기계학습을 고려한 원전 빅데이터 시스템

박재관*, 김택규*, 장귀숙*, 성승환*, 구서룡*
*한국원자력연구원 자율운전연구실
jkpark183@kaeri.re.kr, taekkyukim@kaeri.re.kr, gsjang@kaeri.re.kr, shseong@kaeri.re.kr, srkoo@kaeri.re.kr

A Practice of Nuclear Bigdata System for Machine Learning

Jaekwan Park*, TaekKyu Kim*, Gwi-Sook Jang*, SeungHwan Seong*, SeoRyong Koo*
Korea Atomic Energy Research Institute

요 약

원전 빅데이터를 효율적으로 분석하고 수집된 데이터를 인공지능 서비스에 활용할 수 있도록 제공하기 위해서는 원전 데이터에 특화된 빅데이터 플랫폼이 필요하다. 단순히 시간 순으로 나열된 원시(Raw) 데이터는 의미있는 단위로 논리적으로 구분되어 관리될 필요가 있고, 사건/사고의 발생에 따른 분류가 필요하다. 뿐만 아니라, 다수의 데이터들을 분석하여 수천 개의 계측신호들 중에서 원하는 목적에 적합한 신호가 어떠한 것들인지를 찾아낼 수 있는 데이터 분석이 지원될 필요가 있다. 이는 기계학습 애플리케이션을 개발할 때 필수적인 고품질의 데이터 제공에 크게 기여할 수 있다. 본 연구에서는 원전 데이터를 효과적으로 처리하고 분석하기 위한 원전 데이터 전처리 및 분석 기술을 고안하고 이를 빅데이터 저장 인프라와 통합한 원전 빅데이터 처리 체계를 소개한다. 본 연구의 결과물은 본격적인 원전 빅데이터 시스템 구축 사업에 활용될 것으로 기대된다.

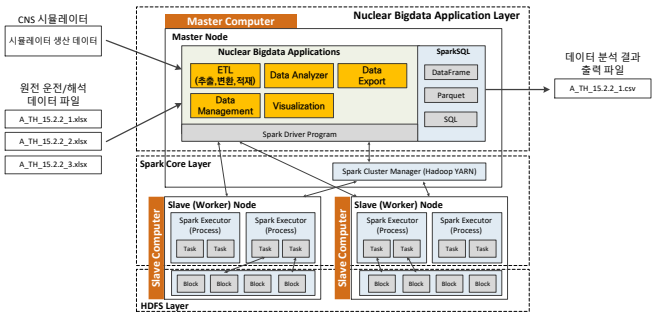
1. 서론

하둡(Hadoop)을 근간으로 하는 빅데이터의 분산 저장 및 병렬 처리 기술은 최근 IT 분야에서 신뢰도를 확보하였다. 신기술 적용에 보수적인 원전 산업도 기존의 DBMS 에서 빅데이터 시스템으로 변화될 것으로 전망된다. 최근, 딥러닝 기술이 발달하면서 인공지능 기술을 원전의 예측 및 진단에 적용하기 위한 다양한 연구가 진행되고 있다[1-2]. 따라서, 원전 빅데이터 시스템은 빅데이터의 저장뿐만 아니라, 인공지능 애플리케이션 연구개발에서의 활용을 고려한 데이터 처리 체계가 요구된다. 이를 위해, 원전 빅데이터 시스템은 원시(Raw) 데이터인 발전소 현장 및 시뮬레이션 데이터를 분석하여 양질의 의미있는 단위로 가공하여 저장 및 관리하는 것이 중요하다[3].

2. 원전 빅데이터 시스템의 구조

빅데이터의 효율적인 처리를 위해 개발된 오픈소스 하둡은 부족한 기능들을 보완하는 하둡 에코시스템들이 등장하면서 발전하고 있다. 최근, 파이썬 언어를 지원하고 머신러닝 모듈을 내재한 스파크 프레임워크가 빅데이터 처리 도구로써 확산되고 있다. 하둡과 스파크는 경쟁보다는 상호 보완적인 구조로 활용된다. 하둡 파일시스템(HDFS)은 분산 데이터 인프라 구조를 제공함으로써, 대량의 데이터를 클러스터 내 복수의 노드들에 분산시켜주는 역할을 수행한다. 스파크는 이러한 인프라 구조 상위에서 동작하는 데이터 프로세싱 툴로써 메인 메모리 내에서 빠른 데이터 처리를 지원한다.

