유전 프로그래밍을 활용한 제조 빅데이터 분석 방법 연구

(Genetic Programming based Manufacutring Big Data Analytics)

오상헌*, 안창욱**

(Sanghoun Oh, Chang Wook Ahn)

요 약

현재 제조 분야 빅데이터 분석을 위하여 black-box 기반 기계 학습 알고리즘을 활용하고 있다. 해당 알고리즘은 높은 분석 정합성 가지는 장점이 있지만, 분석 결과에 대한 해석이 어렵다는 단점이 있다. 그러나 제조업에서는 분석 알고리즘은 제조 공정 원리 기반 해석을 통하여 결과의 근거 및 도출 타당성에 대한 검증이 중요하다. 이러한 기계 학습 알고리즘의 결과 설명력 한계를 극복하기 위하여 유전 프로그래밍을 활용한 제조 빅데이터 분석 방법을 제안한다. 본 알고리즘은 생물학적 진화유전 프로그래밍 알고리즘은 생물학적 진화를 모방한 진화 연산 (선택, 교배, 돌연변이) 반복하면서 최적의 해를 찾아간다. 그리고 해는 수학적 기호를 활용하여 변수 간의 관계로 나타나며, 가장 높은 설명력을 가지는 해가 최종적으로 선택된다. 이를 통하여 입력 및 출력 변수 관계 수식화를 통한 결과를 도출하므로 직관적인 제조 매카니즘에 대한 해석이 가능하며 또한 수식으로 나타낸 변수간의 관계 기반으로 기존 해석이 불가한 제조 원리 도출도 가능하다. 제안 기법은 대표적인 기계 학습 알고리즘과 성능을 비교 분석 결과 동등 또는 우수한 성능을 보였다. 향후 해당 기법을 통하여 다양한 제조 분야 활용 가능성을 검증하였다.

■ 중심어 : 제조 빅데이터 ; 기계학습 알고리즘 ; 유전 프로그래밍

Abstract

Currently, black-box-based machine learning algorithms are used to analyze big data in manufacturing. This algorithm has the advantage of having high analytical consistency, but has the disadvantage that it is difficult to interpret the analysis results. However, in the manufacturing industry, it is important to verify the basis of the results and the validity of deriving the analysis algorithms through analysis based on the manufacturing process principle. To overcome the limitation of explanatory power as a result of this machine learning algorithm, we propose a manufacturing big data analysis method using genetic programming. This algorithm is one of well-known evolutionary algorithms, which repeats evolutionary operators such as selection, crossover, mutation that mimic biological evolution to find the optimal solution. Then, the solution is expressed as a relationship between variables using mathematical symbols, and the solution with the highest explanatory power is finally selected. Through this, input and output variable relations are derived to formulate the results, so it is possible to interpret the intuitive manufacturing mechanism, and it is also possible to derive manufacturing principles that cannot be interpreted based on the relationship between variables represented by formulas. The proposed technique showed equal or superior performance as a result of comparing and analyzing performance with a typical machine learning algorithm. In the future, the possibility of using various manufacturing fields was verified through the technique.

■ keywords : Manufacturing Big Data ; Machine Learning ; Genetic Programming ; Predicion

Ⅰ. 서 론

산업의 핵심 이슈로 부상하여 2012년 및 2013년 Gartner는 빅데이터를 10대 전략기술로 선정하였고, McKinsey에서도 2011년 빅데이터가 경제에 미치는 영향에 대한 보고서를 발표한 데 이어, 2013년 미국에 성장과 재도약의 기회를 제공할 다섯 가지 게임 체인저(game changer) 중의 하나로 빅데이터를 선정하였다[1,2]. 그리고 시장조사기관 IDC 결과에 따르면, 세계 빅데이터 시장은 2010년 32억 달러, 2013년 97억 달러, 2015년 169억 달러 규모로 연평균 약 40%씩 성장하였다[2,3]. 이는 빅데이터가

* 정회원, 한국 방송 통신 대학교 바이오-정보통계학과 대학원생
** 정회원, 광주과학기술원(GIST) AI대학원 교수
이 논문은 2020년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초 연구 사업임 (No. NRF-2019R1I1A2A01057603)

접수일자 : 2020년 05월 27일 게재확정일 : 2020년 07월 04일
수정일자 : 2020년 07월 03일
교신저자 : 안창욱 e-mail : cwan@gist.ac.kr

우리 생활 및 경제에 있어 중요한 요소로 자리매김하고 있음을 의미한다.

