머신러닝을 활용한 스마트 서비스와 금융

이 근 영*

* 금융보안원 보안연구부 보안기술팀(e-mail : kylee@fsec.or.kr)

요 약

최근 금융과 IT업계의 중요 화두 가운데 하나는 핀테크이며, 신기술의 발전 등으로 금융서비스의 모습도 나날이 변화하고 있다. 금융관련 데이터는 폭발적으로 증가하고 있으며 이러한 빅데이터 시대에 새로운 가치를 창출할 수 있는 정보 분석을 위한 머신러닝이 각광받고 있다.

머신러닝은 빅데이터 시대에 보다 직관적인 이해를 돕기 위한 시각화(Visualization)를 지원하며, 차원(Vector) 변경 등 여러 방법을 활용하여 데이터 분석을 통해 숨겨진 데이터를 찾아주기도 한다. 머신러닝의 가장 핵심은 기계를 학습시켜 대량의 데이터에 대해 보다 정교하게 분류, 미래의 예측, 진단 및 탐지할 수 있다는 것이다. 머신러닝의 활용분야는 이미 일상생활에서 의사결정의 지원, 자동 검색과 번역 등 삶의 질 향상에 응용되고 있다.

현재 금융권 데이터 분석에 있어 머신러닝의 활용은 아직 활발하지 못한 실정이다. 머신러닝을 통한 데이터 분석 및 활용분야는 대표적으로 영업 및 마케팅 분야이며, 이외에 국내외 일반/금융/보안 분야에서 활용 및 연구가 계속되고 있으며 향후에는 금융권에서도 활발히 확대될 것으로 예상된다. 금융관련 데이터의 수집, 분석 및 활용은 금융기관 내외부의 업무 효율화 이외에 새로운 서비스를 제공할 수 있는 기반이 될 것이다. 나아가 최근 치열해진 금융시장에서 경쟁우위를 선점하기 위한 기술력으로 작용할 것이다.

또한 작년부터 편리하고 안전한 지급결제서비스로부터 시작된 핀테크 열풍으로 이용자 행태분석을 통한 다양한 서비스 개발 등을 통해 머신러닝 기술은 빠르게 진화하고 이를 활용하여 스마트 서비스들이 더욱 발전할 것으로 예상된다.

빅데이터 시대에 금융의 스마트 서비스를 가능하게 해주고 금융 리스크 관리 능력 제고, 보안 기술에 활용 등 해당 기술의 효과를 높이기 위해서는 금융권에서도 머신러닝을 통한 기술 및 시스템 개발 등 많은 투자와 관심이 필요하다.

머신러닝은 금융회사를 비롯하여 기업의 시장 지위에 변화를 줄 잠재력을 갖출 무기로 동작할 것이라고 감히 예측할 수 있을 것이다.

Ⅰ. 서 론

1959년 “머신러닝(Machine Learning, ML)”이라는 용어가 처음 문헌에 등장[1]한 것을 시작으로 1980년대 머신러닝이 이론적 틀을 형성하고 새로운 학문 및 기술 분야로 정립하는 시기를 거쳐, 1990년대 이후 데이터 마이닝 산업의 등장으로 학문적 경계가 허물어지고 새로운 산업기술로써 재조명 받고 있다.

최근 IT업계의 화두로 떠오른 머신러닝은 인터넷 검색 엔진, 스팸 메일 필터링, 음악, 책, 영화 등의 콘텐츠 추천 시스템과 같은 영역에서 이미 널리 활용되고 있으며, 빅데이터 · 클라우드 컴퓨팅 시대에 맞춰 가치 창출을 위한 연구가 진행 중이다.

금융권에서도 머신러닝의 활용으로는 신용평가모델 개발, 신용카드사에서도 사기를 방지하기 위해 머신러닝을 도입하는 등 데이터의 숨겨진 의미를 파악하고 미래를 예측해야하는 곳에 머신러닝이 활용되고 있다.

특히 금융권에서 빅데이터 산업과 법의 준수의 조화로운 방향이 모색되다가 최근 금융위의 ‘금융권 빅데이터 활성화 방안’에서 신용정보 범위 명확화, 비식별정보 활용 가능여부 명확화를 통해 법령상 제약요건이 어느정도 해소되고 있다. 하지만 금융권에서 개인 및 신용정보는 법률상 해석으로 모두 해결될 수는 없으며 끊임없는 모니터링과 조화로운 해결방안은 계속해서 모색되어야 한다.

따라서 본 연구보고서에서는 “머신러닝을 활용한 스마트 서비스와 금융”에 대해 다음과 같은 순서로 살펴보고자 한다.

우선 머신러닝의 개념을 이해하기 위해서 통계학, 데이터 마이닝, 인공지능 등 다른 분야와의 연계성을 파악하고, 구체적으로 머신러닝의 알고리즘과 각 특징들을 살펴봄으로써 머신러닝의 이론적인 이해도를 돕도록 한다. 이러한 기술적인 이해를 바탕으로 일반적인 머신러닝의 활용분야와 금융/보안 분야까지 활용사례들을 알아보고 최근 빅데이터 시대에 머신러닝 관련 산업 및 기술 동향을 살펴보고 법적 이슈들을 도출한다.

마지막으로 앞서 제기한 이슈들을 고려하여 머신러닝을 통한 금융권 스마트 서비스에서 금융회사가 해결해야할 과제들을 제시하기로 한다.

1) Arthur Samuel의 논문(“Some Studies in Machine Learning Using the Game of Checkers”, IBM Journal of Research and Development, vol. 3(3), pp. 210-219, July 1959)에서 어떤 상황의 정량적 평가를 구하는 평가함수와 이에 관련 파라미터 조정에 기초한 최초의 기계학습을 연구

II. 머신러닝 개요

1. 머신러닝의 개념

머신러닝(기계학습이라고도 한다.)은 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야로써, "환경과의 상호작용에 기반한 경험적인 데이터로부터 스스로 성능을 향상시키는 시스템을 연구하는 과학과 기술"[2]로 정의될 수 있을 것이다.

머신러닝은 미리 결정된 특정 모델을 데이터와 비교하여 테스트하는 것이 아니라, 데이터에서 학습하도록 설계되었다. 경험으로부터 학습할 수 있는 능력은 사람을 포함한 지능적인 시스템의 가장 근본적인 특성 중 하나이며, 초기의 머신러닝은 인공지능(Artificial Intelligence)[3]의 '학습'에 관한 부분을 구체화한 기술로 기계도 인간처럼 학습시키고자 하는 지적탐구에서 시작되었다. 가장 대표적인 예로, 머신러닝을 통해서 수신한 이메일이 스팸인지 아닌지를 구분[4]할 수 있도록 훈련할 수 있다.

최근의 머신러닝 개념은 빅 데이터(Big Data), 클라우드 컴퓨팅(cloud computing) 등의 환경을 포함하여 이해하여야 한다. 나아가 머신러닝은 다양한 확률, 조합 이론과 수학적 최적화 기법, 통계, 알고리즘, 컴퓨터 구조를 활용하여 이상적인 학습 및 예측모델을 구축하는 기술로 연구자의 경험적 지식 습득과 그 응용방법까지 포함하는 융합기술로 발전하고 있다. 즉, 시대의 흐름에 따라 머신러닝의 개념이 재해석 되고 있다.

---

2) 이 정의에서 주목해야 할 것은 학습 시스템이 "환경", "데이터", "성능"의 요소를 가지고 있다는 것이다. "환경"은 학습 시스템이 독립적으로 존재하지 않고 상호작용하는 대상이 있다는 것이며 상호작용의 방법에 따라서 경험하는 "데이터"의 형태가 다르다. 학습 시스템은 또한 문제해결을 수행하며 이 수행의 "성능"이 시간이 감에 따라 향상된다.
장병탁, 차세대 기계학습 기술, 정보과학회지 제25권 제3호, 96쪽, 2007.3.

3) 인공지능(人工知能)은 철학적으로 인간성이나 지성을 갖춘 존재, 혹은 시스템에 의해 만들어진 지능을 뜻한다. 일반적으로 범용 컴퓨터에 적용한다고 가정하며, 이 용어는 또한 그와 같은 지능을 만들 수 있는 방법론이나 실현 가능성 등을 연구하는 과학 분야를 지칭하기도 한다.(출처:위키백과)

4) 스팸 필터링의 기본 알고리즘은 베이즈의 정리(Bayes'Theorem)에 기초하며, 조건부 확률을 이용하여 사전 확률과 사후 확률의 관계를 추정하고 새로운 정보에 대하여 사후 확률의 변동을 예측하는 방법을 사용한다. 베이즈 정리는 수학적으로 식으로 다음과 같이 표현될 수 있다.

A와 B가 사건(event)일 경우, $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$ 이다.

P(A)와 P(B)는 다른 것에 관해서 관계없는 사전 확률이다.

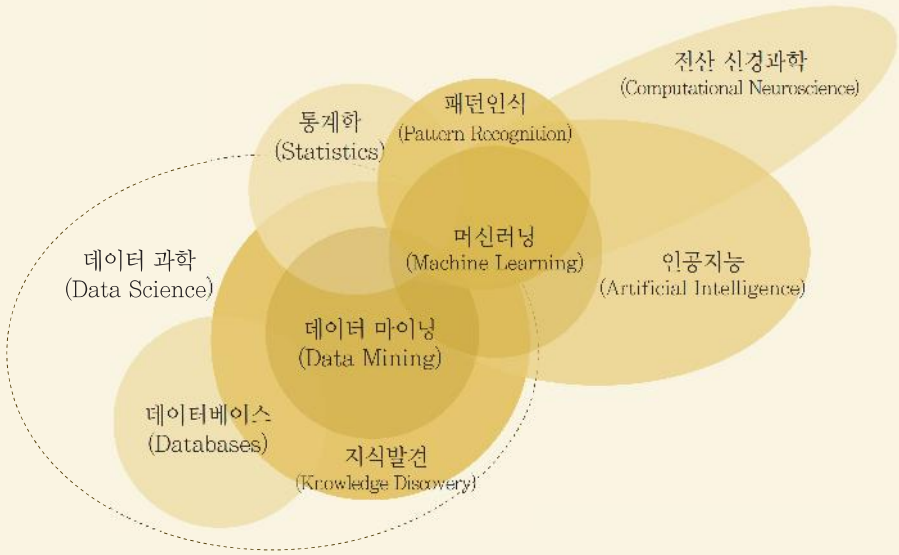
$P(A|B)$는 조건부 확률로 B가 주어진 경우 A의 확률은 참(True)이다.

$P(A|B)$는 A가 주어졌을 때 B의 조건부 확률이다.

2. 머신러닝의 다른 학문분야와의 연계성

오늘날의 머신러닝은 그 역사와 더불어 빅데이터(Big Data)[5], 클라우드(Cloud)[6], 사물인터넷(IoT)[7] 기술 등의 환경과 복합적으로 상호작용[8]하여 여러 학문분야와의 연계성을 가지게 되었고, 통계학(Statistics), 데이터 마이닝(Data Mining), 데이터 과학(Data Science) 등 다양한 영역에 걸쳐있다([그림 1] 참조). 그러나 머신러닝은 인공지능, 패턴 인식 등으로 일반화될 수 없고 기술 간의 연계성 및 차이는 구별될 수 있어야 한다.

[그림 1] 머신러닝의 여러 학문 분야와의 연계성



자료 : SAS Institute Inc., An Overview of Machine Learning with SAS® Enterprise Miner™

5) 데이터를 수집, 저장, 처리, 분석하는 것 뿐만 아니라 이로부터 새로운 가치를 창출하는 전 과정을 포괄한다. 데이터의 특성에 기반하여 V속성(양-Volume, 속도-Velocity, 다양성-Variety, 가치-Value 등)의 개념을 가진다.

6) 클라우드 컴퓨팅(클라우드)은 애플리케이션부터 데이터까지 모든 컴퓨팅 자원을 인터넷 환경에서 원하는 만큼 사용하는 인터넷 기반(cloud)의 컴퓨팅(computing) 기술을 의미한다.

7) 사물 인터넷(Internet of Things, IoT)은 각종 사물(가전제품, 모바일 장비, 웨어러블 컴퓨터 등 다양한 임베디드 시스템)에 센서와 통신 기능을 내장하여 인터넷에 연결하는 기술을 의미한다.

8) 런던 지하철 역사와 지하철 철로에 있는 센서에서 데이터를 취합해 클라우드로 보내고 머신러닝을 돌려서 부품 교체 수명이나 열차 안 온도 등을 예측하는데 활용하고 있다고 한다. 예측 정보는 즉각 역무원 및 직원들에게 모바일로 전송되어 이들이 적절한 조치를 취하게 된다. 런던지하철 사례와 마찬가지로 "사물인터넷(IoT), 빅데이터, 클라우드, 머신러닝, 모바일이 모두 결합된 서비스가 나올 것"이라고 전망했다.
ZDNet Korea, 머신러닝이 몰고 올 IT진화 시나리오, 2014.12.14

가. 머신러닝과 통계학

일반적으로 머신러닝과 통계학과의 연계성은 거의 없어 보일 수 있고, 대부분의 사람들에게 통계학은 자신 회사의 제품이 얼마나 좋은지를 알아보기 위해 사용되는 소수만이 알고 있는 주제[9]에 불과한 것으로만 취급될 수 있다.

머신러닝은 데이터를 정보로 변환해야하며 인간의 참여를 최소화하는 방법론의 개발이 필요하다. 하지만 다차원의 거대한 자료의 출현은 새로운 유형의 자료에 적합한 알고리즘의 개발을 어렵게 하였고 문제를 해결하기 위해 과학적 방법을 적용하여야 했다. 이러한 맥락에서 통계학적 사고는 머신러닝 분야에서 여러 가지 알고리즘들의 원리에 대한 새로운 인식방법으로써 중요한 사고의 도구로 사용되고 있다.