본 논문은 3 계층의 빅데이터 시스템 구조를 소개한다. 가장 아래에 있는 1 계층 HDFS 는 하둡 분산 파일 시스템으로써 대용량의 데이터 파일을 분산 저장하는 역할을 담당한다. 2 계층인 스파크 코어는 스파크 기반의 Job/Task 의 병렬 처리를 담당한다. 가장 상단의 3 계층인 원전 빅데이터 애플리케이션 계층은 스파크를 활용하여 원전 데이터의 저장과 관리, 데이터 분석 및 가시화, 그리고 데이터의 외부 출력을 담당한다.

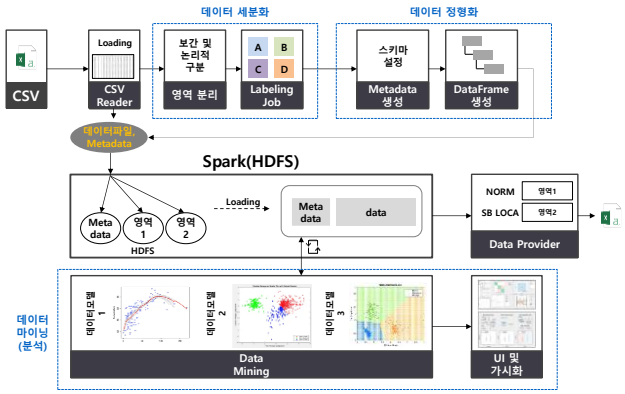


(그림 1) 3 계층의 원전 빅데이터 시스템 구조

3. 원전 애플리케이션 기능

원전 데이터의 정제, 분석, 고품질 데이터의 획득을 위해서는 원전 특성에 적합한 데이터 처리 기능이 원전 빅데이터 시스템에 내재화되어야 한다. 특히, 정상, 과도, 사고 등으로 구분되는 원전 상태에 따른 데이터 라벨링 기능, 타임라인 특성의 계측 데이터 보간(Interpolation) 기능, 루프 구간별 다수의 노달라이제이션(Nordalization)으로 구분되는 열수력 해석 데이터를 인식하고 처리, 공정 변수들 간의 상관관계를 분석할 수 있는 원전 데이터 마이닝 기능이 요구된다.

본 연구에서 도출한 전체적인 원전 빅데이터 처리 체계는 그림 2 와 같다. 먼저, 데이터 세분화는 Raw 데이터가 저장된 엑셀 데이터 파일을 로딩하여 원전의 상태가 변하는 이벤트 시점을 식별하여 논리적 영역을 구분하고 각 영역에 대한 세부적인 라벨(Label)을 부여함으로써 학습 데이터의 가장 중요한 부분인 라벨을 결정하는 기능을 수행한다. 데이터 정형화는 입력된 데이터 파일에 대한 스키마 정보를 생성하고 데이터 보간을 통하여 일정한 저장 형태를 결정하는 기능을 수행한다. 데이터 세분화 및 정형화가 완료된 후에는 원본 데이터 파일과 메타 데이터를 스파크(HDFS)에 저장한다.