다양한 제조 산업 분야에서 빅데이터 활용하여 제품의 개발 및 생산 비용을 절감하고 생산성과 시장점유율을 높이는 등 기업과 소비자의 후생을 늘려 GDP를 증가시킴으로써 경제 성장의 원동력이 되고 있다. McKinsey에서는 미국의 경우 빅데이터 활용을 통하여 2020년까지 GDP가 약 6,100억 달러 증가할 것으로 예측하였다. 그러나 기존 제조 분야에서는 공정 및 공급망 관리등에서 생성되는 대량 데이터를 이미 축적 및 IT 활용한 자동화가 진행된 산업으로 인하여 제조업에서 빅데이터 활용하여 생산성을 높이고 있는 분야인지에 대한 상반된 의견이 존재한다[1]. 하지만 최근 Internet of Thing 기술 기반으로 제조업체의 부품이나 완성품에 부착될 센서의 방대한 데이터 기반 새로운 빅데이터 활용 분야가 창출되고 있다. 예를 들어 트위터에서 생성되는 일단위 피드의 양이 80GB인 반면, GE가 제조한 가스터빈 엔진의 날개에 있는 하나의 센서가 하루에 수집하는 데이터의 양은 520GB로 하루 기준으로 4배 이상 많다. 또한 제조업 상품과 관련하여 서비스를 지속적으로 제공하고 소프트웨어 판매에까지 나섬에 따라, 기존의 서비스업이 빅데이터 분석에 가지는 강점들(고객/협력업체 수, 높은 거래 빈도수 등)을 제조업에 활용 가능하게 되었다. 그러므로 빅데이터의 가치를 감지한 글로벌 제조 대기업들은 빅데이터 분석에 관심을 가지고 투자를 시작하고 있는 실정이다. 실제 사례로 GE는 2012년에 프로세스의 자동화, 최적화, 정지시간 감축, 고장 시기 예측 등을 통해서 총수입에서 450억 달러의 경영 성과를 이루었다[2]. 그리고 Intel은 빅데이터 기반 예측 분석을 활용하여 1개의 칩 생산라인에서 2012년 300만 달러의 제조원가 절감을 하였고 2013~14년에는 해당 프로세스를 다른 chip 생산라인으로 확장하여 3천만 달러의 추가 원가절감을 달성하였다. 현재 제조 환경에서 수집된 빅데이터 분석을 통하여 제조업의 첨단화가 이루어지고 있다[3].

다음 장에서는 제조업에서 수집된 많은 데이터로부터 의미 있는 정보를 도출하는 분석 방법론에 대하여 다루고자 한다. 대표적인 기계 학습 기법의 경우 입력에 대한 출력 결과를 예측하지만 분석 모델링의 결과 도출 설명력 한계로 인하여 제조 매커니즘 기반 직관적 해석이 불가하여 결과 활용 및 적용에 어려움이 있다. 본 연구에서는 이러한 한계점 극복을 위하여 진화 기법 기반 자동 프로그래밍 기법을 제안하다. 그리고 대표적인 제조업 데이터 활용하여 제안 기법 및 대표적인 기계 학습 기법에 대한 성능을 비교 분석한다. 마지막으로 본 논문의 결론으로 마무리하고자 한다.

Ⅱ. 관련 연구

본 장에서는 다양한 문헌 조사를 통하여 제조 분야 빅데이터 활용 분석 방법에 대하여 자세하게 설명한다. 일반적으로 제조업 분야 빅데이터 분석은 생산품에 대한 양품/불량품 선별하는 분류(classification) 방법과 설비 상태 또는 제품 수요를 예측(prediction)하는 두 가지 방법으로 나눈다. 먼저 분류 기법은 품질 관리에서 많이 사용되고 있으며, 이는 생산 제품에 대한 고객 신뢰도와 직접 연결되는 부분으로 제조 과정의 경우 다양한 제조 과정의 조합으로 이루어진다. 이를 기반으로 제품의 양불 여부 판단하여 최종 생산 제품에 대한 품질 보증이 가능하다. 다음 예측 분야의 경우 실시간 데이터 수집이 가능한 제조 설비 대상으로 설비 이상 여부를 진단하여 사전 조치함으로써 지속적 생산성을 확보에 목적이 있다. 이를 통하여 기존 수행하고 있는 주기적 설비 정비에서 설비 상태 예측 분석 기반 설비 가동률 향상 및 최적의 부품 교체를 통하여 기회비용 손실 최소화 가능하다. 다음 장에서는 빅데이터 분석 방법에 대하여 상세하게 설명한다.