머신러닝 분야 중에서 통계학이 가장 활발하게 적용되고 있는 분야는 교사학습 분야[10]이며, 교사학습 방법론 중 SVM(Support Vector Machine)[11]과 부스팅(Boosting)[12] 알고리즘의 개발은 많은 실증적 연구를 통하여 예측력 측면에서 기존의 머신러닝 방법론을 질적으로 향상시켰음이 밝혀졌다. 실증적 연구 이후에 이 두 개의 알고리즘이 왜 예측력을 급격하게 향상시켰는가에 대한 연구가 시작되었으며, 이 연구에 통계학자들이 많은 기여를 하고 있다.[13]

나. 머신러닝과 데이터 마이닝 그리고 지식발견

데이터 마이닝(Data Mining)이란 대규모로 저장된 데이터 안에서 체계적이고 자동적으로 의미있는 데이터(정보, 지식, 규칙, 패턴, 특성 등)를 추출, 분석하는 과정이며([그림 2]), 머신러닝은 이러한 데이터를 자동으로 추출 및 분석하는 기술로 활용된다.

데이터 마이닝은 데이터베이스 안의 지식 발견(Knowledge-Discovery in Databases, KDD)이라고도 일컬어지며, '지식발견 및 데이터마이닝 국제학술대회(1955년)' 등에서 다양하게 그 개념이 제시되고 있다.

9) 더럴 허프(Darrell Huff)의 '통계로 거짓말하는 방법(How to lie with statistics)'은 1954년 저서이며, 통계로 사기치는 방법을 알려주는 통계학 분야의 역대 베스트 셀러 중 하나이다. 저자는 해당 저서에서 통계전문가들이 즐겨 사용하는 모든 형태의 통계를 제시하여, 표본 연구, 도표화, 인터뷰 기법, 숫자로부터 결론을 추출하는 방법 등을 분석했다.

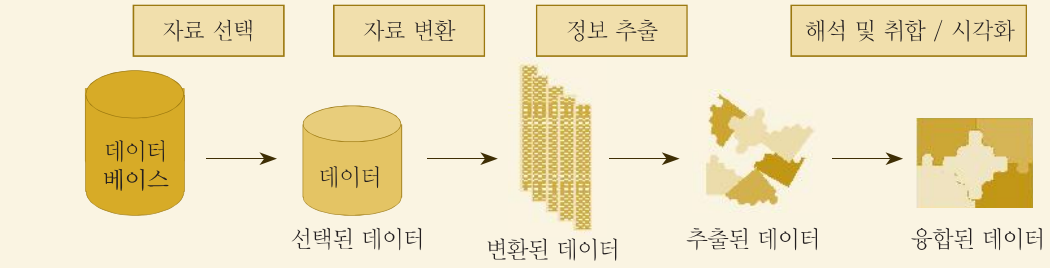
10) 자세한 내용은 Ⅲ. 머신러닝 관련 기술 및 특징에서 후술한다.

11) SVM의 특징은 주어진 자료들의 마진(margin, 주어진 자료가 분류경계에서 떨어진 거리)의 최소값을 최대로 하는 분류경계(decision boundary)를 최적분류모형을 정의하는 방법이다.

12) 기본 아이디어는 여러 개의 나쁘지 않은 분류 모형을 결합하여 아주 좋은 분류 모형을 만드는 것이며, Adaboost 알고리즘은 여러 개의 분류모형을 만들기 위하여 연속적으로 자료의 가중치를 조절한다.

13) 김용대, 기계학습과 통계학, 정보과학회지 제25권 제3호, 90쪽, 2007.3.

[그림 2] 데이터 마이닝 과정



자료 : 김진형, 한국과학기술정보연구원(KORDIC), 정보추출기술(Data Mining Techniques)

데이터 마이닝의 기법은 ①발견할 지식의 종류에 따라서 분류(Classification), 요약(Summarization), 군집화(Clustering) 등 ②탐사할 데이터베이스의 종류에 따라 관계형(Relational) DB, 객체지향(Object-Oriented) DB 등 ③탐사 기법에 따라서 기호처리식 인공지능적 방법론, 신경망적 방법 등이 있다.

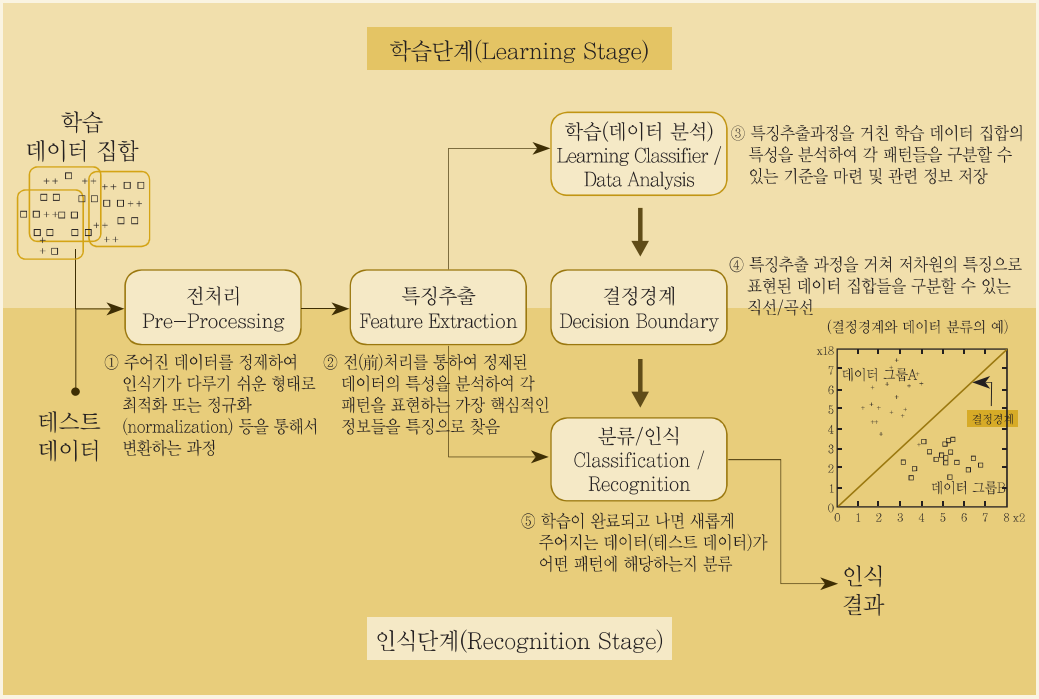
다. 머신러닝과 패턴인식

패턴인식(Pattern Recognition)이란 주어진 데이터의 집합에 대해 입력 값을 바탕으로 특정 기준에 따라 여러 개의 그룹으로 분류(인식)하는 것을 말한다.

예를 들어 숫자인식, 얼굴인식 등의 문제를 해결하기 위해 각 데이터의 구조적 특성에 따라서 패턴을 분류해야하는데 얼굴인식의 경우 각 사람들의 얼굴 특징을 일일이 분석하여 정의하고 특징 자체를 정의하는 것은 매우 힘든 일이다. 이러한 데이터의 구조적 특징에 의해 패턴의 정의 및 인식 방법과 템플릿 매칭 방법이 가장 기본적인 패턴인식 기법이라고 할 수 있다. 하지만 실세계에서는 패턴의 다양한 변형이 존재하고 패턴인식 문제의 핵심은 이러한 변형을 효과적으로 표현하고 구분하는 보다 정교한 방법을 설계하는데 있다고 볼 수 있다. 이러한 패턴의 변형에 따른 문제를 해결하기 위해 머신러닝 분야의 다양한 방법론들이 적용될 수 있다.

머신러닝 기법을 사용하는 패턴인식에는 크게 두 가지인 학습단계와 인식단계가 존재하며, 패턴인식의 전체적인 처리과정은 [그림 3]과 같다.

[그림 3] 패턴인식 처리과정



자료 : 박혜영, 이관용, "패턴인식과 기계학습" 재구성

먼저 학습단계에서는 주어지는 데이터 집합(학습 데이터)을 이용하여 패턴의 특성을 분석하고 서로 다른 패턴들을 구분하기 위한 핵심정보를 추출한다.

학습이 완료되고 나면 인식단계에서는 새롭게 주어지는 데이터[14]가 어떤 패턴에 해당하는지 분류하고 인식하는 단계가 수행된다. 인식단계에서는 먼저 전처리와 특징추출과정이 학습단계와 동일하게 수행되고 추출된 특징에 대하여 학습된 분류기를 이용한 인식(분류)과정을 통해 최종 인식 결과를 얻게 된다.

라. 머신러닝과 딥러닝 그리고 인공지능

우리는 어떤 지식을 다양한 경험과 데이터를 통한 학습과정으로 축적하는 경우가 많으며, 이런 문제를 접근하는 것이 '머신러닝'이다.

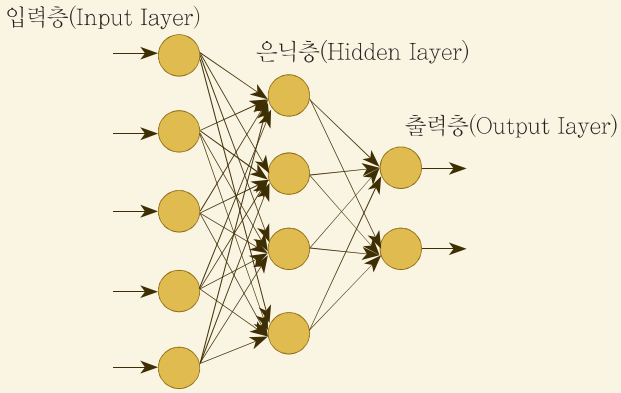
이러한 학습을 위한 또 다른 접근방식으로 '인공신경망(Artificial Neural Networks,

14) 주로 테스트 데이터(Test Data)라고 불린다.

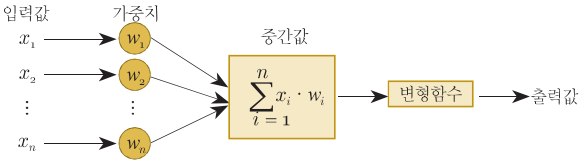
ANN)[15]'이라는 학습방식이 연구, 발전되어 오고 있다. 인공신경망의 발상은 인간의 신경망 구조를 모방하면서 입력으로 들어오는 정보를 각각의 입력 노드에 배정하고, 이들을 주어진 함수를 표현하는 네트워크로 출력 노드를 통해 결과를 얻어낸다는 아이디어에서 출발했다. 이러한 신경망을 학습의 주요 방식으로 사용하는 것을 '딥러닝(Deep Learning)'이라고 부른다.

딥러닝은 ①입력층(Input layer) ②출력층(Output layer) ③은닉층(Hidden layer)을 갖고 있는 계층 구조(Layer Structure)로 구성되며, 심층 신경망(DNN, Deep Neural Network)을 지칭하는 머신러닝의 모델 또는 알고리즘의 집합이다. 큰 틀에서는 사람의 사고방식을 컴퓨터에게 가르치는 인공지능의 한 분야라고 할 수 있으며, 음성인식, 이미지 분류 등 패턴인식 분야에서 높은 성능을 보인다.

[그림 4] **층상구조로 연결된 신경망의 예**



15) 인공신경망은 X_1부터 Xn까지 입력 값에 각각 W_1부터 Wn까지의 가중치를 곱하고 그 모든 합이 변형함수를 통해 임계치가 초과되면 출력값이 발생하게 된다.



Ⅲ. 머신러닝 관련 기술 및 특징

1. 머신러닝의 분류

머신러닝의 다양한 기법과 알고리즘들에 대해 명확하게 구분하는 것은 상당히 어려우며, 입력데이터와 적용환경, 학습방법, 요구되는 결과로 세분화되어 구분될 수 있다.[16]

일반적으로 머신러닝 알고리즘은 학습의 방법에 따라 ①교사 학습(Supervised Learning) ②비교사 학습(Unsupervised Learning) ③반교사 학습(Semi-Supervised Learning)으로 구분된다.[17]

①교사 학습(Supervised Learning)은 학습 시에 인식기에서 출력해야할 결과 값을 미리 알려주는 "교사(supervised)"가 존재하는 형태이며, 학습 시 인식기의 원하는 출력 값에 대한 정보 없이 학습이 이루어지는 형태를 ②비교사 학습(Unsupervised Learning)이라고 한다. ③반교사 학습(Semi-Supervised Learning)은 목표 값이 표시된 데이터와 표시되지 않은 데이터를 모두 훈련에 사용하는 것으로 교사 학습과 비교사 학습 사이에 위치한다.

[표 1] 교사 학습, 비교사 학습, 반교사 학습의 차이

구분	내용
교사 학습	학습 Training: 학습 데이터 Training Data → 특징 추출 Feature Extraction → 머신러닝 알고리즘 Machin Learning Algorithm; 목표 결과값 Lavels 예측 Predict: 새로운 데이터 Test Data → 특징 추출 Feature Extraction → 예측 모델 Predictive Model → 예측 Expected Lavels
비교사 학습	학습 Training: 학습 데이터 Training Data → 특징 추출 Feature Extraction → 머신러닝 알고리즘 Machin Learning Algorithm 분석 Analysis: 새로운 데이터 Test Data → 특징 추출 Feature Extraction → 예측 모델 Predictive Model → 가능성, 나은 표현 Likelihood or Better Representation
반교사 학습[18]	목표 결과값 Lavels → 학습 데이터 Training Data → 교사 학습 Supervised Learning; 비교사 학습 Unsupervised Learning; 데이터 Data 교사 학습(Supervised Learning) 비교사 학습을 교사 학습에 추가 자가 학습(Self-Training) 반교사 학습 Semi-Supervised Learning

16) 이재구 외 2명, Big Data 분석을 위한 Machine Learning, 한국통신학회지, 제31권 제11호, 15쪽, 2014.10

17) O. Chapelle et al., Semi-supervised learning, vol. 2,MIT press Cambridge, 2006.

18) Stefan Uhlmann, Semi-Supervised Learning for Ill-Posed Polarimetric SAR Classification, remote sensing, 2014.6

2. 머신러닝 알고리즘과 특징

머신러닝 알고리즘별 특징에 대한 이해는 주어진 데이터 분석에 최적화된 알고리즘 적용 및 데이터를 활용한 비즈니스 문제에 대한 적절한 답을 찾는데 도움을 준다. 국제 데이터 마이닝 컨퍼런스(IEEE International Conference on Data Mining, ICDM)[19]에서는 가장 영향력 있는 알고리즘을 식별하기 위한 노력의 일환으로 '데이터 마이닝을 위한 알고리즘 TOP 10'[20] 이 발표되기도 하였다.

