기획 집중

데이터 마이닝의 이해 및 기계 설비 분야의 활용

최근 빅데이터를 비롯한 데이터 분석 기술인 데이터 마이닝에 대한 소개와 기계 제어 분야에서의 활용 가능성에 대해서 소개한다.

구원용
(주)이마이닝 대표
oais7koo@gmail.com

서 론

최근 딥 런닝(Deep Learning)이 이슈가 되면서 기계학습(Machine Learning)이 다시 주목을 받고 있다. 기계학습은 인공지능의 한 분야로, 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야로서, 데이터 마이닝(Data Mining)에서 많이 활용되고 있다. 최근에 많이 회자되고 있는 빅데이터 분석도 결국 데이터 마이닝의 분석 알고리즘을 하둡 등의 IT 분산 처리 기술과 접목해서 새로운 패러다임으로 등장했으며, 딥 런닝도 기존의 기계학습 알고리즘 중의 하나인 신경망 이론(Neural Network Theory)의 이론적 단점을 보완하며, 또한 빅데이터 분석의 대용량 데이터 처리 기술을 활용함으로써 새로운 패러다임으로 떠오르고 있다. 이러한 빅데이터-딥 런닝 패러다임의 시대를 이해하기 위해서 데이터 마이닝, 기계학습 이론 및 설비제어 부분에서 어떻게 활용되고 있는지에 대해서 소개하고자 한다.

데이터 마이닝과 기계학습

데이터 마이닝은 대량의 데이터로부터 새롭고 의미 있는 정보를 추출

해 의사결정에 활용하는 일련의 프로세스를 의미한다. 데이터를 채굴(mining)한다는 의미로서 숨어있는 의미 있는 데이터를 발견하기 위해서 대량의 데이터를 파헤치고 필터링하는 과정을 거친다. 데이터 마이닝은 일반적으로 대량의 데이터를 대상으로 하나, 반드시 대량의 데이터를 대상으로 하지 않더라도 비정형의 복잡한 데이터를 대상으로 하는 경우도 있다. 이렇게 데이터 마이닝은 대규모로 저장된 데이터 안에서 체계적이고 자동적으로 통계적 규칙이나 패턴을 찾아내는 것으로 KDD(Knowledge Discovery in Databases), 데이터 사이언스(Data Science)로도 불린다. 데이터 마이닝에서는 기계학습 방법론은 물론 통계 등 다양한 분야의 이론을 도입해서 활용을 하고 있기 때문에 그 경계가 모호하나, 결국 데이터 속에서 지식을 발견하기 위해서 다양한 알고리즘과 기법을 적용하는 일련의 프로세스를 지칭한다고 할 수 있다.

데이터 마이닝의 주요 기법은 아래와 같이 분류가능하다.

- 분류(classification) : 일정한 집단에 대한 특정한 정의를 통해서 대상 집단을 분류한다. 보통 설명변수(explanatory variable)와 목적변수의 페어로 이루어진 데이터 세트(data set)에 대해서 입력 데이터인 설명변수로 출력 데이터인 목적 변수를 예측하며, 이와 같이 예측 대상인 목적 변수가 있는 경우를 지도학습이라고 한다.
- 군집화(clustering) : 입력 데이터를 몇 개의 군집으로 분류하는 기법으로 설명변수는 존재하나, 목적 변수는 존재하지 않는다. 대표적인 기법으로 SOM 등이 있다.
- 연관 규칙(association rule) : 데이터의 항목들 간의 조건-결과 식으로 표현되는 유용한 패턴을 찾아내는 기법으로 대표적인 예는 고객이 구입한 쇼핑카트와 같이 한 액션을 통해서 구축된 일련의 그룹화된 정보인 바스켓 데이터(basket data) 즉, 트랜잭션(transaction) 데이터의 분석을 통해서 패턴을 추출한다.
- 추정(estimation) : 목적 변수가 있다는 점에서는 분류와 동일하나 주어진 설명변수로 연속형이나 수치형의 목적 변수를 예측하는 것이 분류와 다른 점이다.

데이터 마이닝의 기법은 크게 데이터 세트의 종류에 따라서 교사 학습(supervised learning)과 비교사 학습(unsupervised learning)으로 나뉠 수 있다. 교사 학습은 데이터 세트에 예측이나 분류하고자 하는 성질의 정답을 가지고 있는 경우로 교사 학습 관련 알고리즘에서는 예측 혹은 분류 모델이 정답률을 최대화하는 것을 목적으로 하며, 관련 알고리즘으로 회귀분석(regression analysis), 의사결정나무(decision tree), 신경망 분석(neural network analysis) 등이 있다. 비교사 학습은 정답이 없는 상태에서 데이터-설명변수 간 유사도를 최대화하는 군집을 구하는 것을 목적으로 하며, 관련 알고리즘으로 K-means, SOM 등이 있다.