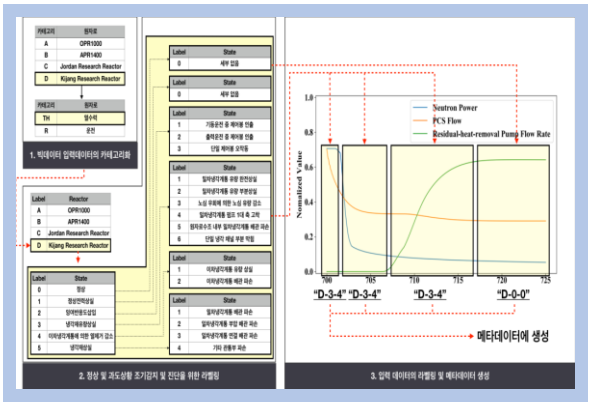


(그림 2) 원전 빅데이터 처리 체계

원전 빅데이터 마이닝 기능은 수천 개의 계측 변수들 중에서 인공지능 애플리케이션이 식별하려는 상태에 대하여 데이터들을 분석하여 각 변수들이 식별 대상 라벨에 기여하는 정도(Ranking)를 분석하여 최적의 학습 대상 변수들을 도출하는데 사용할 수 있다. kNN, Linear Regression, Random Forest, Correspondence Analysis, Distance Matrix, PCA 기법 등 다양한 방법들이 상황에 맞게 선택되어 분석에 사용될 수 있다. 또한, 이러한 분석 결과는 GUI 기반으로 쉽게 이해할 있는 형태로 표시되어야 하므로 마이닝 결과 가시화 기능도 동반된다. 스파크는 인메모리 데이터 처리를 지원하기 때문에 배치작업이 아닌 형태의 데이터 작업과 동일한 데이터를 여러 번 반복해서 엑세스하는 작업에 좋은 성능을 보인다. 데이터 마이닝에서는 다양한 데이터 분석 기법들을 적용하여 특정 그룹의 데이터 셋을 다수 엑세스하기 때문에 스파크는 빠른 분석 결과를 제공해준다. 데이터의 외부 제공 기능으로써 인공지능 학습을 위한 데이터의 출력이 있다. 원전 빅데이터 시스템에 잘 정제되어 쌓여있는 데이터로부터 사용자가 원하는 원자로 노형, 데이터의 종류(발전소 현장데이터, 시뮬레이터 데이터, 해석데이터), 상태 라벨을 지정하여 손쉽게 추출해주는 기능이다. 이와같이, 원전 빅데이터 시스템은 기존 빅데이터 인프라(하드웨어 및 소프트웨어)와 더불어 원전 특성에 맞는 데이터 처리 모듈들이 모두 포함되는 하나의 시스템으로 구성되는 것이 효율적이다.

3.1 원전 데이터 전처리 기능

원전 빅데이터 시스템의 전처리 단계의 정형화를 위한 데이터 포맷은 그림 3 과 같이 정의하였다. 이것은 원전의 열수력 모델을 이용해 추출된 원전 모사 데이터를 빅데이터에 저장 및 관리하고 데이터 분석을 통해 상관분석 결과를 사용자에게 제공하기 위한 데이터 포맷이다. 또한, 기타 세부 정보를 수록하여 사용자에게 다양한 정보를 제공하기 위해 여러 개의 시트로 입력데이터를 구성하므로 입력데이터 파일 포맷은 엑셀 파일 포맷(.xlsx)으로 정의하였다.



(그림 3) 입력데이터 전처리 과정

빅데이터 전처리 단계의 입력데이터는 Nodalization, Initial Condition, Event Time Table, Node Description, Pressure, Total Temperature, Void Fraction, Total Density, Mass Flow Rate, 그리고 Other 시트로 구성된다. Nodalization, Initial Condition 시트는 정보로써 열수력 데이터를 추출하기 위해 사용자가 사용한 열수력 모델과 초기조건 정보를 알 수 있다. Event Time Table 시트는 열수력 데이터가 시간 순으로 어떠한 이벤트들에 의해 생성되었는지를 사용자에게 정보로 제공한다. 또한, 이 정보는 시계열로 구성된 각 계측 데이터

(Pressure, Total Temperature, Void Fraction, Total Density, Mass Flow Rate 등)에 어떤 상태 라벨을 붙여야 하는지에 대한 정보를 제공한다. Node Description 시트는 빅데이터 시스템의 데이터 저장 및 관리에 사용할 스키마 생성하는데 사용된다. Pressure, Total Temperature, Void Fraction, Total Density, Mass Flow Rate, 그리고 Other 시트는 계측 값을 가지는 데이터 시트이다.