학습의 방법에 따른 주요 머신러닝 알고리즘별 특징은 [표 2]와 같다.

[표 2] **머신러닝 알고리즘 분류와 특징**

구분	내용		
교사 학습	특징	· 주로 인식, 분류, 진단, 예측, 회귀분석 등의 문제 해결에 적합 · 학습모델은 정답으로 알려진 라벨에 의한 수정 과정을 통해 일정 수준의 정확도를 얻을 때까지 진행 · 비교사 학습 방법에 비해 성능은 좋으나 원하는 결과를 데이터에 포함하기 위한 시간과 구축 비용이 증가	
	예시	회귀분석 (Regression Analysis)	· 주어진 데이터와 선택된 학습 모델에 의해 얻어진 예측값 간의 오차를 최소화하기 위한 반복적인 과정을 수행하면서 데이터들간의 관계를 모델링 · 주요 알고리즘 : Ordinary Least Squares, Logistic Regression, Ridge Regression 등
		의사결정나무 (Decision Tree)	· 데이터의 속성(Feature)에 따라 나무형태의 의사결정 학습모델을 만들고, 반복을 통해 주어진 문제에 대한 최종 결정을 도출 · 주요 알고리즘 : Gradient Boosting, Random Forest 등
		인공신경망 (Artificial Neural Networks)	· 생물의 신경망 구조와 기능을 모방한 알고리즘 · 입력층(Input Layer), 중간 연결층(Hidden Layer), 결과 출력층(Output Layer)의 구조로 각 Layer의 노드들을 상호 연결하는 가중치를 갱신함으로써 결과를 출력 · 주요 알고리즘 : Perceptron, Restricted Boltzman Machine(RBM) 등
		베이지안 방법 (Bayesian Methods)	· 군집과 예측 문제를 풀기 위해 특성들 사이의 독립을 가정하는 베이즈 정리를 확장, 적용한 알고리즘 · 주요 알고리즘 : Naive Bayes, Bayesian Belief Network (BBN) 등
		서포트 벡터 머신 (support vector machine, SVM)	· 최적의 결정 경계를 찾기 위해 마진*을 이용하여 학습의 목적 함수를 정의 * 학습데이터들 중에서 결정 경계에 가장 가까운 데이터로부터 결정경계까지의 거리

19) http://www.cs.uvm.edu/~icdm/

20) ICDM 컨퍼런스에서 2006년 12월에 발표되었으며, TOP 10 알고리즘에는 ① C4.5 ② 2. k-Means ③ Support Vector Machines(SVM) ④ Apriori ⑤ Expectation Maximization(EM) ⑥ PageRank ⑦ AdaBoost ⑧ k-Nearest Neighbors(kNN) ⑨ Naive Bayes ⑩ Classification and Regression Tree(CART)가 있다.
이후, 2008년 1월「Knowledge and Information Systems」저널(vol 14, issue1, pp 1-37)에 "Top 10 Algorithms in Data Mining(Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, Dan Steinberg)"라는 제목으로 수록되었다.

<table>
<tr><td rowspan="4">비교사
학습</td><td>특징</td><td colspan="2">· 군집화(비슷한 관측치끼리 군집하는 작업), 밀도추정, 차원축소(데이터간의 연관규칙을 찾음), 특징 추출 등이 필요한 문제에 적합
· 교사 학습에 비해 성능은 좋지 않으나, 원하는 결과가 표현되지 않은 학습 데이터를 이용하기 때문에 학습 데이터 구축이 용이</td></tr>
<tr><td rowspan="2">예시</td><td>군집화
(Clustering)</td><td>· 주어진 데이터간의 유사성을 최대로 하는 군집 생성을 통해 데이터를 분류하는 방법
· 주요 알고리즘 : Connectivity-based Clustering(데이터 객체간의 거리 근접 특성을 활용), K-means Clustering(중점 벡터 이용), Centroid-based Clustering, Distribution based Clustering(Gaussian통계 분포 모델 이용), Density-based Clustering(데이터 밀도간 차이 활용) 등</td></tr>
<tr><td>차원축소
(Dimensionality Reduction)</td><td>· 군집화와 유사하게 데이터의 고유 구조를 찾지만, 차원축소의 목적은 원 정보보다 낮은 차원의 데이터로 요약하거나 표현
· 주요 알고리즘 : 주성분 분석(Principal Component Analysis, PCA) 등</td></tr>
<tr><td rowspan="4">반교사
학습</td><td>특징</td><td colspan="2">· 라벨(목표 결과값)이 없는 대용량 데이터에 적은 수의 라벨이 있는 데이터가 포함된 입력데이터로부터 예측을 요구하는 빅데이터 분석에 주로 적용
· 예측을 목적으로 한 교사학습 특성과 학습모델이 라벨 없는 데이터간의 관계 구조도를 알아야 하는 비교사학습의 특징을 함께 가짐</td></tr>
<tr><td rowspan="3">예시</td><td>예측과 분류
(Prediction and Regression)</td><td rowspan="2">· 반교사학습, 교사학습(예측과 분류) 알고리즘은 종종 군집화(클러스터링)와 함께 결합됨
※ 본 표 상기의 교사학습의 알고리즘들과 비교사 학습의 군집화(클러스터링) 내용 참조</td></tr>
<tr><td>군집화
(Clustering)</td></tr>
<tr><td>기대값 최대화
(expectation-maximization, EM)</td><td>· 관측되지 않는 잠재변수에 의존하는 확률 모델에서 최대가능(maximum likelihood)나 최대사후확률(maximum a posteriori, MAP)을 갖는 매개변수를 찾는 반복적인 알고리즘
· 기대값(E)단계와 기대값을 최대화하는 변수값을 구하는 최대화(M)단계를 번갈아가면서 적용하며, 최대화 단계에서 계산한 변수값은 다음 기대값 단계의 추정값으로 사용</td></tr>
</table>

Ⅳ. 머신러닝의 활용

1. 머신러닝의 활용분야

머신러닝의 이제 일상생활에까지 깊숙이 파고들었다. 최근 에어비앤비[21]는 에어로솔브[22]를 활용해 '프라이스팁스'라는 기능을 개발하였는데, 이는 여행 트랜드와 날짜를 분석하여 집주인에게 알맞은 예약 가격을 추천한다. 또한 집주인은 프라이스팁스로 특정 가격 전후로 여행객이 얼마나 모일지 예측할 수 있다.

[그림 5] 머신러닝을 활용한 숙박 예약시스템

자료 : 에어비엔비

현재 머신러닝은 문자/문서 인식에서부터 인터넷 정보검색, 음성인식/언어처리, 생체인식, 컴퓨터 그래픽, 금융데이터 분석, 의료정보, 로보틱스 등 거의 모든 분야에서 활용되는 기술로 각광받고 있다.

머신러닝은 다양한 응용분야에서 실용적 가치가 크다. 많은 데이터로부터 규칙성을 발견하는 문제(Data Mining), 문제의 성격 규명이 어려워 효과적인 알고리즘을 개발할 지식이 없는 문제 영역(Human Face Recognition), 변화하는 환경에 동적으로 적응하여야 하는 문제 영역(Manufacturing Process Control) 등 다양한 분야로 더욱 확대될 것으로 전망된다.

21) 남는 공간이 있는 사람과 머무를 곳을 찾는 사람을 연결해주는 커뮤니티 마켓플레이스이다.
22) 데이터를 분석해주는 소프트웨어로 스칼라, 자바 등이 활용됐으며, 데이터 간의 우선순위를 정해준다. 예를 들어 수많은 데이터를 분석해 서비스 가격과 수요에 대한 상관관계를 분석한다.

머신러닝과 관련된 다양한 활용분야의 대표적인 예는 [표 3]과 같다.

[표 3] **머신러닝의 활용 분야 및 예**

활용 분야	내용 및 적용사례
인터넷 정보검색	· 텍스트 마이닝, 웹로그 분석, 스팸필터, 문서 분류, 여과, 추출, 요약, 추천 등에 활용 · 활용 예 : 다음(Daum)의 '바로이거', 구글의 '대화형 검색', '지식 그래프(Knowledge Graph)', 이용자의 다음 질문 예측 결과를 보여주는 검색 등
문자/문서 인식	· 숫자 인식과 문자인식은 초기 패턴인식의 연구 대상 · 활용 예 : 상용 컴퓨터의 운영체제나 전자사전 등에서 문자 인식 기능, 은행 ATM의 자동 지로 납부 기능 등
컴퓨터 시각	· 문자/패턴/물체/얼굴 인식, 장면전환 검출, 화상 복구 등에 활용 · 활용 예 : 페이스북의 '팬더'프로젝트(사진에서 인물의 성별, 헤어 · 옷 스타일, 얼굴 표정을 식별하는 연구로 사진 태그, 타게팅된 광고를 제공) 등
음성인식/ 언어처리	· 음성 인식, 단어 모호성 제거, 번역 단어 선택, 문법 학습, 대화 패턴 분석 등에 활용 · 활용 예 : 네이버 음성 인식 및 음성 통역기, 개인 비서 서비스인 구글 나우에 내장된 음성 인식의 정확도 향상 기능 등
모바일 HCI*	· 모바일 기기의 각종 센서를 통한 정보인식, 상황 판단 및 입력 해석을 위한 지능형 처리기술, 동작 인식, 제스쳐 인식 등에 활용 * 모바일 환경에서 인간과 컴퓨터간 자연스러운 상호작용(Human-Computer Interaction, HCI)은 지능적, 능동적으로 사용자의 의도와 입력을 파악하여 정보를 처리 · 활용 예 : 애플의 '시리(음성인식과 자연어 처리기술, 음성합성기술이 융합)', 구글의 '구글 나우(개인비서 서비스)' 등
생물 정보학	· 유전자 인식, 단백질 분류, DNA 칩 분석, 질병 진단, 염기서열 분석 등에 활용 ※ 생물정보학의 초기 단계에서는 주로 통계적 데이터 분석법이 많이 활용되었으나, 점점 머신러닝을 활용한 연구가 주목받음
바이오 메트릭스 (생체인식)	· 홍채 인식, 심장 박동수 측정, 혈압측정, 당뇨치 측정, 지문 인식 등에 활용 ※ 최근 활발히 상용화가 이루어지고 있는 패턴인식 응용분야 중의 하나임 · 활용 예 : 생체정보(지문, 얼굴, 홍채, 망막, 손금 등)에 머신러닝을 적용한 신원 확인 등
뇌신호 처리	· 인간의 뇌 신호를 분석하여 그 의미를 알아내고, 뇌와 컴퓨터의 인터페이스 수단으로 사용하고자 하는 연구(Brain Computer Interface, BCI) 등에 활용 ※ 뇌과학 연구가 활발해짐에 따라 관심을 모으고 있는 주제 중의 하나임
의료정보	· 의료현장에서 얻어지는 임상 데이터나 최근 개발된 다양한 의료 영상기기(MRI, CT, 초음파 등)로부터 얻어지는 데이터들을 분석하여 질병진단 등에 필요한 의미있는 정보 추출 등에 활용
금융데이터 분석	· 홈쇼핑 데이터, 주식데이터, 보험회사의 고객 정보 등 다양한 금융데이터를 분석하여 의미있는 정보 추출에 활용
컴퓨터 그래픽	· 데이터기반 애니메이션, 캐릭터 동작제어, 행동 진화, 가상현실 등에 활용
로보틱스	· 장애물 인식, 물체 분류, 지도 작성, 무인자동차 운전, 경로 계획, 모터 제어, 객체인식, 초음파/적외선 신호 분석 등에 활용
서비스업	· 고객 분석, 시장 클러스터 분석, 고객관리(CRM), 마케팅, 상품추천 등에 활용 · 활용 예 : 유튜브에서 영상을 추천하는 알고리즘 등
제조업	· 이상 탐지, 에너지 소모 예측, 공정 분석 계획, 오류 예측 및 분류 등

자료 : 박혜영, 이관용, "패턴인식과 기계학습", 이한출판사, 2011과 장병탁, "차세대 기계학습 기술", 정보과학회지 제25권 제3호, 2007.3. 재구성

2. 금융권 머신러닝의 활용

선도 금융기관들은 마케팅, 투자 관리 및 트레이딩, 리스크 관리, 고객 서비스 등 경영활동의 다양한 분야에 빅데이터를 활용하고 있으며, 나아가 사내에 축적된 대량의 데이터 분석 결과를 외부에 제공하여 신규 수익 창출 기회로도 활용[23]하고 있다.

실제 금융권에서 머신러닝을 활용하여 데이터를 분석할 수 있는 분야는 광범위하다. 하지만 금융 데이터 특성, 데이터의 구조화 수준 및 처리 기술, 비즈니스 활용 목적에 맞는 분석 기법, 전문 인력 등 자원 및 역량 확보 등의 문제 등으로 모든 데이터 분석에 있어서 머신러닝을 활용할 수는 없다.

본고에서는 최근 금융권과 핀테크 기업 등에서 기존의 단순 통계기반의 데이터 분석이 아닌 실제 머신러닝 알고리즘 적용 및 예측 모델링 등을 통하여 고객 이탈 경향 분석, 투자 관리 및 트레이딩, 사기 및 부정방지, 신용 평가 및 심사 등 머신러닝이 활용되고 있는 분야에 대하여 설명하고자 한다.

가. 영업 및 마케팅

제품 추천(Product Recommendation)이나 최적 대안 제시(Next Best Action)등 최신 마케팅 기법에서는 데이터 분석을 통해 구매 가능성이 가장 높은 제품을 예측한다. 추천시스템은 협업필터링 알고리즘[24]을 이용하여 현재 대표적인 전자상거래 업체인 아마존의 상품추천은 판매의 35%가 추천으로 발생[25]하는 등 대형 서비스와 함께 지속적으로 영역이 확대되어 가며 점점 중요성이 높아지고 있다.

또한 미국 퍼스트 테네시(First Tennessee) 은행은 데이터의 통계분석(2년간의 마케팅 ROI와 고객 대응 데이터를 분석) 및 모델링을 통해 마케팅에 활용하여 메일 발송비용은 20% 감소시키면서 고객 대응률은 3.1%를 증가시켰다. 결과적으로 예측 분석 투자 비용 대비 600%의 수익을 창출하였다.[26]

이렇듯 금융회사들은 이미 통계 기반 분석모델을 이용하여 고객정보를 분석하고 있지만, 기존 통계적 기법의 한계를 보완하고 정확히 예측하기 위해서는 머신러닝을 활용하여 고객

23) KB금융지주 경영연구소, 금융업의 빅데이터 활용, 2013.7.