산업에서의 데이터 마이닝 활용

데이터 마이닝 입문서에서 가장 많이 소개되는 케이스가 기저귀-맥주 관련 케이스이다. 앞에서 소개한 연관 규칙을 활용하여 마트에서의 소비자의 구매 패턴을 분석하던 중 우연히 기저귀와 맥주를 동시에 구입하는 소비자가 많다는 패턴을 발견하게 되고 이에 대해서 분석을 한 결과, 퇴근길에 아내에게 기저귀 심부름을 부탁받은 남편이 기저귀를 구입하면서 자신이 마실 맥주를 같이 구입한다는 패턴을 발견하고, 기저귀와 맥주를 인접 진열대

에 진열함으로써 매출 상승의 효과를 얻었다는 내용으로 데이터 마이닝의 효과를 설명하는데 가장 많이 사용되고 있다. 여기에서 데이터 마이닝과 기계학습의 차이가 나타나는데, 기계 학습은 기저귀와 맥주의 상관관계가 높다는 분석 결과를 정량적으로 얻음으로써 분석이 완료된다고 볼 수 있는 공학적 분석이라고 본다면, 데이터 마이닝에서는 그러한 분석 결과가 왜 나타나는지에 대한 사회, 인문학적 분석까지 포함을 하는 보다 포괄적인 개념이라고 볼 수 있다.

이러한 데이터 마이닝 및 기계학습은 많은 분야에서 활용되고 있다. 초창기의 필기 인식에서 시작하여 최근에는 안면, 지문 인식 등의 생체 인식과 차량 번호판 인식 등 다양하게 활용되고 있는 영상 인식 분야와 로봇 AI를 위한 컴퓨터 비전과 자연어 처리, 음성인식, 자동차 자율 주행 등 거의 모든 산업분야에서 활용이 되고 있다. 현재 우리는 IT의 시대에서 데이터의 시대로 넘어가는 과도기에 서 있다고 볼 수 있다.

기계/제어부분에서 데이터 마이닝 및 기계학습은 많은 분야에서 활발하게 활용되고 있다. 과거 출퇴근 시간 등의 이용객 수가 많은 시간대에 대한 자가학습을 통해서 상시 운영을 할 것인지 혹은 이용객이 있을 경우에 한해서 운영을 할 것인지, 운영 스케줄을 이용객 정보를 활용해서 학습하는 단순한 에스컬레이터 자율 운행에 신경망 이론 및 대기행렬 이론(queueing theory)이 이용되었다. 최근에는 유비쿼터스 기반 인텔리전트 빌딩의 운영을 통해서 축적된 데이터에 기반한 기계 학습 모델에 의해서 경제적이며, 안정적인 최적운영 제어시스템 개발에 많은 관심이 집중되고 있다.[1] 또한 시설물 유지 보수에도 데이터 마이닝이 활용되고 있는데, 대표적인 예로 글로벌 엘리베이터 회사가 전 세계에 시공/운행 중인 엘리베이터에 부착된 센서들을 통해 모터의 온도, 속도 등의 각종 정보들을 실시간으로 수집해 엘리베이터가 고장 나기 전에 수리할 부분을 미리 찾아내기 위해서 머신 런닝을 도입해서 80%의 서비스 정확도를 달성하고 있다. 또한 항공분야에서 기체의 형태, 크기, 무게 등의 정적 및 동적 변화에 따라 스스로 비행 계수를 조정하여 목표 비행 궤적을 정확하게 따라가도록 자율제어를 하는 기계학습 모듈을 추가한 무인기의 자동 비행 제어 시스템 개발 등에 기계학습이 활용되고 있다.[1)]

최근에는 사물인터넷(Internet of Thing, IoT)이 급부상하고 있다. 현재의 컴퓨터나 스마트폰은 사용자의 입력에 의해서 작동되며, 사용자의 입력은 디바이스 내의 프로세서가 처리를 해서 결과를 스크린을 통해 사용자에게 보이며, 이러한 사용자 인터페이스(UI)를 통해서 프로세스가 진행된다. 사물인터넷 환경에서 사물인터넷의 구동 과정도 이와 유사하지만 프로세스가 진행되는 방식은 크게 다르다. 사용자의 입력은 필수적이지 않으며, 사물 인터넷의 하드웨어는 평소에 자율적으로 현실계의 정보들을 센서로 인식해서 정보를 인지하며, 사용자의 명령 없이도 정해진 프로세스에 의해서 정보를 처리한다. 정보를 처리한 이후 결과를 자율적으로 분석해서 이후 프로세스에 피드백하는 자율판단 및 운영/제어를 수행하는 독립적인 시스템으로 존재할 수 있게 되며, 최근 화제가 되고 있는 구글-애플-MS의 스마트카 경쟁은 사물인터넷 시대의 도래에 대한 전조라고 볼 수 있다.