1. 제조 빅데이터 전처리

제조업에서는 연속형, 수치형, 날짜/시간형 텍스트의 정형 데이터 및 이미지, 비디오의 비정형 데이터등의 다양한 형태의 데이터가 수집된다. 이러한 여러 가지 형태의 데이터로 부터 의미 있는 정보 도출을 위해서는 데이터 효율적으로 처리하는 데이터 전처리 방법이 선행되어야 한다.

첫번째 단계는 주어진 다양한 데이터 특성 및 타입을 파악하여 하나의 데이터로 병합하는 것이 제일 먼저 수행하는 데이터 전처리 기법이다. 예를 들어 반도체 제조업에서 chip 품질 이상 분석을 위하여 텍스트 형태 공정 이력 데이터와 설비에서 수집하는 수치형 제조 데이터를 병합을 통하여 분석 수행이 가능하다.

두번째 단계로 서로 다른 다양한 단위를 가지는 데이터에 대한 단위를 표준화 전처리 기법으로 일반적으로 빅데이터 분석 알고리즘의 경우 가장 큰 단위를 가진 데이터로 분석 결과가 편향되는 특징을 가지고 있다. 해당 전처리의 경우 정형화된 방법이 없으므로 현장 전문가의 지식 및 경험 기반으로 전처리 방향 수립이 필수적이다.

세번째 단계로 기술 통계량을 활용하여 제조 현장에서는 수집되므로 방대한 양의 데이터 검증 및 데이터의 특징 파악하는 전처리 기법이다. 여기서 평균(mean) 및 중앙값(median)을 활용하여 데이터 위치 정보 파악이 가능하며, 퍼짐 정도는 표준편차(standard deviation) 및 범위(range)를 활용하여 검증 가능하다. 특히 이렇게 도출한 정보들은 특징 변수(feature variable) 정의하여 분석에서 사용한다. 추가로 비대칭 정도인 왜도(skewness) 및 뾰족한 정도 첨도(kurtosis) 값을 활용하여 변수의 분포 정보도 파악 가능하다. 왜도에서 평균값이 왼쪽 치우친 경우 제곱 또는 지수화를 통한 데이터 증폭하여 정규 분포로 변형한다. 반대로 오른쪽으로 치우친 경우 로그 변환하여 값의 범위를 작게 변형한다. 이러한 데이터 정규 분포 변화를 통하여 분석 정합성 향상이 가능하기 때문이다.

네번째 단계로 제조 데이터에서 빈번하게 발생하는 결측을 처리하는 데이터 전처리 기법은 매우 중요하다. 특히 제조 현장에서 네트워크 부하 및 설비 이상의 외부 요인으로 인하여 데이터 결측이 빈번하게 발생하지만, 그대로 분석 수행한 경우 분석 정합성 확보가 불가능하다. 이러한 결측을 해결하기 위해서 제거 또는 보간의 데이터 전처리 방법을 사용한다. 여기서 결측 데이터를 제거하는 경우는 활용 가능한 데이터 모수가 줄어들어 분석 정합성 확보가 어렵우며, 반대로 결측 데이터를 보간하는 경우 데이터 의미를 왜곡할 가능성이 있다. 데이터 결측에 대한 적합한 방법이 없기 때문에 해당 방법에 대한 검증 후 더 좋은 방법을 선택하는 것이 필요하다.

