고 이후 해킹, 빅데이터, 모바일 순으로 비율이 추출되었다. 단어 13%는 보안 관련된 단어를 포함하지 않는 제외영역 부분이다.



[Fig. 3] Frequency and rate of words

[Fig. 4]은 데이터를 세분화하기 위해 각 데이터 유형별 단어 그룹을 분류한 의사 결정트리를 나타낸다.



[Fig. 4] Decision Tree(Type by values)

총 87%의 단어를 기준(13% 제외)으로 첫 번째 과정에서 긍정·부정 유형, 두 번째 과정에서 이슈, 예방, 사건·사고 유형을 변수로 군집화 결과를 추출하였다.

긍정 유형은 52%로 나타났고 이 중 88%가 보안 관련(이슈:67%, 예방:22%), 11%가 사건·사고 데이터로 분류되었다. 부정 유형은 48%로 나타났고 이 중 66%가 보안 관련(이슈:35% 예방:31%), 34%가 사건·사고 데이터로 분류 되었다. 2015년 1월 기준으로 긍정·부정의 비율은 전체 기준(52%, 48%)로 비교적 같은 비율로 나타났다.

세부적으로 긍정적인 성향은 대부분 보안 이슈(67%)에 관련된 데이터로 나타났고, 부정적인 성향에서 사건·사고 비율이 긍정(11%)에서 34% 약 3배 이상 높아 사건·사고에서 부정적인 견해가 높은 것으로 분석되었다.

<Table 5>는 각 데이터 유형 별 긍정·부정 용어와 보안용어를 하나도 포함하지 않는 데이터의 거리(평균)을 나타낸다.

〈Table 5〉 Unrelated data and Positive, Negative Distance(Average)

Title	Positive	Negative	Distance	Type
T1~TN	0.67	0.52	0.78	Issue
T1~TN	0.22	0.31	0.33	Prevention
T1~TN	0.11	0.34	0.33	Incidents
?	0.07	0.01	Prediction	Unrelated

분석 결과 때 긍정·부정이 포함하지 않은 데이터의 유형은 가장 거리가 짧은 예방, 사건·사고 데이터로 예상할 수 있다. 두 가지 유형 목록 중 가장 많이 나타난 유형(다수결)은 사건·사고 데이터로 분석되었다. 이는 나머지 관련 유형이 없는 데이터가 보안 데이터에 포함될 경우 사건·사고 데이터일 가능성이 높다는 것을 예측할 수 있다.

<Table 6>은 키워드 빅데이터, 보안, 해킹, 모바일 단어와 클라우드, 이상징후, FDS, 취약성 단어를 결합하여 관심도를 분석하였다.

〈Table 6〉 Interest results by keyword

Condition	Result	Approval rating	Confidence	Result
Big Data	Cloud	0.01	0.5	12.315
Security	Threat detection	0.01	0.5	7.647
Hacking	FDS	0.01	0.5	1.162
Mobile	Vulnerability	0.01	0.5	2.681

본 논문의 데이터 7073개에서 Condition에서 Result가 포함될 확률이 상대적으로 낮았기 때문에 지지도를 최소 0.01로 설정하고 최소 신뢰도를 0.5로 설정한 결과 빅데이터->클라우드 관심도 결과가 12.315로 가장 높은 것으로 나타났고, 이외 항목들은 보안->이상징후 이외에 결과 값의 신뢰도가 저조한 것으로 나타났다.

결론적으로 보안관련 키워드와 앞서 분석한 분석결과와 관심도 지표를 적용할 경우 2015년 1월 이후 IT 데이터 기준으로 수집된 보안관련 데이터에서는 긍정적 성향으로 보안 이슈에 대한 데이터가 주를 이루었고, 부정적 성향으로 사건·사고 데이터가 주를 이루었으며 주요 키워드로 빅데이터와 클라우드의 관심도가 높게 나타났다는 것을 추론할 수 있다.

5. 결론

최근 기계학습을 효율적으로 활용하려는 다양한 제품들이 상용화 예정 중에 있다. 이는 최근 몇 년간 빅데이터 분석이 각광 받으면서 기계학습에 관심이 높아졌기 때문이다. 본 논문에서는 기계학습에서 활용되는 알고리즘을 응용하고 보안 성향을 분석 하였다. 성향 분석 결과 특정 목적에 대해 추론할 수 있다는 것을 증명할 수 있었다. 빅데이터를 활용한 이러한 사용자의 보안 성향분석은 카드사, 뱅킹, 온라인 마켓 등 금융업종 별 정보보안체계를 강화하는데 활용될 수 있을 것으로 예상한다.

그러나 본 논문의 분석 대상 데이터 7073개는 데이터 분석의 측정 범위가 부족한 것으로 나타났다. 실질적인 결과 활용을 위해서는 보안뿐만 아니라 사용자로부터 수

집된 다양한 데이터 학습 결과가 복합적으로 응용되어야 할 것이다.