24) 추천의 근간이 되는 유사점을 분류하는 방식이며, 사전에 누적된 데이터를 분류하고 새로운 데이터를 대입하여 분류하는 방법이다.
Greg Linden, Brent Smith, Jeremy York, Amazon.com Recommendations Item-to-Item Collaborative Filtering, IEEE Computer Society, 2003.1.

25) McKinesy&Company, The Secret of Amazon : Lessons for Multichannel Retailers, 2012.10.

26) IBM, 지식기반의 은행 업무 구현, 2010.5.

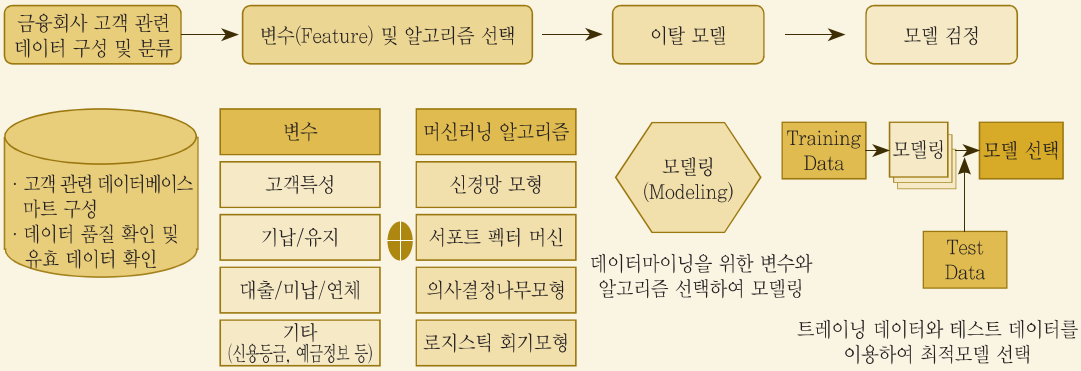
이탈 경향 분석 및 예측이 가능하다. 고객 이탈 경향은 제품추천이나 최적 대안제시 등의 마케팅 기법의 사례와 매우 유사하지만, 고객의 이탈 경향을 추정한다는 점이 다르다.

금융회사는 이탈 가능성이 큰 고객을 예측하고 안내 서비스, 타겟 마케팅 또는 특별 관리 적용 등을 통해 해당 고객의 요구사항을 해결해야한다. 즉, 비즈니스 측면에서 특별 고객에 대한 유지와 관리의 노력을 최소화하기 위해 고객 이탈 경향을 예측하는 것은 금융회사의 핵심 비즈니스 영역이다.

보험사의 경우 예측모델을 수립하면 이탈가능 고객[27], 갱신 고객의 보험 갱신율도 예측하여 활용할 수 있다. 갱신 대상 고객이 보험을 갱신하지 않을 것으로 예측될 경우 타깃 마케팅을 통해 고객을 유지하여 궁극적으로 이탈을 방지하는 것이 목적이다.

머신러닝을 활용하여 이탈 가능 고객예측 모델 수립 프로세스는 일반적으로 ①금융회사 고객 관련 데이터를 구성 및 분류하고 ②고객특성 등 변수(Feature) 및 머신러닝 알고리즘을 선택하여 ③이탈 모델을 만들고 ④모델 검정을 통해 최적화된 모델을 선택하는 과정을 거친다.([그림 6])

[그림 6] 머신러닝을 활용한 이탈가능 고객예측 모델의 예



자료 : 데이터솔루션, 금융기업 경쟁력 강화를 위한 예측분석 활용방안(제2회 빅데이터리더스포럼) 재구성

나. 투자 관리 및 트레이딩

증권권역 머신러닝은 주로 트레이딩(Trading)[28] 시스템에서 예측 정확도와 수익률 향상을 도모하는데 활용된다.

27) 예를 들어 보험회사의 경우 고객과의 거래가 정상적으로 유지되지 않고 중단된 상태로 즉, 보험 상품에 가입 후 해약 또는 장기간 보험금 납입을 연체하는 경우를 의미한다.

28) 트레이딩은 파생금융상품거래에서 사용되는 용어로 외환, 채권, 주식 등의 가격변동을 예측하여 이로부터 매매차익을 획득하려는 목적의 거래이다.

트레이딩 시스템[29]에서 매매체결의 이익을 얻기 위해서 각종 트레이딩 기법[30]들이 있으며, 트레이딩이 전산을 통하여 이루어지는 만큼 트레이딩 기법은 최신 정보 기술의 영향을 받으며 트레이딩 기법에 머신러닝이 활용되기도 한다.

실제 SVM, 신경망 등의 머신러닝 알고리즘을 통해 기존 주가 등락율의 매수, 유지, 매도를 분석한 예측결과(종목 추천 등)를 모바일 주가예측 애플리케이션으로 제공[31]하는 서비스도 이용되고 있다. 또한 주가예측 모형 개발에 있어 빅데이터가 활용[32]되기도 하였으며, 주가 데이터에 머신러닝을 활용한 주가 등락 예측 연구들도 지속적으로 수행되고 있다.

다. 사기 및 부정 방지

금융권에서 머신러닝은 사기 및 부정방지 기능을 고도화하여 사후 뿐 아니라 사전 대응을 위해 활용된다. 특히 이상거래 탐지시스템(Fraud Detection System, 이하 FDS라고 한다.)[33]에서 머신러닝은 일반적으로 현재 진행 중인 거래의 위험도와 특정 거래의 발생 가능성을 예측하는데 사용된다.

FDS에서 "분석 및 탐지 기능"은 수집 시스템에서 전달받은 수집 정보를 활용하여 이상 탐지 여부를 판단하는 기능으로 탐지방법은 탐지 모델별로 상이하며, 데이터베이스에 탐지패턴을 저장하여 관리한다. 탐지모델은 서비스 유형에 따라 단일 또는 복합적으로

29) 보통은 구매자, 판매자의 전문 투자자가 사용하는 거래 시스템을 통칭하지만 국내에서는 증권사(판매자) 법인영업 및 상품운용부서가 현물과 파생상품을 거래하기 위한 시스템으로 한정하는 경우가 많다. 주식거래시스템 대부분의 트레이딩 시스템은 의사결정을 내리기 위해 필요한 시세데이터(data source), 의사결정알고리즘, 거래소와의 접속으로 구성된다.
코스콤, IT용어사전

30) 전통적인 기법들은 금융공학과 관련되며 통계를 비롯한 수학에 그 뿌리를 두고 있다. 그 이후 특정 조건 하에서의 주문 실행을 규칙화하여 매매를 자동화하는 지표추종형 전략으로 개선되었다.
트레이딩 기법에는 ①전통적 트레이딩 기법(기업의 재무적 평가기준을 이용한 트레이딩, 기술적 분석을 이용한 트레이딩, 두 상품간의 가격 차이를 이용한 트레이딩), ②알고리즘 트레이딩(TWAP, VWAP, POV, Iceberg 등 주문실행 알고리즘을 이용한 트레이딩), ③고빈도 매매(시장조성 전략, Stuffing, Smoking, Spoofing 등), ④인텔리전스 트레이딩(다양한 사람들의 아이디어를 수집하고 가공하여 트레이딩 전략을 수립한 다음 신속하게 트레이딩 어플리케이션에 반영, 머신러닝/인공 신경망 등 빅데이터 기술도 활용되는 경우가 있음) 등이 있다.
코스콤 홈페이지(http://www.koscom.co.kr)

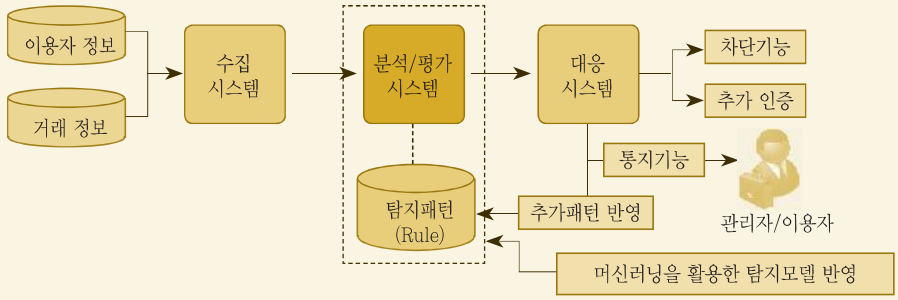
31) 지디넷코리아, 데이터 기반 개인용 주가 예측 통할까?, 2013.12.10

32) 코스콤은 29일 "지난 1월 빅데이터 기술을 활용한 '주가분석 및 예측 시스템' 연구에 착수해 10개월여 만에 성공적으로 개발을 완료했다"며 "SNS에서 사용되는 단어와 블로그와 카페 등에 쓰인 단어 등을 수집 · 분석해 이를 주가 예측에 활용하는 기법"이라고 설명했다. SNS에서 쓰이는 5만9000개의 긍정, 부정 단어를 포함한 감성 사전과 뉴스 · 블로그 · 카페 등에 기재된 주요 단어 25만개의 형태소 사전을 수집, 분석에 활용하였다.
NEWSIS 뉴스, 'SNS 빅데이터로 주가 예측한다'…코스콤, 내달 시범서비스, 2013.11.29

33) 전자금융거래에 사용되는 단말기 정보 · 접속 정보 · 거래 내용 등을 종합적으로 분석하여 의심거래를 탐지하고 이상금융거래를 차단하는 시스템을 의미한다.
금융보안연구원, "이상금융거래 탐지시스템 기술 가이드", 2014.8

이용되며 크게 ①오용탐지모델(Misuse Detection Model) ②이상탐지모델(Anomaly Detection Model)[34]기법이 있다.

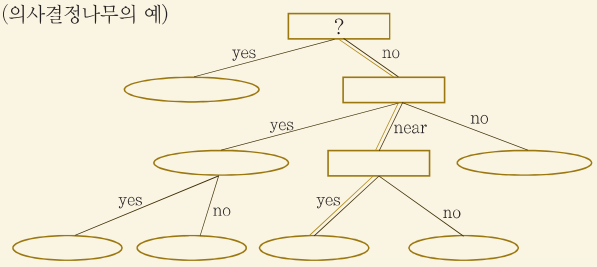
[그림 7] FDS에서 분석 및 탐지기능의 범위와 머신러닝



FDS 탐지 모델에 활용되는 머신러닝 알고리즘의 예는 아래의 [표 4]와 같으며, 다양한 알고리즘이 최적의 탐지 효과를 보이기 위해서는 각 금융사 서비스의 실제 데이터와 결합되어 커스터마이징(Customizing) 되어야 한다.

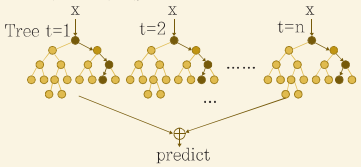
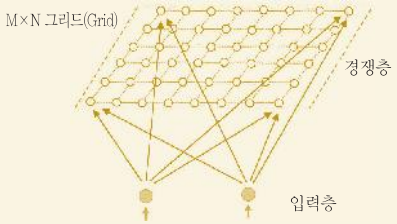
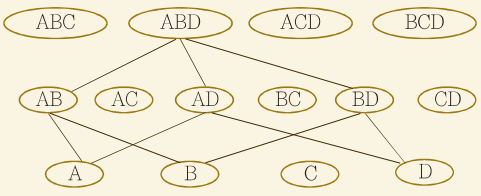
FDS에서 정상적인 거래와 부정거래를 구분 또는 예측하기 위해 장기간의 자료축적과 분석이 필수이며, FDS의 고효율을 위해 페이팔(PayPal)에서는 이상거래를 판별하는 인공지능에 딥러닝을 적용[35]하였다.

[표 4] 이상금융거래 탐지 모델에 활용되는 머신러닝 알고리즘의 예

구분	내용	
의사결정 나무 (Decision Tree)	개념	의사를 결정하거나 분류 · 예측하는데 사용하는 트리로 가장 큰 조건의 트리 뿌리를 만들고, 세부 조건의 트리 가지를 만들며, 해결 방안은 트리의 잎(Leaf) 노드로 의사결정 나무를 형성하여 분석하는 알고리즘 (의사결정나무의 예) 
	장점	실시간 적용이 가능하고 분류 과정이 트리 구조에 의한 추론규칙으로 표현되기 때문에 쉽게 이해하고 설명 가능
	단점	특성 개수에 따라 트리의 모양이 많이 달라질 수 있으며, 출력이 다양할 경우 트리는 매우 복잡하여 예측 결과가 떨어짐
	적용사례/ 시나리오	실시간 부정 IP차단 적용, 엔트로피를 이용한 IP 오염도, IP Address 및 Action 로그 등

34) 알려지지 않은 부정거래행위에 대한 사전 탐지가 가능하나, 오탐률이 높으며 수집된 정보를 분석하는데 많은 학습 시간이 소요되며, 주로 "통계모델 · 데이터 마이닝 모델" 등을 이용한다.