기계설비 분야에서의 데이터 마이닝의 활용

사물인터넷 시장이 활성화됨에 따라서 특히 기계설비 분야에서 센서 데이터를 중심으로 데이터 마이닝 및 빅데이터 분석 기법을 도입하려고 하는 시도가 활발해지고 있다. 복잡하고 고가인 장비를

1 마이크로소프트의 애저 머신 런닝 서비스, 삼성SDS ICT 스토리 (http://www.ictstory.com/800).

유지 보수하기 위해서 최근 몇 년 사이에 초 단위 혹은 밀리초 단위로 장비 상태를 모니터링하기 위해 항공기 엔진부터 탱크까지 모든 장비에 내장 센서가 이용되기 시작했다.[5)]

엔진을 예로 들 경우, 센서는 온도, 분당 회전수, 연료 소모율, 오일 압력 상태에 이르는 모든 정보를 원하는 간격으로 수집 가능하다. 측정 빈도, 측정 지표의 개수, 모니터링할 장비의 개수가 늘어남에 따라 데이터양은 기하급수적으로 늘어날 것이다. 이러한 센서 데이터는 특히 개발 단계에서 중요한 역할을 하는데, 엔진은 고온에서 작동을 해야 하며, 다양한 작동 조건을 경험하며, 지속적이고 안정적인 성능을 내는 것이 중요하기 때문에, 장비가 중단되는 시간을 최소화하는 것이 지극히 중요하다.

이러한 중단 시간을 단축하는 방안으로 유지보수가 필요한 장비에 신속하게 교체할 수 있는 여분의 부품이나 엔진을 상비해 놓거나, 교체해야 하는 부품을 신속하게 확인할 수 있는 진단법을 개발하거나, 문제가 발생한 부품보다 신뢰성이 높은 부품으로 교체하는 전략이 있다. 이러한 전략은 모두 데이터가 있어야만 효과적으로 시행할 수 있으며, 최근에는 이러한 센서 데이터를 진단 알고리즘을 이용해서 처리함으로써 장비의 유지 보수는 물론 기존 제품의 개량 및 신제품 개발에 기여할 수 있는 기술개발이 이루어지고 있다.

장비의 운영에서 나오는 센서 데이터 중 고장직전 데이터를 축적하고 데이터 패턴을 발견함으로써 성능저하-고장-수리로 이어지는 사이클에 대한 예측을 가능하게 하고, 이를 통해서 미연에 고장을 방지할 수 있는 선수적인 조치를 취할 수 있게 된다.

최근 GE는 이러한 산업 기계에서 발생하는 대규모 데이터를 수집/분석하는 산업 클라우드 솔류션인 프레딕스 클라우드(Predix Cloud)[2]의 출시 계획을 발표하였다. 프레딕스 클라우드는 소프트웨어 애플리케이션 개발을 지원하는 서비스형 플랫폼(PaaS)으로서 GE는 향후 자사의 소프트웨어 분석업무뿐만이 아니라 데이터 관리와 응용을 필요로 하는 기업 고객들을 대상으로 상용화를 계획하고 있다.

이러한 서비스를 통해서 일반 기업은 빅데이터 분석 시스템을 개별적으로 구축함으로써 발생하는 비용을 최소화하고 전문적인 서비스를 통해서 높은 수준의 분석 기법을 제공 받음으로써 제품의 효율 향상을 꾀할 수 있게 된다.

데이터 마이닝 및 기계학습 SW

데이터 마이닝 및 기계학습의 중요도가 증가함에 따라서 이를 지원하기 위한 SW도 상용 및 오픈 소스를 중심으로 다양하게 개발되어 왔다. 상용으로는 Microsoft Azure,[3] matlab,[4] SPSS-Modeler[5] 등이 있고, 오픈 소스로는 파이썬 기반 Orange,[6] 자바 기반의 Weka,[7] R[8] 등이 제공되고 있다. 특히 Microsoft Azure는 SaaS(Software as a Service)로 제공되어서 사용자 입장에서 따로 설치할 필요 없이 웹을 통해서 간단하게 서비스를 받을 수 있는 장점이 있으며, 오픈 소스 SW의 경우, 관련 연구자들이 연구성과가 빠르게 SW에 피드백됨으로써 최신 알고리즘을 상용 SW보다 빠르게 사용할 수 있는 장점을 제공하고 있다. R의 경우 CRAN에 약 7,000개의 패키지가 관리되고 있으며, 이 중 많은 패키지가 데이터

2 http://www.ge.com/digital/predix.

3 https://azure.microsoft.com/ko-kr/services/machine-learning/.