3.2 데이터 분석 기능

수천 개의 계측 변수 전체를 머신러닝에 사용하여 결과를 도출하면 정확도가 높아질 것이라고 기대하지만, 오히려 결과를 잘못되게 도출하는 경우가 많다. 이는 통계분석에서 선형함수의 독립변수가 많다고 해서 종속변수의 기댓값의 정확도가 항상 올라가는 것은 아닌 이유라고도 할 수 있다. 즉, 머신러닝의 성능은 어떤 데이터를 학습의 입력에 사용하는지에 굉장히 의존적이기 때문에 먼저, 충분한 데이터를 모으고 어떤 특징이 유용한지 아닌지 확인하는 과정을 거치는 것이 중요하다. 특징이 유용한지 아닌지 확인하는 과정을 특징선택(feature selection) 또는 특징 추출(feature extraction) 이라고 하며, 특징선택은 특징 랭킹(feature ranking) 또는 특징 중요도(feature importance)라고도 불린다. 특징 선택의 목적은 모든 특징의 부분집합을 선택하거나 불필요한 특징을 제거하여 간결한 특징 집합을 만드는 것이다[4].

본 연구에서는 수천 개의 계측 변수들을 목표로 하는 라벨, 즉 타겟과의 관련성을 측정하는 점수에 따라 정렬하는 기능을 제공한다. 이를 통해, 라벨과의 관련성이 높은 변수들을 확인할 수 있고 머신러닝에 입력으로 사용할 변수를 최적화할 수 있다. 기본적으로 피어슨 상관관계 기법을 사용하지만, 사용자가 적용할 분석기법을 선택할 수 있다.

(그림 4) 데이터 분석 결과의 예

그림 4 는 데이터 분석 예제를 보여준다. 특정 사고(A-TH-15-6-5)에 대하여 계측 변수들의 상관관계 점수에 따른 분석 결과이다. P_110(압력)부터 O_005(온도)까지가 유의미한 관계성이 있음을 확인할 수 있다. 즉, 이러한 유의미한 데이터만 추출하여 머신러닝에 활용한다면 학습의 효율성을 극대화할 수 있을 것으로 기대된다.

4. 결론

최근 원자력 분야의 산업계, 학계, 연구소에서도 인공지능 및 빅데이터와 같은 첨단 기술을 기반으로 예측 진단, 지능형 감시, 자율 제어를 실현하기 위한 많은 연구를 수행하고 있다. 특히, 딥러닝을 활용한 기계학습 기반의 인공지능 기술 확보에 많은 노력을 기울이고 있다. 이러한 인공지능 기술은 고품질의 데이터에 기반한 진단과 판단을 수행하는 것이 특징이다. 따라서, 이러한 첨단기술을 원자력 분야에 적용하기 위해서는 다양한 소스에서 확보되는 빅데이터를 의미있는 단위로 가공 및 정제하여 안정적인 분산저장 시스템에 저장하는 것이 중요하다. 본 연구에서는 원전 인공지능 기술에 선행되어야 할 원전 빅데이터 시스템을 위한 기반 기술을 연구하였다. 이를 위해 원전 빅데이터 시스템 구성을 정립하였으며, 원전 데이터에 대한 전처리 기능 및 데이터 분석 기능을 정의하였다.

Acknowledgement

이 논문은 과학기술정보통신부의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2018M2B2B1065652 and No. 2019M2C9A1055903).

참고문헌

[1] Yang, J., Kim, J., “An accident diagnosis algorithm using long short-term memory”, Nuclear Engineering and Technology, vol. 50, pp. 582–588, 2018

[2] T.-K. Kim et al., “deep-learning-based alarm system for accident diagnosis and reactor state classification with probability value”, Annals of Nuclear Energy, vol. 133, pp. 723–731, 2019

[3] 안성원, “빅데이터를 제대로 활용하기 위한 조건: 데이터 확보와 비즈니스 발굴”, 월간 SW 중심사회, 소프트웨어정책연구소, pp. 7-13, 2017.

[4] Mehul Ved, “Feature Selection in Machine Learning: Variable Ranking and Feature Subset Selection Methods”, Internet: medium.com/@mehulved1503, 2018