35) ZDNet Korea, 페이팔, 결제사기 막으려 '딥러닝' 도입, 2015.3.10

랜덤 포레스트 (Random Forest)	개념	데이터의 일부를 추출하여 의사결정나무를 만드는 작업을 반복하고, 만들어진 다수의 의사결정나무들의 투표로 최종 결과를 도출하는 방식 (랜덤 포레스트의 예) 
	장점	빠른 학습 속도를 가지므로 많은 양의 데이터 처리 능력을 가지며, 단계별 노드의 수를 조절하여 멀티클래스로 쉽게 확장이 가능함
	단점	노이즈 데이터가 많은 경우 과적합(Overfitting) 될 수 있으며, 해석하는데 어려움이 존재함
	적용 시나리오	결제 종류, 금액, 시간, 지역, 횟수 등 여러 가지 특징들 중 랜덤으로 특징과 특징의 개수가 선택됨
자가 조직도 (Self-Organization Map, SOM)	개념	주어진 입력패턴에 대하여 해답을 미리 주지 않고 자기 스스로 학습하며, 샘플들을 상호 비교하며 스스로 군집을 조직해냄 (자가 조직도 신경망 구조의 예) 
	장점	전방향(Feed-forward) 인공 신경망으로 구성되어 수행 속도가 빠르며, 입력 데이터의 분류와 자료에 숨겨져 있는 패턴을 감지
	단점	많은 입력 데이터에 대해 전처리 과정이 필요하며, 블랙박스와 같은 신경망의 각 층들 간의 학습을 통해서 결과를 계산하기 때문에 결과 값에 대한 과정 설명이나 추론이 어려움
	적용 시나리오	· 개인정보 도용 후 다량의 거래를 수행하기 전 개인정보를 변경 · 거래 전 잘못된 로그인의 수가 증가 · 블랙리스트 처리된 지역에서의 거래가 발생 · 기존의 거래가 발생한 시간대와 다른 시간대에서 거래가 발생
연관 규칙 (Association Rule)	개념	하나의 항목 집합과 다른 항목 집합 사이의 연관성을 나타내는 것으로 일련의 거래나 사건들의 연관성에 대한 규칙을 분석하는 알고리즘 (연관규칙의 기본개념의 예) 
	장점	많은 양의 데이터를 대상으로 하거나 변수의 개수가 많은 경우에도 쉽게 사용 할 수 있으며, 계산이 용이하여 결과 값이 분명함
	단점	품목 수의 증가에 따라 계산량이 폭증하고, 자료의 속성에 따라 제한 사항이 존재함
	적용 시나리오	· "특정행위나 특징이 있는 사용자는 부정 거래를 한다."라는 규칙을 생성할 수 있음 · 결제 금액, 기간별 거래 횟수, 로그인 횟수 등 연관 규칙 생성에 적합하지 않을 수치 데이터를 퍼지이론(Fuzzy Logic)을 이용해 정규화하고 연관 규칙을 생성할 수 있음 · 사기 거래 탐지 속도가 실시간 서비스 적용에 적합하며 언어 형태규칙을 제공하므로 사람이 직관적으로 이해할 수 있고 사기 거래 탐지에 대한 근거로 사용할 수 있음

라. 신용 평가 및 심사

신용 평가 및 심사에서 머신러닝을 활용한 데이터 분석은 대출 신청자의 사회경제적인 특성, 신용정보와 지급이력과 같은 세부정보를 바탕으로 신용도, 특정 대출에 대한 채무불이행 가능성 예측 등을 수행한다.

금융권 관계자들에 따르면 국내 신평사들은 대략 1000개의 신용거래 정보를 받는다면 이 가운데 100~200개 정도만 사용한다. 현재의 신용평가 모델은 회귀분석방식이다. 하나나 그 이상의 독립변수를 가지고 종속변수를 추정하는 식이다. 분석이 빠르다는 장점이 있지만 분석이 정형화돼 있고 일부 신용정보로 10등급까지만 나눈다는 점이 한계라는게 전문가들의 평가다. 이에 머신러닝은 회귀분석의 대안으로 떠오르고 있다.[36]

핀테크 기업인 제스트 파이낸스(Zest Finance)[37]는 빅데이터와 머신러닝을 활용하여 신용 평가에 접목시켰다. 제스트 파이낸스는 개인이 파산 이후 어떤 노력을 했는지 등을 포함해 거의 1만개 이상 변수로 신용도를 분석하고 대출 여부를 결정한다.[38] 이를 통해 일반 신용등급 평가에서 낮은 등급을 받아 대출을 받지 못하는 사람들을 주 고객으로 확보할 수 있다. 또한 신용평가 모델링에 있어서 머신러닝 알고리즘을 활용하여([그림 8]) 머신러닝의 특징 중 하나인 관계와 분류의 정확성을 높일 수 있다.

[그림 8] 제스트 파이낸스의 머신러닝을 활용한 신용평가 모델 적용 사례

자료 : GIGAOM, Where machine learning and 'human artistry' meet your wallet 재구성

36) 조선비즈, [신용평가의 진화]①빅데이터의 묘기 "대출 받는 걸 와이프가 알고 있나요?", 2015.06.10
37) http://www.zestfinance.com/
38) 전자신문, 미국 핀테크 스타트업, 중국에 빅데이터-머신러닝 들인다, 2015.6.29

V. 머신러닝 관련 동향 및 이슈

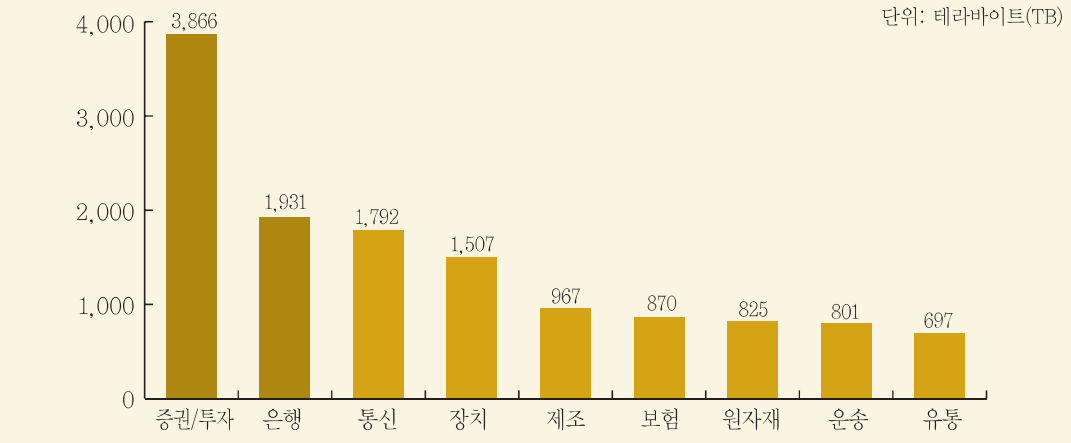
1. 머신러닝 관련 동향

MIT가 '2013년을 빛낼 10대 혁신기술' 중 하나로 선정[39]하고 가트너(Gartner, Inc.)가 '2014 세계 IT 시장 10대 주요 예측'[40]에 포함시키는 등 머신러닝에 대한 관심과 성장은 최근 빠르게 높아지고 있다.

머신러닝에 대한 관심과 성장을 주도하는 요인은 주로 ①빅데이터의 발달 ②정보처리(연산, 저장) 능력의 향상 ③딥러닝 알고리즘의 특징 ④편리한 클라우드 기반 머신러닝 솔루션의 등장 등에 있다고 볼 수 있다.

머신러닝 관련 동향은 ①빅데이터의 발달과 관련 동향을 함께 살펴보아야 한다. 증권/투자, 은행, 보험사가 보유한 데이터량은 총 6667TB로 파악되며 전체의 약 50%를 차지하고 있고[41]특히 글로벌 금융기업은 타 산업 대비 높은 데이터 보유량을 기록하며 이를 경쟁우위로 활용하기 위한 방안을 강구하고 있다.

[그림 9] 미국 산업별 평균 데이터 보유량



자료 : KB금융지주 경영연구소, 2012.6.27

39) MIT 선정 올해의 10대 혁신기술로써 학습과 추론을 통해 의사소통이 가능한 인공지능 기술이 포함되었다.
"MIT, 올해의 10대 혁신기술 선정", 동아일보, 2013.4.26

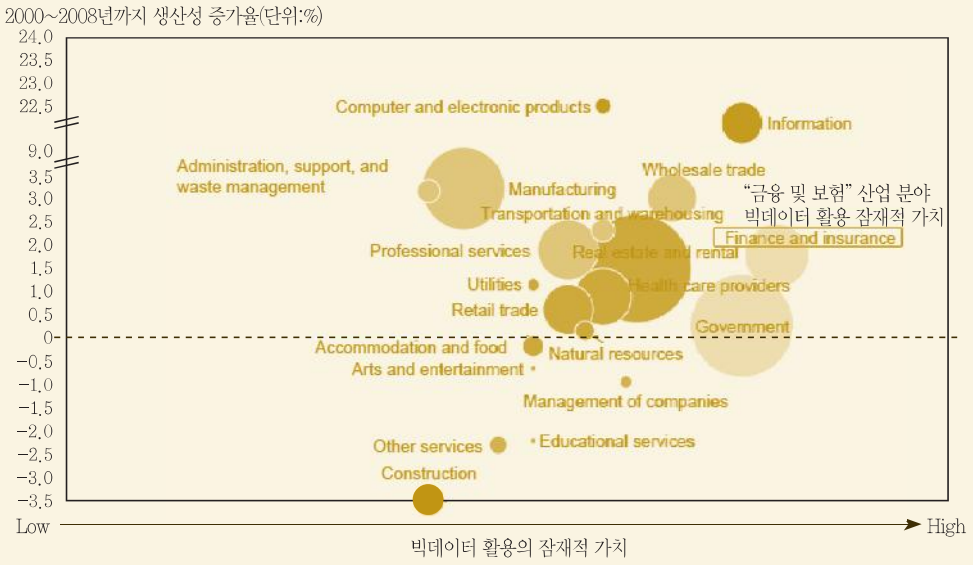
40) 2017년에는 컴퓨터의 10%는 '처리'보다는 '학습'을 하게 될 것이다. 2014년 심층 신경(neural) 네트워크 알고리즘을 운영하는 음성 인식 애플리케이션의 수는 배가 될 것이다. DNN(Deep Neural Network)을 바탕으로 한 심화 학습 방법이 일부 사물 인식 애플리케이션은 물론이고, 음성 인식 시스템에도 적용되고 있다. 인터넷에서 막대한 비정형 데이터를 수집해 유용한 정보를 획득 할 수 있을 때 삶의 질은 개선된다. 학습 컴퓨터가 갖는 가장 중요한 의미는 복잡한 패턴 인식에 훨씬 적은 에너지를 사용한다는 것이다.
"Gartner, Gartner Reveals Top Predictions for IT Organizations and Users for 2014 and Beyond", Gartner newsroom, 2013.10.8

41) 미래창조과학부, 한국정보화진흥원, 빅데이터 전략센서, "2015년 빅데이터 글로벌 사례집", 2015.5

빅데이터 시장규모는 계속 커지고 있는 상황[42]이며, [그림 10]과 같이 금융 및 보험업 분야에서 빅데이터 활용의 가치는 상대적으로 다른 산업군에 비해 큰 위치를 차지하고 있음을 알 수 있다.

머신러닝은 빅데이터 시대에 전통적인 시스템 공학보다 데이터 숨겨진 정보와 가치를 효율적으로 찾을 수 있다는 점에서 특히 그 가치를 증명 받고 있다. 특히 금융권에서는 포트폴리오 분석, 트레이딩, 리스크 관리, 마케팅, 보안 등으로 머신러닝 활용 수준을 확대해가고 있다.

[그림 10] 산업별 빅데이터 활용의 잠재적 가치 (미국 사례)



* 주) ◯원의 크기는 상대적으로 산업별 GDP 기여도를 나타냄

자료 : McKinsey&Company, Big data:The next frontier for innovation, competition, and productivity, 2011.6

②정보처리(연산, 저장) 능력의 향상은 머신러닝의 발달을 가능하게 한 요인이다. 머신러닝은 거대한 데이터를 연산하는 작업이므로 상당한 기술적 토대가 마련되어야 가능하다. 이러한 기술적 토대는 대규모 데이터를 빠른 속도로 처리하기 위해 아파치 하둡(Apache Hadoop) 등과 같은 분산 처리(컴퓨팅 및 저장)[43]의 발달과 그래픽

42) Wikibon은 향후 빅데이터 시장 규모가 2012년 51억 달러에서 2017년 534억 달러로 보다 높은 성장률(연평균 60%)을 달성할 것으로 예상하였다.
http://wikibon.org/wiki/v/Big_Data_Market_Size_and_Vendor_Revenues

43) 분산 컴퓨팅(Distributed Computing)은 인터넷에 연결된 여러 컴퓨터들의 처리 능력(분산 처리 기술, 분산 컴퓨팅 기술)을 이용하여 거대한 계산 문제를 해결하려는 분산처리 모델이다.

처리장치(Graphics Processing Unit, 이하 GPU라고 한다.)[44] 등과 같은 고성능 프로세서의 발전이 있었기 때문이다.

기존에 컴퓨터 성능은 주로 중앙처리장치(Central Processing Unit, 이하 CPU라고 한다.)에만 의존하는 형태로 한계를 보였으나, 멀티미디어 콘텐츠 등을 다루는 고속 그래픽 처리에 특화된 전용 프로세서인 GPU를 (멀티)탑재하여 성능을 높이는 환경을 구현하기도 한다. 또한, GPU상의 범용 처리(General-Purpose Computing on Graphics Processing Units, GPGPU)는 일반적으로 컴퓨터 그래픽스를 위한 계산만 맡았던 GPU를, 전통적으로 CPU가 맡았던 응용 프로그램들의 계산에 사용하는 기술이다.

이를 통하여 정보처리 성능의 비약적 향상을 가져왔고, GPU를 활용한 머신러닝, 딥러닝이 적극 연구[45] 및 활용[46]되고 있다.

③딥러닝 알고리즘의 특징도 머신러닝에 대한 관심이 높아지고 있는 이유 중의 하나이다. 왜냐하면 머신러닝의 방법 중의 하나인 딥러닝 알고리즘은 시뮬레이션의 크기를 늘릴수록 대량의 데이터를 흡수하는 능력이 좋아지는 특징을 가진 초고용량 학습 알고리즘이다. 딥러닝은 학습 모델링과 예측률을 높이기 위해 풍부한 과거 데이터[47]인 빅데이터의 발달과 복잡한 딥러닝의 함수에 대한 모델링 연구와 함께 딥러닝을 활용할 수 있는 가능성이 열리고 있다. 구글, 페이스북, 트위터 등과 같이 딥러닝 기술을 사용하는 곳은 빅데이터를 가진 곳이다. 글로벌 IT기업의 딥러닝 활용 동향은 [표 5]와 같다.