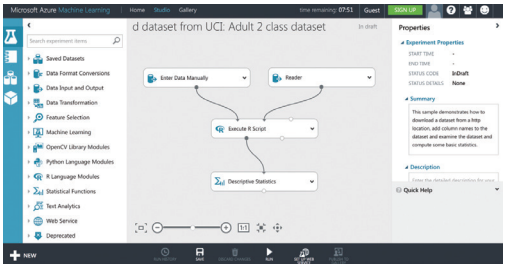
4 http://kr.mathworks.com.

5 http://www-03.ibm.com/software/products/ko/spss-modeler.

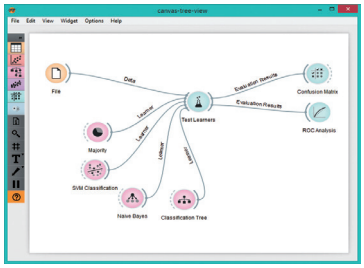
6 http://orange.biolab.si.

7 http://www.cs.waikato.ac.nz/ml/weka/.

8 https://cran.r-project.org.



(a) Microsoft Azure Machine Learning



(b) Orange Data Mining Tool

[그림 1] Data Mining Tools

마이닝 및 기계학습과 관련이 있으며 상시 업데이트되고 있다(**그림 1**).

파이썬에서는 사이킷-런(Scikit-learn)[9]이 유명하다. 수학과 과학 분야에서 넘피(NumPy), 사이피(SciPy), matplotlib 등의 패키지를 폭넓게 이용하며, 결과물인 라이브러리는 인터랙티브 워크벤치 애플리케이션에 이용하거나, 다른 소프트웨어에 탑재시켜 재사용할 수 있다.

다른 기계학습 라이브러리로 쇼군, MLlib 등이 있으며, 가장 오래된 기계 학습 라이브러리들 가운데 하나인 쇼군(Shogun)은 1999년에 개발됐으며, C++에 기반을 두고 있지만, SWIG 라이브러리 덕분에 자바, 파이썬, C#, 루비, R Lua, Octave, Matlab에서도 이용할 수 있다.

최근에는 설치형 SW가 아닌 SaaS,[10] PaaS[11] 형태의 툴들도 많이 제공되고 있다. 앞에서 소개한 GE의 프레딕스가 산업 부분에 특화된 PssS라면 Microsoft Azure는 범용적인 PaaS로서 제공 기능 중 빅데이터 분석, 기계학습, 스트리밍 분석 등 다양한 기능들을 제공함으로써 고객이 손쉽게 각종 기업 데이터를 분석할 수 있는 서비스 환경을 제공하고 있다.

맺음말

현재 기계 및 제어 분야에 데이터 마이닝 및 기계학습 알고리즘의 활용에 관한 다양한 시도가 이루어지고 있으며, 기계 설비 제어 및 운영 최적화 등에서 성과를 나타내고 있다. 특히 사물인터넷이 활성화됨에 따라서 데이터 마이닝 및 기계학습의 중요도가 더욱 높아질 것으로 예상되며, 글로벌 기업들의 산업 데이터 분석 서비스 시장이 활성화됨으로써, 각 기업들은 이에 대해서 보다 적극적인 기술 도입의 검토가 필요하다. 제품의 가치를 높이기 위한 노력은 하드웨어뿐만이 아니라 그 하드웨어에 내재된 소프트웨어의 가치도 하드웨어만큼이나 중요하다는 것을 굳이 애플의 지속적인 성장과 애플 생태계를 예를 들지 않아도 자명하다는 점에서 향후 데이터 마이닝과 기계학습 혹은 빅데이터와 딥런닝 분야의 활성화를 기대해본다.

참고문헌

1. 문미선, 송강, 송동호, 2010, 기계학습 알고리즘을 이용한 UAS 제어계수 실시간 자동 조정 시스템, 한국항행학회논문지, Vol. 14. No. 6.
2. 김석태, 2015, 영상인식의 이해, 한국학술정보.

9 https://github.com/scikit-learn/scikit-learn.

10 Software as a Service의 약자로, on-demand software로 불리우며 소프트웨어 및 관련 데이터는 서비스 프로바이더에서 호스팅되고, 사용자는 웹 브라우저 등의 클라이너트를 통해서 접속하는 형태의 서비스.

11 Plantform as a Service로 Saas의 개념을 개발 플랫폼에 확장한 방식으로 개발을 위한 플랫폼 구축을 할 필요없이 필요한 개발 요소를 웹에서 빌려 쓰는 개념의 서비스.

3. 양순옥, 김성석, 정광식, 2015, 유비쿼터스컴퓨팅 개론, 한빛미디어.
4. 이안 위튼 외, 이승현 옮김, 2015, 데이터 마이닝(Data Mining), 에이콘출판.
5. 빌 프랭크스, 2014, 빅데이터에서 천금의 기회를 캐라, 에어콘.