마지막 단계로 제조업 데이터 자주 발생하는 종속 변수 클래스 불균형(class imbalance)을 처리하는 데이터 전처리 방법이다. 예를 들어 전구 제조업에서 발생되는 불량률은 0.01%로 매우 낮은 경우 불량이 거의 발생하지 않기 때문에 불량을 진단하는 것은 불가능하다. 그러므로 데이터 분석을 통하여 불량 진단 목적 달성이 불가능하다. 이러한 클래스의 불균형 해결하기 위하여 일반적으로 두 가지 방법이 사용된다. 데이터 셋이 충분한 경우 언더 샘플링을 사용하여 많은 데이터 셋을 적은 데이터 셋 수준으로 감소시키는 방식이다. 가령 정상 레이블을 가진 데이터가 10,000건, 비정상 레이블을 가진 데이터가 100건이 있을 경우 정상 레이블 데이터를 100건으로 줄이는 방식이다. 반대로 오버 샘플링은 비정상 데이터와 같이 적은 데이터 셋을 증식하여 학습을 위한 충분한 데이터를 확보하는 방법이다. 동일한 데이터를 단순히 증식하는 방법은 과적합이 되기 때문에 의미가 없으므로 원본 데이터의 피처 값들을 아주 약간만 변경하여 증식하며 Synthetic Minority Over-Sampling Technique (SMOTE)이 대표적 방법이다. 해당 경우 데이터 학습 시간이 많이 걸린다는 단점이 있으므로 가능한 모든 방법에 대한 검증을 통하여 최적의 방법을 선택하는 것이 중요하다.

2. 제조 빅데이터 분석 알고리즘

일반적으로 제조 데이터 분석의 목적은 숫자, 문서, 이미지, 음성, 영상 등의 다양한 입력 및 출력 데이터 간에 의미 있는 정보를 도출하는 것이다. 즉, 불량품에 대한 이상 원인 분석을 통한 예측을 통한 불량 사전 예방하는 것이 목적이다. 최근 제조업에서는 설비 상태를 미반영한 주기적 설비 유지 보수를 수행하고 있지만, 설비에서 수집 가능한 인자들과 고장과의 관계 분석을 통하여 예측 설비 유지을 통한 생산성 개선에 대한 연구가 활발하게 진행되고 있다. 또한 수요 및 재고 관리 분석 기반 예측을 통하여 불필요한 재고품을 많이 보관하고 있을 필요가 없으며, 재고 및 공급 프로세스를 효율적으로 조절할 수 있게 되어 공급자와 제조자 모두 이익 창출이 가능하게 하고 있다. 마지막으로 제품의 신뢰도와 품질에 대한 가치 있는 정보를 활용하여 품질보증 분석을 통한 제품의 결함에 대한 주요 인자를 찾아내는 역할을 한다.

이러한 효율적인 분석을 위하여 활용하고 있는 대표적인 기계학습 알고리즘에 대하여 설명한다. 첫번째 양방향 선형 회귀(stepwise linear regression)방법은 종속변수가 수치형일 경우 변수들 간의 관계를 선형으로 모델링하는 방법이다. 여기서 하나의 설명 변수인 경우 단순 선형 회귀, 둘 이상의 설명 변수를 사용하는 경우 다중 선형 회귀라고 한다. 해당 알고리즘의 경우 종속변수와 설명변수간선형적 관계 함수로 설명하기 때문에 이들 간의 관계를 확인하기에 용이하지만, 종속변수와 설명변수들 간에 선형 관계만을 가정하기 때문에 정확성을 높이는 데는 한계가 있다[6].

두 번째 Least Absolute Shrinkage and Selection Operator (Lasso) 회귀 방법의 경우 기존 선형 회귀에서 적절한 가중치와 편향을 찾아내는 것에 가중치의 절대값 합을 최소화하는 것을 제약 조건으로 추가한다. 하지만 변수들끼리 강한 상관관계가 있다면 해당 알고리즘의 경우 단 하나의 변수만 채택하고 다른 변수들의 계수는 0으로 바꾸는 특징을 가지고 있으며, 이는 원본 데이터에 대한 정보가 손실됨에 따라 정확성이 떨어지는 한계가 있다[6].

세 번째 Ridge 회귀모형에서는 선형 회귀 모형에 가중치들의 제곱합을 최소화하는 것을 추가적인 제약 조건으로 추가하였다. 이를 통하여 모든 변수를 활용하면서 계수값을 줄일 수 있지만 회귀 모델에서 설명변수가 증가하여도 그대로 변수의 수를 유지하여 성능 저하에 영향을 준다[6].