44) GPU라는 용어는 엔비디아(NVIDIA)社에서 1999년에 '지포스(GeForce)'라는 이름의 그래픽 컨트롤러(Graphics Controller:그래픽카드용 칩)를 내놓으며 처음 붙여진 이름이다. 지포스는 CPU의 도움 없이 자체적으로 폴리곤(Polygon: 3D 그래픽을 구성하는 도형)의 변형(Transform) 및 광원(Lighting)효과를 구사하는 기능을 갖추고 있다. 이는 이전까지 사용했던 그래픽 컨트롤러와는 다른 개념이었기 때문에 GPU라는 이름으로 구분하게 되었다. 2000년에 ATi(현재의 AMD)사에서 '라데온(Radeon)'이라는 GPU를 출시하게 되면서 양사의 GPU 경쟁이 본격화된다.

45) 앤드류 응 교수와 엔비디아는 16대의 GPU 가속화 서버를 사용해 112억 개의 파라미터를 갖춘 신경회로망을 구축하고, 인공 신경회로망에서 발생하는 대량의 데이터를 CPU가 아닌 GPU가 처리하도록 했다. 이는 2012년 구글 브레인의 신경회로망보다 6.5배 큰 것이나, 구축 비용은 현저히 줄어들었다. 그 결과는 국제기계학습학술대회(ICML) 2013, "Deep Learning with COTS HPC Systems" 논문을 통해 발표했다.
디지털데일리, GPU 컴퓨팅이 기계학습 주도, 2014.3.28

46) 중국 바이두에서는 머신러닝 알고리즘인 신경망 학습에 GPU를 적용하였다.
WIRED, Chinese Google' Unveils Visual Search Engine Powered by Fake Brains, 2013.6.13

47) 보통 트레이닝 데이터(Training Data)라고 부른다.

[표 5] 글로벌 IT기업의 딥러닝 활용 동향

회사	내용
구글 (Google)	· 2011년 앤드류 응 (스탠포드대)교수는 구글 안에 딥 러닝 프로젝트를 구성, 음성인식과 구글 플러스의 사진 태깅에 딥러닝 기술을 활용하기 시작 · 2012년 응 교수의 팀은 1만 6000개의 컴퓨터 프로세서로 10억 개 이상의 연결을 갖는 뉴럴 네트워크를 이용한 자율학습 방식의 딥러닝 기술을 적용해 유튜브 안에 있는 1000만 개의 이미지 중에서 고양이를 알아내는 연구수행 · 2012년 젤리빈(jellyBean)부터 음성인식서비스에 딥러닝 활용 · 2013년 3월 제프리 힌튼 교수와 토론토대학의 연구자들 영입 및 힌튼교수의 회사인 DNN 리서치를 인수함. 구글 나우의 음성인식, 유튜브 추천, 이미지 물체에 대한 자동 태깅 등 다양한 영역에서 딥러닝 기술을 이용 · 구글은 딥마인드(DeepMind)라는 회사를 4억 달러가 넘는 금액으로 인수* * Re/code NEWS, "Google to Buy Artificial Intelligence Startup DeepMind for $400M", 2014.1.26
페이스북 (Facebook)	· 2013년 얀 레쿤(뉴욕대) 교수를 영입 및 '인공지능 연구그룹'을 출범 · 연구그룹에서 '딥 페이스 기술'*을 발표하여 인간과 유사한 97.35% 정확도로 다양한 각도 · 조명에서 사람 얼굴을 인식할 수 있는 기술을 선보임 * Conference on Computer Vision and Pattern Recognition(CVPR), Yaniv TaigmanMing, YangMarc'Aurelio, Ranzato, Lior Wolf, "DeepFace:Closing the Gap to Human-Level Performance in Face Verification", 2014.6.24
트위터 (Twitter)	· 매드비츠(Madbits)라는 딥러닝 회사를 인수해 사진 이미지 분석기술을 확보하고자 함* * TC News, "Twitter Acquires Image Search Startup Madbits", 2014.7.29
마이크로소프트 (MS)	· 2014년 MS 윈도폰용 음성인식 개인 비서 서비스인 '코타나(Cortana)' 출시 · 2014년 7월 'MS 리서치 학술회의 2014'에서 AI 프로젝트인 '아담'을 통해 견종을 컴퓨터가 분류하는 등 시각적 정보를 활용해 사물을 인식하는 딥러닝 기술을 공개
바이두	· 2014년 5월 미국 실리콘밸리에 인공지능(AI) 연구소를 개설하고, 2013년 설립된 베이징 바이두 연구소와 함께 앤드류 응 교수를 총책임자로 영입 · '2014 바이두 월드 컨퍼런스(World Conference)'에서 안경형 웨어러블 디바이스인 '바이두 아이(Baidu Eye)'를 공개하여 물체를 스캔하고 인식하는 이미지 인식기술을 선보임

④편리한 클라우드 기반 머신러닝 솔루션의 등장도 머신러닝에 대한 성장을 주도하는 배경적 요인이다. 금융회사 내부에 데이터 분석가가 있다면 클라우드 기반의 머신러닝 서비스를 적용하여 수요예측, 의사결정 등에 활용이 가능하다.

이러한 대표인 서비스로는 구글의 프리딕션 API, 아마존의 아마존 ML, 마이크로소프트의 애저(Azure)서비스 등이 있으며, 그 특징과 내용은 [표 6]과 같다.

[표 6] 클라우드 기반 머신러닝 서비스 제공 현황

<table>
<tr><th>구분</th><th colspan="4">내용</th></tr>
<tr><td rowspan="11">구글
프리딕션 API
(Google
Prediction API)
  </td><td>개발사</td><td colspan="3">· 구글(Google)</td></tr>
<tr><td>서비스</td><td colspan="3">Google Cloud Platform https://cloud.google.com/prediction</td></tr>
<tr><td>개념</td><td colspan="3">· 데이터를 분석할 수 있는 클라우드 기반의 예측모델 구축 지원 서비스</td></tr>
<tr><td>시작</td><td colspan="3">· 2010년 5월 구글 I/O 컨퍼런스에서 클라우드 머신러닝 기능 소개
· 2011년 말 서비스 상용화 시작</td></tr>
<tr><td>내용</td><td colspan="3">· 고객 심리분석, 스팸 탐지, 판매증가 기회 분석, 추천 시스템, 의심스러운 활동 식별 등과 같은 특징에 따라 데이터를 분석 및 예측
· Prediction API(샘플, 라이브러리, 데이터 전(前)처리 등)와 Prediction Tool(구글 개발자 콘솔)을 제공
※ 전제조건 : 프리딕션 API와 클라우드 스토리지 API가 활성화된 상태로 구글개발자 콘솔 사용
· 프리딕션 API 구현 3단계 :
① 업로드(upload) : 구글 스토리지에 해당 데이터 업로드
② 학습(Train) : 데이터로부터 모델을 구축
③ 예측(Predict) : 해당 데이터를 이용한 새로운 예측</td></tr>
<tr><td>특징</td><td colspan="3">· 사용자인터페이스(UI)가 없으나, 마법사 형태로 사용이 용이
· 모델을 구축하는데 선택할 수 있는 알고리즘을 여러 개 제공</td></tr>
<tr><td rowspan="3">사용
기술</td><td colspan="3">· 프레딕션API를 생성하기 위해 구글 클라우드 스토리지에 연결, 빅쿼리(BigQuery, 구글 빅데이터 분석 서비스) 결과 사용
· 구글의 핵심서비스인 검색에 예측 로직 기술 기반 활용</td></tr>
<tr><td rowspan="2">관련
기술</td><td>분석</td><td>BigQuery(SQL등 사용, 빅데이터 온라인분석처리 플랫폼)</td></tr>
<tr><td>저장</td><td>cloud storage/Datastore(NoSQL)/SQL(관계형 MySQL)</td></tr>
<tr><td rowspan="4">비용</td><td></td><td>무료</td><td>유료</td></tr>
<tr><td>기본요금</td><td>6개월간 무료</td><td>개발자 콘솔 프로젝트 별 매월 10달러</td></tr>
<tr><td>예측(Predictions)</td><td>100 predictions/일</td><td>10,000 predictions/월 : 무료,
초과시 1,000 predictions 당 0.50달러</td></tr>
<tr><td>학습(Training)</td><td>5MB trained/일</td><td>0~10,000 스트리밍 업데이트 : 무료,
초과 시1000 업데이트 당 0.05달러</td></tr>
<tr><td rowspan="6">아마존 머신러닝
(Amazon ML)
  </td><td>개발사</td><td colspan="3">· 아마존(Amazon)</td></tr>
<tr><td>서비스</td><td colspan="3">amazon web services http://aws.amazon.com/ko/machine-learning</td></tr>
<tr><td>개념</td><td colspan="3">· 데이터를 읽어 머신러닝 모델 생성, 신규 데이터 처리 및 애플리케이션 예측</td></tr>
<tr><td>시작</td><td colspan="3">· 'AWS 샌프란시스코 서밋 2015(4월9일)'에서 '아마존 머신러닝'서비스 공개</td></tr>
<tr><td>내용</td><td colspan="3">· 머신러닝 구축과 예측 생성을 지원하는 관리형 서비스
· 아마존 머신러닝을 이용한 머신러닝 모델 구축 프로세스 :
① 데이터 분석 : 데이터 배포를 컴퓨팅 및 시각화
② 모델 학습 : 변환된 데이터에서 예측 패턴을 찾아 저장
③ 평가 : 모델의 정확도 평가(선택사항)</td></tr>
<tr><td>특징</td><td colspan="3">· 데이터 시각화 지원 및 입력 데이터에 대한 데이터 변환(transformation) 지원
· 바이너리 속성(바이너리 분류), 범주별 속성(다중 분류) 또는 수치 속성(Regression, 회귀) 등의 값을 예측하는 모델을 생성
· 많은 기업들이 AWS에 상당한 양의 데이터를 저장하고 있고 S3 스토리지에 연결되어 있는 장점</td></tr>
</table>

<table>
<tr><td rowspan="5"></td><td>사용
기술</td><td colspan="3">· 아마존 S3(Simple Storage Service, 클라우드 스토리지)의 데이터 이용, 아마존 레드시프트(Redshift, 클라우드 기반 데이터 웨어하우스) 또는 아마존 RDS(관계형 데이터베이스 서비스)에 있는 MySQL 데이터베이스에 쿼리하여 머신러닝 모델을 생성 및 사용
· 아마존 전자상거래 비즈니스 예측분석 모델(상품 추천 기술) 활용</td></tr>
<tr><td rowspan="2">관련
기술</td><td>분석</td><td>Amazon EMR(Elastic Map Reduce, 하둡 인터페이스 지원), Amazon Redshift(페타바이트 규모의 데이터 웨어하우스 솔루션, 기존 비즈니스 도구를 사용하여 데이터 분석 지원) 등</td></tr>
<tr><td>저장</td><td>Amazon S3(클라우드 기반 스토리지), RDS(아마존 웹서비스에서 관리하는 MySQL, Oracle 등 지원하는 관계형 DB) 등</td></tr>
<tr><td rowspan="4">비용</td><td colspan="2"></td><td>무료</td><td>유료</td></tr>
<tr><td colspan="2">분석 및 모델 구축</td><td rowspan="3">–
※ AWS 프리티어에서 Amazon S3, RDS 등 일부 서비스의 정해진 한도내에서 무료</td><td>시간당 0.42달러</td></tr>
<tr><td rowspan="2">예측</td><td>배치</td><td>1,000 예측당 0.1 달러</td></tr>
<tr><td>실시간</td><td>예측당 0.0001 달러</td></tr>
<tr><td rowspan="6">마이크로소프트
애저 머신러닝
(MS Azure ML)
  </td><td>개발사</td><td colspan="3">마이크로소프트(MS)</td></tr>
<tr><td>서비스</td><td colspan="3">http://azure.microsoft.com/ko-kr/</td></tr>
<tr><td>개념</td><td colspan="3">애저는 MS에서 관리/지원하는 데이터 센터에 호스팅된 인터넷 규모의 컴퓨팅 및 서비스이며, 애저 머신러닝은 클라우드 기반 예측 모델 구축 지원 서비스</td></tr>
<tr><td>시작</td><td colspan="3">· 2014년 6월 16일 미국에서 '애저 머신러닝'공개프리뷰 발표
· 2015년 1월 21일 한국MS 본사에서'애저 머신러닝'국내 공식 론칭</td></tr>
<tr><td>내용</td><td colspan="3">· 애저 머신러닝 서비스는'머신러닝 스튜디오'*와'머신러닝 API서비스'**등으로 이루어짐
* 데이터에 대한 예측 분석 솔루션을 빌드, 테스트, 배포할 수 있는 공동 작업 시각적 개발 환경
** '머신러닝 스튜디오'에서 제공되는 예측 모델 등을 확장 가능한 웹 서비스로 배포
· 애저 머신러닝을 이용한 예측분석 모델 구축 프로세스 :
① 모델 만들기 : 데이터 가져오기, 데이터 전처리, 기능 정의
② 모델학습 : 학습 알고리즘 선택 및 적용
③ 모델 점수 매기기 및 테스트 : 새 데이터 예측</td></tr>
<tr><td>특징</td><td colspan="3">· 모델을 구축 시 분류, 회귀, 클러스터링 등의 여러 알고리즘을 제공
· 순서도 스타일의 데이터 플로우를 제공
· R, Python 등의 개발 프로그램으로 확장 가능</td></tr>
<tr><td rowspan="3">사용
기술</td><td colspan="3">· 빙(bing)같은 검색 서비스의 머신러닝 기술을 적용
· 애저 HD인사이트(HDInsight)를 포함해 애저 데이터 애셋의 기존 데이터를 머신러닝에 활용
· 클라우드 기반 예측 분석</td></tr>
<tr><td>분석</td><td colspan="2">애저 HDInsigh(클라우드에서 제공되는 Apach Hadoop 기반 서비스, 페타바이트급 지원) 등</td></tr>
<tr><td>저장</td><td colspan="2">SQL데이터베이스를 활용한 관계형 DaaS(Database-as-a-Service), 애저 Blobs 등에 저장된 애저 클라우드 저장소 등</td></tr>
<tr><td rowspan="5">비용</td><td colspan="2"></td><td>무료</td><td colspan="2">유료</td></tr>
<tr><td colspan="2">가입제</td><td rowspan="4">30일간 무료
※초기 발생되는 200 달러 크레딧 한도 내에서 무료</td><td colspan="2">매월 seat당 9.99달러</td></tr>
<tr><td colspan="2">스튜디오 실험</td><td colspan="2">시간당 1달러</td></tr>
<tr><td rowspan="2">API 서비스 예측</td><td>시간기준</td><td>2달러</td></tr>
<tr><td>트랜젝션 기준</td><td>1000개 API 당 0.5 달러</td></tr>
</table>

2. 머신러닝 관련 법적 이슈

가. 빅데이터 활용 관련 이슈

IT기업들이 금융서비스 산업을 시작하면서 핀테크[48]시장이 형성되고, 머신러닝 등을 적용한 빅데이터 분석을 통해 가치 창출을 위한 금융서비스들이 크게 변화하고 있다.