수집 데이터는 기계학습을 적용하기 위해 일차적으로 가공되어야 할 필요가 있기 때문에 데이터 준비에 장시간을 들여야 하는 단점이 존재했다. 또한 기계학습을 위해 필요한 학습 데이터는 오랜 기간 동안 축척되어야 할 필요가 있기 때문에 앞으로 많은 데이터 분석이 요구될 것이다.

향후 연구로는 이러한 기계학습에 관한 학습데이터를 연계하고 공유할 수 있는 효율적인 기계학습 알고리즘의 응용에 대한 연구가 필요할 것으로 예상된다.

REFERENCES

[1] TechNavio, Global Threat Intelligence Security Market 2014-2018, TechNavio (Infiniti Research Ltd.), 2014.

[2] Lee-Moongoo, Bae-Chunsock, Next Generation Convergence Security Framework for Advanced Persistent Threat, Journal of The Institute of Electronics Engineers of Korea, Vol. 50, No. 9, pp 92-99, 2013.

[3] Jeon-Deokjo, Park-Donggue, Analysis Model for Prediction of Cyber Threats by Utilizing Big Data Technology, JKIIT, Vol. 12, No. 5, pp. 81-100, 2014.

[4] Chung-Yongwook, Noh-Bongnam, The weight analysis research in developing a similarity classification problem of malicious code based on attributes, Journal of The Korea Institute of Information Security & Cryptology, Vol. 23, No. 3, pp. 501-514, 2013.

[5] Park-Hyeongyu, Situation awareness based intelligent security technology research and development trends, Institute for Information & communications Technology Promotion, p.18, ICT Planning Series Week Technology Trends, 2015.

[6] Im-Sujong, Min-Okgi, Machine Learning Technology Trends for Big Data Processing, Electronics and Telecommunications Research Institute, p.56, Electronics and Telecommunications Trends, 2012.

[7] Mitchell, An Introduction to Genetic Algorithms, p.48, The MIT Press, 1996.

[8] Lee-Jaegu, Lee-Taehoon, Yoon-Sungro, Machine Learning for Big Data analysis, Korean Institute of Communication and Information Sciences, Vol. 31, No. 11, pp 14-26, 2014.

[9] Jang-Byeongtak Next-Generation Machine Learning Technologies, Korean Institute of Information Scientists and Engineers, Vol. 25, No. 3, pp 96-107, 2007.

[10] Steven Bird, Ewan Klein, and Edward Loper, Natural Language Processing with Python, p.201, O'Reilly Media, 2014.

[11] Ethem Alpaydin, Introduction to Machine Learning, second edition, pp 20-32, The MIT Press, 2010.

[12] Mitchell, Tom Michael, The discipline of machine learning, Machine Learning Department technical report, p.6, 2006.

[13] Andrew McCallum, and Kamal Nigam, A comparison of event models for naive bayes text classification, AAAI-98 workshop on learning for text categorization, Vol. 752, pp. 41-48, 1998.

[14] S. B. Kotsiantis, Supervised machine learning: A review of classification techniques, An International Journal of Computing and Informatics, Vol. 31, No. 3, pp. 3-24, 2007.

[15] Blum, Avrim L and Pat Langley. Selection of relevant features and examples in machine learning, Artificial intelligence 97.1, pp. 245-271, 1997.

[16] Dietterich, Thomas G, An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization, Machine learning 40.2, pp. 139-157, 2000.

[17] Zhang, Min-Ling, and Zhi-Hua Zhou, ML-KNN: A lazy learning approach to multi-label learning, Pattern recognition 40.7, pp. 2038-2048, 2007.

[18] Jovanoski, Viktor, and Nada Lavrac, Classification rule learning with APRIORI-C, Springer Berlin Heidelberg, pp. 44-51, 2001.

최 도 현(Choi, Do Hyeon)

- 2008년 2월 : 동서울대학 컴퓨터소프트웨어 공학사
- 2010년 8월 : 숭실대학교 컴퓨터학과 석사
- 2010년 9월 ~ 현재 : 숭실대학교 컴퓨터학과 박사과정
- 관심분야 : 모바일보안, 가상화, PKI
- E-Mail : cdhgod0@ssu.ac.kr

박 중 오(Park, Jung Oh)

- 2000년 7월 : 성결대학교 컴퓨터공학과 졸업
- 2003년 3월 : 명지대학교 전자계산교육 석사
- 2011년 8월 : 숭실대학교 컴퓨터공학 박사
- 2013년 3월 ~ 현재 : 동양미래대학교 조교수
- 관심분야 : PKI, Network security, 암호학
- E-Mail : jopark13@dongyang.ac.kr