네 번째 Elastic Net은 Lasso 및 Ridge 회귀 방법의 조합으로 가중치의 절대 값 및 제곱 합을 동시에 제약 조건으로 가지는 모형이다. 여기서 입력 변수들이 독립적으로 구성된 경우 Elastic Net은 먼저 상관된 변수로 구성된 그룹을 형성하여 이 그룹의 변수 중 하나가 종속 변수와 강한 관계가 있다면 이 전체 그룹을 모델에 포함된다[6]. 같은 그룹 내에 속하면서도 강력한 예측 변수 하나가 아닌 그 외 변수들을 모두 제거하면 해석에 있어 정보 손실이 발생하여 결과적으로 모델 성능은 떨어지게 된다.

다섯 번째 주성분 회귀 분석(principal component regression)은 주성분 분석을 이용해 독립변수를 새로운 좌표축으로 이동시킨 후에 압축된 독립변수를 이용해 종속변수와의 관계를 다중회귀를 이용해 분석한다[15]. 독립변수는 주성분분석을 거친 후에 변환된 좌표 중 몇 개의 주성분을 독립변수로 다중회귀를 하게 된다. 즉, 주성분 분석을 통해 필요한 주성분만을 독립변수로 선정해서 회귀분석을 하는 것이다. 여기서 해당 기법은 주성분 분석을 사용으로 인하여 높은 상관관계 변수끼리 동일한 주성분으로 구성하기에 다중 공선성 문제가 해결 가능하다. 그리고 주성분 분석 기법을 통하여 변환한 주성분 변수들 중에서 상위 변수들만 선택할 경우 Lasso 회귀분석처럼 정규화 효과를 줄 수 있어 모델의 과적합 현상을 완화시킬 수 있는 장점이 있다. 반대로 각 주성분 변수들은 실제 독립변수들의 전체 영향력을 부분적으로

반영하여 이를 통하여 각 조건의 영향력을 파악은 불가능하여 도출한 모델에 대한 해석이 불가능 한 단점이 있다[14].

다음으로 대표적인 두 가지 black box 기반 기계학습 기법을 설명한다. 처음으로 support vector machine(SVM)은 두 그룹에서 각각의 데이터 간 거리를 측정하여 두 개의 데이터 사이의 중심을 구한 후에 그 가운데에서 최적의 초평면을 구함으로써 두 그룹을 나누는 방법으로 학습한다. 여기서 직선으로 나눌 수 있다면 선형 분류 모델을 적용하고, 직선으로 나눌 수 없는 경우 비선형 분류 모델을 사용하게 된다. 주어진 데이터를 이진 분류해 내는 데에 있어 다른 알고리즘 보다 정확도가 뛰어나지만, 데이터 셋의 크기에 따라 복잡도가 증가하여 연산 속도가 느려진다는 단점이 있다[9,14,16]. 마지막 random forest(RF)는 두 가지 이상의 의사결정나무(decision tree)를 만든 후 최빈값을 기준으로 예측 또는 분류하는 알고리즘이다. 하나의 의사결정나무만을 사용하면 과적합(over fitting)의 확률이 높아 이를 효율적으로 해결하기 위하여 무작위하게 여러 개의 트리를 만들어 각각 어떤 결과를 내는지 보고 각 트리별 결과를 취합하여 결과를 예측한다. 해당 기법의 경우 분석에 있어 높은 정합성을 가진다. 그러나 동일한 결과를 도출하기에는 어려움이 있으며, 텍스트 데이터와 같은 매우 차원이 높고 희소한 데이터에는 잘 작동하지 않는다[6].

다음 장에서는 기존 제조 데이터 분석 방법이 가지고 있는 분석 결과에 대한 해석이 불가한 한계를 극복하기 위하여 유전 프로그래밍(genetic programming) 활용 제조 데이터 분석 기법을 제안한다.