하지만 머신러닝을 활용한 빅데이터 분석에는 개인정보보호 등의 법적 이슈가 존재한다. 특히 개인정보 대량 유출사고에 따른 개인정보보호 이슈가 화두가 되면서 빅데이터의 활용은 위축되었고, 머신러닝과 같은 빅데이터 분석 기술이 나날이 정교해지고 기업들은 광범위한 데이터로부터 가치있는 정보를 추출하는데 집중하면서 개인정보, 프라이버시 침해 등에 대한 우려는 점차 커지고 있다.

금융산업에서 빅데이터의 활용과 개인정보보호 관련 법 제도상의 제약과의 조화로운 균형점에 대해 찾기 어려운 실정이다. 또한 빅데이터 활용 시 개인 및 신용정보 이용에 대해 개인정보보호법 등 현행 정보보호 관련 법규[49]들이 산재되어 있어 해당 법률의 해석에 있어 어려움이 존재한다.([표 7])

[표 7] 개인정보보호 관련 대표적인 법률

구분		법률	주요 내용
일반		개인정보보호법	· 개인정보의 수집, 처리 및 보호에 관한 사항 ▷ 개인정보보호 일반법
민간부문	정보통신	정보통신망법[50]	· 정보통신망의 이용 촉진 및 정보통신 서비스를 이용하는 자의 개인정보보호 규정 ▷ 빅데이터 처리 등 이용자에게 공개 ▷ 처리시스템에서 기술적 · 관리적 보호조치 등
		정보통신기반보호법	· 주요정보통신기반시설의 지정, 금융ISAC의 운영 ▷ 주요정보통신기반시설 준수 법률
	상거래	전자문서 및 전자거래기본법	· 전자문서 및 전자거래의 안전성과 신뢰성 확보 ▷ 전자거래이용자의 개인정보 수집/이용/제공 및 관리에 관한 사항
		전자상거래 등에서의 소비자보호에 관한 법률	· 전자거래시 소비자의 의사표시 확인 ▷ 소비자에 관한 정보이용, 신원 및 거래조건에 대한 정보 제공 등
		전자서명법	· 전자서명에 관한 기본 사항 ▷ 공인인증서, 인증업무의 안전성 및 신뢰성 확보
		산업기술의 유출방지 및 보호에 관한 법률	· 산업기술의 부정한 유출 방지 및 보호

48) 금융을 뜻하는 파이낸셜(financial)과 기술(technique)의 합성어다.
49) 「개인정보보호법」, 「정보통신망 이용촉진 및 정보보호 등에 관한 법률」 등 현행 개인정보 관련 법규는 공공기관, 정보통신서비스제공자 등 정보처리 주체에 따라 다른 법규가 적용된다.
50) 「정보통신망 이용촉진 및 정보보호 등에 관한 법률」의 약칭이다.

		전자서명법	· 전자서명에 관한 기본 사항 ▷ 공인인증서, 인증업무의 안전성 및 신뢰성 확보
		산업기술의 유출방지 및 보호에 관한 법률	· 산업기술의 부정한 유출 방지 및 보호
	금융 · 신용	신용정보의 이용 및 보호에 관한 법률	· 신용정보 전산시스템의 기술적, 물리적 보안대책 ▷ 금융분야 개인(신용)정보보호 법률
		금융실명거래 및 비밀보장에 관한 법률	· 실지명의에 의한 금융거래 및 비밀보장
		전자금융거래법, 전자금융감독규정	· 전자금융거래의 안전성과 신뢰성 확보(안전성 확보 의무, CISO 지정 등) ▷ 금융분야 IT 및 정보보호 법률
		특정 금융거래 정보의 보고 및 이용 등에 관한 법률	· 자금세탁방지를 위한 금융거래 모니터링

이에 개인정보는 보호하면서 빅데이터 활용을 높일 수 있는 대안으로 비식별화 기술에 대한 관심이 높아지게 되었으며, 이에 공공 및 민간에서는 빅데이터 활용 시 참고할 수 있도록 '빅데이터 활용을 위한 개인정보 비식별화[51]' 관련 기술 활용 안내서[52] 및 사례집[53]이 발간되었다.

한편 금융권의 빅데이터 활성화 제약요인으로 ①(법령상 제약) 신용정보법령상 불명확한 규정 등으로 인해 금융회사 등은 개인 신용정보 활용이 어려움 ②(인프라 미흡) 핀테크 기업은 금융상품을 만들고 새로운 서비스를 제공하기 위해 필요한 금융정보의 확보가 어려움 ③(지침 미비) 금융회사가 정보를 비식별화할 때 이에 대한 명확한 지침이 없어 비식별화 정보 활용에 주저 등의 크게 3가지 요인[54]이 존재한다.

금융권에서도 산업과 법의 준수의 조화로운 방향이 모색되다가 최근 금융위의 '금융권 빅데이터 활성화 방안'[55]에서 ①신용정보 범위 명확화 ②비식별정보 활용가능여부 명확화를

51) 비식별화란 정보에 포함되어 있는 개인정보의 일부 또는 전체를 삭제하거나 다른 정보로 대체함으로써 다른 정보와 결합하여도 특정 개인을 식별하기 어렵도록 하는 일련의 조치이다. 원칙적으로 그 자체로 개인을 식별할 수 있는 정보*는 삭제(또는 개인을 식별할 수 있는 정보의 삭제처리 대신 다른 정보로 대체)한다.
* 그 자체로 개인을 식별할 수 있는 정보 예시
① 쉽게 개인을 식별할 수 있는 정보(이름, 전화번호, 주소, 생년월일, 사진 등)
② 고유식별정보(주민등록번호, 운전면허번호, 외국인등록번호 여권번호)
③ 생체정보(지문, 홍채, DNA 정보 등)
④ 기관, 단체 등의 이용자 계정(등록번호, 계좌번호, 이메일 주소 등)

52) 미래창조과학부, 한국정보화진흥원(NIA), 빅데이터전략센터(KBiG), "빅데이터 활용을 위한 개인정보 비식별화 기술 활용 안내서", Ver 1.0, 2014.5.8

53) 미래창조과학부, 한국정보화진흥원(NIA), 빅데이터전략센터(KBiG), "빅데이터 활용을 위한 개인정보 비식별화 사례집", 2014.5.1

54) 금융위, "금융권 빅데이터 활성화 방안" 중 빅데이터 활성화 제약요인, 2015.6.3

55) 금융위 보도자료, "빅데이터를 활성화하여 금융회사와 핀테크 기업의 동반성장 토대 구축", 2015.6.3

통해 법령상 제약요건이 어느 정도 해소되었다.

신용정보 범위 명확화를 위해 시행령에서 비식별정보는 개인신용정보에서 제외[56]하고, 비식별정보 활용가능여부 명확화를 위해 개인 정보보호법에 따라 비식별화할 경우 동의 목적 외 이용이 가능하다고 유권해석[57]을 하였다. 또한 정책적으로 빅데이터 활성화 인프라 구축을 위한 정책으로써 신용정보법 개정으로 기존 5개 협회의 신용정보집중기관이 종합신용정보집중기관[58]으로 통합('16.3월까지 통합 완료)되어 금융권, 핀테크 기업 등의 빅데이터 업무 활용을 지원하는 역할을 수행할 예정이다.

하지만 금융권에서 개인 및 신용정보는 위와 같은 법률상 해석으로 모두 해결되는 것이 아니라, 빅데이터 분석 등을 위하여 비식별화된 정보를 활용하는 경우에는 접근통제, 관련 정보의 추가 이용 제한 등 비식별화 처리 전에 보유한 개인관련 정보를 활용 · 연계하여 개인을 식별할 수 없도록 내부 규정 등을 보완해야한다. 왜냐하면 비식별화된 정보를 비식별화 처리 전에 습득한 개인관련 정보와 매칭하여 사용하는 경우 개인정보의 목적 외 이용에 해당될 수 있기 때문이다.

나. 이상금융거래탐지 관련 이슈

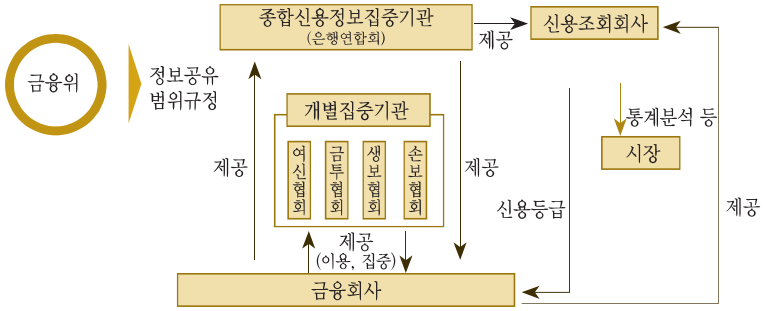
빅데이터 활용 이외에도 머신러닝은 관련 법적 이슈는 FDS에서도 발생한다. 정확한 사기탐지를 위해 FDS에서 머신러닝을 활용 시 빅데이터 활용에서처럼 개인정보를 모두 비식별화 처리를 할 수 없는 문제가 존재한다.

FDS에서 수집 정보는 ①금융거래유형 정보, ②사고유형 정보, ③이용자 매체환경 정보가 있으며, 수집되는 이용자 매체환경 정보는 아래 [표 8]과 같다.

56) 신용정보법 : 신용정보의 구체적인 범위를 시행령에서 정하도록 위임한다.
시행령 초안 : 식별성이 전제되지 않은 개별 거래내용, 신용도, 신용거래능력 판단정보를 개인신용정보 범위에서 제외한다.(9.12일 시행예정)

57) 신용정보법(특별법)에 규정되지 않은 사항은 개인정보보호법이 적용된다.

58) 신용정보집중기관은 신용정보를 금융회사, 신용조회회사에 제공하고, 금융회사는 신용정보집중기관의 정보 등을 활용하여 여신심사, 보험계약 인수 등에 활용하여 신용조회회사는 신용정보를 분석 · 가공하여 금융회사 등에 판매한다.



[표 8] 수집되는 이용자 매체환경 정보

구분	수집정보	
	PC계열	스마트폰 계열
하드웨어 정보	· 물리적 MAC정보 · HDD 정보(S/N, 모델 등) · CPU정보(코어〈cpu core〉 수 등) · 메인보드 정보(제조사, Product Name, Product S/N 등) 등	· UUID(Universally Uniue IDentifier) 정보 · 디바이스 모델명 등
OS 및 애플리케이션 정보	· 가상화 소프트웨어 사용 정보 · 브라우저 정보(종류, 언어 등) 등	· OS 버전 정보 · 제조사 정보 등
네트워크 정보	· IP정보(공인/사설, 국가, 지역 등) · Proxy IP 정보(설정여부, 국가 등) · VPN 정보(설정여부, 국가 등) 등	· 연결된 네트워크 정보 등

하지만 FDS에서 “개인정보 수집 및 활용”은 다른 법률과의 충돌문제가 존재한다. 즉, 이상금융거래 정보 중 개인정보 수집 및 활용은『전자금융거래법』제22조에서는 ‘전자금융기록 보관’에 근거하고 있으나,『개인정보보호법』,『위치정보보호법』,『정보통신망법』등에서 개인정보 수집 및 활용을 제한하고 있는 실정이다. 금융권에서 이상금융거래 정보 수집 관련 법률은 [표 9]와 같다.

[표 9] 이상금융거래 정보 수집 관련 법률

법	조항
전자금융거래법	· 제22조(전자금융거래기록의 생성 · 보존 및 파기) 제1항
개인정보보호법	· 제15조(개인정보의 수집 · 이용) 제1항 1호 · 제16조(개인정보의 수집 제한) 제1항, 제2항
위치정보법	· 제12조(이용약관의 신고 등) 제1항 · 제15조(위치정보의 수집 등의 금지) 제1항 · 제18조(개인위치정보의 수집) 제1항
정보통신망법	· 제22조(개인정보의 수집 · 이용 동의 등) 제1항

또한 금융권에서 해석될 수 있는 대표적인 현행 법률기반 개인식별정보는 [표 10]과 같다.

[표 10] 법률 기반 개인 식별 정보

구분		법률 근거	개인 식별 정보 항목
일반		· 개인정보보호법 제18조, 제23조, 제24조 제1항, 제24조 제3항	· 주체자의 사생활을 침해할 수 있는 식별정보(예: 의료정보, 정신적 성향 등) · 주체자의 신분확인을 위한 일반 식별정보(예: 이름, 주민번호, 주소 등)
민간부문	정보통신	· 전자금융거래법 제25조	· 휴대폰 결제서비스 수행을 위한 식별정보(예: 결제수단별 개인정보, 카드번호, 비밀번호 등)
		· 위치정보보호법 등	· 업무 수행 및 처리를 위한 통신상의 식별정보(예: 접속 IP정보, GPS 정보 등)
	상거래	· 전자문서 및 전자거래기본법 제12조 · 정보통신망법 제23조, 제24조 · 전자상거래 등에서의 소비자보호에 관한 법률 제12조	· 전자문서 서비스를 위한 식별정보(예: 공인전자주소, 송신자, 수신자 등) · 통신의 안전한 조치를 위해 확인할 수 있는 식별정보(예: 비밀번호, 계좌번호, 주민등록번호 등) · 거래 기록 및 배송을 확인하기 위한 식별정보(예: 배송주소지, 수령인 연락처 등)
		· 전자서명법 제24조	· 정당한 사용자임을 인증하는 식별정보(예: 가입자 이름, 전자서명검증정보, 인증서 일련번호)
	금융·신용	· 신용정보의 이용 및 보호에 관한 법률 제33조 · 금융실명거래 및 비밀 보장에 관한 법률 제4조	· 신용정보 및 거래능력을 판단할 수 있는 식별정보(예: 재산, 소득, 대출 보증 등) · 금융기관의 거래내역을 판단할 수 있는 정보(예: 주민등록번호, 계좌번호, 거래실적 자료 등) · 이용자 및 거래내용의 정확성을 확인하기 위한 식별정보(예: 전자금융업자에 등록된 이용자번호, 이용자의 생체정보 등)
		· 전자금융거래법 제26조 · 전자금융 감독규정 제5조의 3	· 신용정보 및 거래능력을 판단할 수 있는 식별정보(예: 재산, 소득, 대출 보증 등) · 금융기관의 거래내역을 판단할 수 있는 정보 (예: 주민등록번호, 계좌번호, 거래실적 자료 등) · 이용자 및 거래내용의 정확성을 확인하기 위한 식별정보(예: 전자금융업자에 등록된 이용자번호, 이용자의 생체정보 등)
		· 특정 금융거래 정보의 보고 및 이용 등에 관한 법률 제5조의 3	· 자금이체를 수행을 위한 식별정보(예: 송금인 성명, 계좌번호, 수취인의 정보)

이상금융거래 정보 수집 관련 법적 이슈에 대해서는 ①금융회사 등이 의무를 이행할 수 있는『전자금융거래법』상에서 '개인정보 수집 및 활용' 관련 예외조항 추가 등 법령의 통일적인 개정 여부에 대해 지속적으로 검토 및 모니터링이 필요하다.

또는 ②다른 법률인『개인정보보호법』,『위치정보보호법』,『정보통신망법』등에서 이상금융거래 정보 관련한 '개인정보 수집 및 활용' 항목의 개정에 대한 심도있는 검토가 필요하다고 할 수 있다.

Ⅵ. 결론

본고에서는 머신러닝 관련 동향을 파악하고 해당 기술 및 활용 사례들을 살펴보았다. 이에 머신러닝을 통한 금융권 스마트 서비스에서 금융회사가 해결해야 할 몇 가지 과제 및 시사점에 대해 크게 3가지를 제시하는 바이다.

첫째, 빅데이터 시대에 금융회사의 비즈니스의 목적과 규모에 맞추어 머신러닝 기술 활용을 통해 금융경쟁시장에서 돌파구를 찾아야 한다.

빅데이터 분석은 먼 미래에 도입을 고려할 분야가 아닌 현재의 이슈이며, 빅데이터 분석을 통한 가치창출을 이루기 위해 경쟁이 치열한 것은 사실이다. 특히 금융권은 타 산업 대비 기업의 데이터 보유량과 활용 잠재 가치가 높은 것으로 분석됨을 확인하였다. 증권 · 투자, 은행, 보험 순으로 데이터 보유량이 많으며, 특히 은행은 오디오, 비디오, 이미지 등의 비정형 데이터의 보유 비중이 증권 · 투자, 보험보다 높다.[59]

머신러닝을 활용한 빅데이터의 분석을 통해 스마트 금융서비스의 제공은 작년부터 시작된 핀테크 열풍과 함께 금융서비스의 패러다임 변화에 기여할 것이다. 이는 모든 금융회사들이 핀테크 서비스를 시작해야한다는 뜻이 아니라, 머신러닝과 같은 기술 분야를 얼마나 적절한 방법으로 현업과 금융서비스에 적용시키느냐가 경쟁력의 핵심인 것이다.

회사가 이미 보유한 트랙잭션 데이터를 새로운 방법으로 활용하여 이용자와 기업 모두에게 유용한 서비스를 제시하되 이는 기술 활용을 통해 이루어질 수 있으며, 지금부터 머신러닝 기술 활용을 위한 기술적 토대 마련과 역량 강화가 필수적이다. 또한 금융회사는 머신러닝을 활용한 빅데이터의 막대한 활용가치에 대한 공감대를 형성하고 장기적인 로드맵의 수립 등 기술 역량의 단계적 배양이 필요한 시점이다.

현재 기본적인 수준에 머물러 있었던 머신러닝의 활용은 빅데이터의 출현과 이를 처리할 수 있는 기술(정보처리 능력의 향상 및 머신러닝 알고리즘에 대한 연구)과 인프라가 구축됨에 따라 그 실현 가능성이 점차 높아지고 있다.

특히 딥러닝 기술은 신경망을 어떻게 디자인하고 어떤 종류의 신경망과 연결하느냐에 따라 과거 상상할 수 없었던 서비스를 개발해 낼 수 있는 기회를 제공하고 있다.[60] 구글이나 마이크로소프트 같이 압도적인 데이터 양과 컴퓨팅 파워를 가진 업체들이 딥러닝

59) KB금융지주 경영연구소, “KB daily 지식 비타민: 빅데이터(Big Data)의 이해와 금융업에 대한 시사점”, 2012-68호, 2012.6.27

60) 머니투데이 뉴스, “[딥러닝②] 상상 속 기계가 스스로 학습 기계로”, 2015.3.9

기술을 활용할 때 국내 업체들이 어떻게 경쟁력 우위를 확보할 수 있을지 고민해야 할 것이다.

둘째, 머신러닝을 이용한 데이터 사이언스의 운영 사이클을 이해하고 당면한 비즈니스 문제에 잘 활용할 수 있어야 한다.

머신러닝은 데이터의 숨겨진 가치를 찾는데 유용한 기술이다. 이러한 머신러닝은 잘 활용한다면 당면한 비즈니스 문제를 해결하고 장기적으로는 신규 가치 창출을 위한 방법이 될 수 있다.

하지만 머신러닝을 통한 데이터 분석 등은 당면한 문제를 해결하기 위해서 활용할 수 있는 방법들 중의 하나일 뿐이며, 전적으로 모든 문제에 대한 예측이나 분석을 할 수는 없다는 점을 유의하여야 한다. 우선적으로 금융회사가 현재 가지고 있는 데이터에 대한 분석과 데이터 관리 체계 등을 파악하는 과정을 통해 조직의 정보체계를 정확히 이해하고 비즈니스 적용 운영원리를 파악하여야 한다. 운영되는 데이터 안에서 문제의 답을 찾고 해당 데이터를 최적화할 수 있는 방법을 모색할 수 있어야 한다. 즉, 머신러닝을 통한 영업 및 마케팅 금융서비스 모델, 위험 및 사기 관리 등 많은 비즈니스 모델의 활용 및 시스템 구축 시 전체 운영 라이프 사이클을 이해하고 이후 관리 및 최적화 작업이 중요하다.

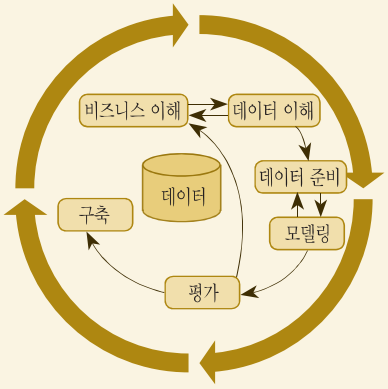
기업에서는 머신러닝을 산업에 활용하여 대규모의 데이터 마이닝을 효율적으로 수행하기 위해서는 여러 산업에 적용 가능한 데이터 마이닝 표준 프로세스(Cross Industry Standard Process for Data Mining, 이하 CRISP-DM이라고 한다.) 방법론[61] 등을 참고하여 비즈니스의 이해부터 시작하여 데이터의 분석, 적합한 모델링을 선정 및 적용까지 적절한 단계를 수행하여야 한다.(〈참고 1〉 참조)

61) SPSS, NCR, Daimler-Chrysler 등 여러 업계의 선도 회사들이 데이터마이닝 작업의 표준화를 연구하여 발표한 포괄적인 데이터마이닝의 방법론이자 프로세스로서 현재 전 세계의 데이터마이닝 프로젝트의 40%이상이 CRISP-DM의 프로세스에 따라 실행되고 있을 정도로 보편화된 data mining의 한 방법론이다.

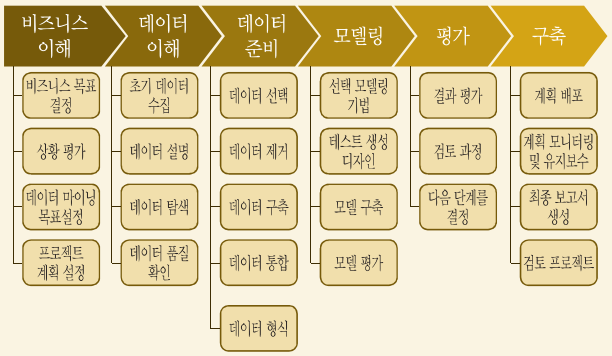
〈참고 1〉 머신러닝과 CRISP-DM 방법론에 따른 데이터 사이언스의 운영 사이클

여러 산업에 적용 가능한 포괄적인 데이터마이닝의 방법론이자 표준 프로세스인 CRISP-DM 방법론은 머신러닝의 활용과 함께 대규모의 데이터마이닝을 효율적으로 수행할 수 있도록 도와준다. CRISP-DM은 총 여섯 단계로 구성되어 있으며([그림 11], [그림 12] 참조), 각 단계별로 수행해야할 작업 내용이 정의되어 있다. ([표 11] 참조)

[그림 11] CRISP-DM의 운영 사이클



[그림 11] CRISP-DM의 운영 사이클



[표 11] CRISP-DM의 단계별 작업 내용

단계	내용
비즈니스의 이해 (Business Understanding)	· 당면한 해당 비즈니스 문제와 관련 비즈니스 프로세스에 대한 기본적인 이해가 필요한 단계 · 이 과정에서 데이터마이닝으로 접근할 수 있는 문제를 파악하는 단계
데이터의 이해 (Data Understanding)	· 현업이 보유 및 관리하고 있는 데이터를 이해하는 단계 · 한 조직의 정보체계를 정확히 이해하는데는 많은 시간이 소요
데이터의 준비 (Data Preparation)	· 정확한 예측과 진단값을 제공하는 가장 연관성 있는 데이터에 접근 · '데이터 준비'와 데이터가 머신러닝 시스템에 적용되는 방식과 관련 깊은 '피처(Feature) 엔지니어링' 단계를 거침
모델링 (Modeling)	· 동일한 데이터 마이닝 문제 유형을 위해 다양한 모델링 기법이 선택되어 적용 · 데이터 형태상의 특정 요구사항을 갖고 있으며, 적합한 모델링 기법 적용을 위해 데이터 준비 단계로 되돌아가는 것이 종종 필요함
평가 (Evaluation)	· 생성한 모델이 잘 해석되는지, 독립적인 새 자료에 적용은 얼마만큼 시킬 수 있고, 재현 가능한지를 알아보는 단계 · 데이터 분석 관점으로부터 높은 품질을 갖는 모델을 만듦
구축 (Deployment)	· 검토가 끝난 모형을 현업 비즈니스 인프라에 적용하는 단계 · 머신러닝에서 도출한 패턴을 비즈니스 환경에 적용하기 위한 목적으로 실행

자료 : Pete Chapman(NCR), Julian Clinton(SPSS), Randy Kerber(NCR), Thomas Khabaza(SPSS), Thomas Reinartz (DaimlerChrysler), Colin Shearer(SPSS) and Rudiger Wirth(DaimlerChrysler), "CRISP-DM 1.0" 재구성

셋째, 법적 허용범위 안에서 기술적 역량을 키워야 한다.

금융권은 이전에 과도한 규제에 비해 최근 규제개혁의 움직임으로 법적 허용 범위 안에서 핀테크 열풍과 함께 간편결제, 인터넷전문은행, 관계기반 신용대출 등 크게 변화의 물결을 맞이하고 있다.

최근 빅데이터 활성화 방안으로 인해 법령상 제약요건이 어느 정도 해소되었으나, 금융권에서 개인 및 신용정보는 위와 같은 법률상 해석으로 모두 해결되는 것이 아니다. 또한 비식별화와 익명성, 비식별화된 정보를 개인정보로 취급할 것인가의 문제(다른정보와 쉽게 결합해 개인을 식별할 수 있는 개인정보의 식별성의 쉬운 정도) 등 아직 정보의 활용에 있어서 해결해야할 사항들은 많이 존재한다.

또한 머신러닝을 적용하여 빅데이터 분석을 통해 금융서비스를 제공하는 사례가 증가하고 있지만, 국내는 빅데이터 활용이 중요한 관건임에도 불구하고 법적 제약, 기술적 역량의 부족 등으로 제대로 활용되지 못하고 있다.

금융회사들은 정보와 데이터의 특성을 잘 이해하고 분석하여 IT정보들을 활용할 수 있어야만 단순 해당 법률의 해석이 아닌 법적 허용 범위 안에서 기술적 역량을 키울 수 있는 조건이 마련될 수 있을 것이다. 그래야만 머신러닝의 활용에 있어 개인정보 유출 등 보안사고에 대해 금융기관의 감독과 규제만 면피하는 것이 아니라 금융 고객에게 신뢰를 줄 수 있는 것이며, 금융회사에서 빅데이터의 활용이 심화될수록 개인정보보호와의 조화로운 균형 속에서 IT 정보들을 관리할 수 있는 기술적 역량을 키워나갈 수 있을 것이다.

〈참고문헌〉

[1] Alpaydin Ethem, Introduction to machine learning, 2nd ed., MIT Press, 2010.

[2] SAS Institue Inc, An Overview of Machine Learning with SAS® Enterprise Miner™

[3] 이재구 외 2명, Big Data 분석을 위한 Machine Learning, 한국통신학회지(정보와 통신), 2014.10,

[4] 박혜영, 이관용, 패턴인식과 기계학습, 이한출판사, 2011.

[5] 클라우드 기반 머신러닝 서비스 각 홈페이지

[6] 미래창조과학부, 한국정보화진흥원, 빅데이터전략센터, 빅데이터 활용을 위한 개인정보 비식별화 기술 활용 안내서, Ver 1.0, 2015.6.10.

[7] 금융위 보도자료, “빅데이터를 활성화하여 금융회사와 핀테크 기업의 동반성장 토대 구축”, 2015.6.3

[8] 컴퓨터월드, [전문가 기고] 기계 학습, 디지털 비즈니스를 이끌다, 2015.1.31