III. 유전 프로그래밍 활용 데이터 분석 방법 연구

본 장에서는 대표적인 진화 알고리즘의 일종인 유전 프로그래밍을 활용한 제조 빅데이터 예측 분석 방법을 제안한다. 해당 분석 기법의 경우 기존 활용하고 있는 분석 알고리즘의 도출 결과와 제조 매카니즘 간의 부재한 설명력을 강화하는 것이 목적이다. 활용도가 높은 회귀 분석의 경우 제조 인자 간의 관계를 선형적 함수로 해석이 가능하나, 복잡한 변수간 비선형 관계에 대한 해석이 불가능한 단점이 있다. 반면에 black box 기반 기계학습 알고리즘들은 높은 정합성 결과 도출이 가능하지만, 결과에 대한 변수간의 관계 설명 불가하여 제조 원리 기반 재해석이 필요하다. 이러한 한계를 극복하기 위하여 제조 과정 동안 수집되는 입력 및 출력 변수 관계를 유전 프로그래밍을 활용하여 수식화 가능한 예측 모델링 방법을 연구한다. 이를 위하여 활용하는 유전 프로그래밍 알고리즘의 원리 및 분석 방법을 구체적으로 설명한다.

먼저 유전 프로그래밍은 생태계의 자연 선택(natural selection) 또는 적자 생존(survival of the fittest)을 기반으로 하여, 개체군(set of population)의 전역적인 탐색을 통한 최적의 해(optimal solution)를 찾는 확률적 탐색 알고리즘(stochastic search algorithm)의 일종이다[7]. 여기서, 자연 선택 또는 적자 생존이란, 우성 형질의 염색체는 다음 세대로 전이되고, 열성 형질의 유전자

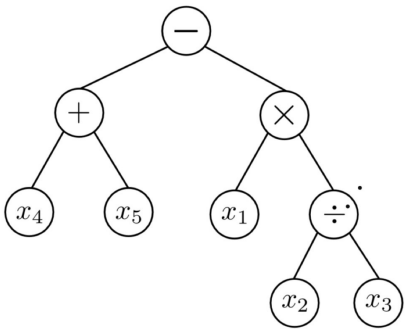

그림 1. 유전 프로그래밍 염색체 표현 방법

는 다음 세대로 전이되지 못하고 도태되는 것을 의미한다[11]. 진화 알고리즘은 문제의 해를 표현하는 염색체 부호화 방법(encoding), 적합도 함수(fitness function), 그리고 유전 연산자(genetic operator)들의 세 단계를 거쳐 진화함으로써 최종 수렴 해를 얻게 되는 동작 과정으로 구성된다. 여기서 유전 연산자는 선택 또는 재생산, 교배, 그리고 돌연변이의 세 동작 과정으로 구성된다. 제조 데이터 분석을 위하여 활용하는 유전 프로그래밍은 입력 및 출력 변수 간의 관계를 수학적 기호를 활용하여 입력 변수의 관계로 부호화되어 직관적 해석에 용이한 트리 구조로 표현한다. 여기서 트리 구조의 루트 노드 및 중간 노드에는 수학적 연산 기호 할당 되고, 단말(leaf) 노드들에는 입력 변수들이 할당된다. 이러한 해의 표현 방법을 통하여 제조 원리 기반 결과 해석이 가능하여 현장 적용이 용이하다[8,10,12]. 그림 1에서는 수학적 기호로 부호화된 해의 예를 표현하며, 다음과 같은 $(x_4 + x_5) - (x_1 \times \frac{x_2}{x_3})$의 수식으로 표현된다. 다음으로 입력 및 출력 변수간의 관계를 수식으로 표현된 염색체는 적합도 함수를 통하여 물리적 수치로 평가한다. 그러므로, 적합도 함수의 설계는 유전프로그래밍의 성능을 평가하는 매우 중요한 요소이다. 입력 변수에 대한 예측 결과와 출력 결과에 대한 차이를 비교 평가를 위하여 평균 제곱근 편차(root mean square error) 활용하여 적합도 함수를 설계한다. 다음으로 해의 다양성 확보를 위한 유전 연산자에 대하여 설명한다. 먼저 선택(selection) 또는 재생산(reproduction) 연산자는 다윈의 적자생존 원칙을 바탕으로 우수한 품질의 염색체가 다음 세대에 유전되는 가능성을 증가시켜 개체군의 평균 품질을 향상 시키는 역할을 한다[7-8].

여기서 선택 기법은 해-표면상에서 최적 해가 존재 가능한 영역을 집중 탐색 할 수 있도록 하는 것이 목적이다. 본 논문에서는 무교체 승차 진출 선택(tournament selection without replacement)을 사용하여 서로 다른 염색체를 개체군으로부터 선택하여 가장 우수한 적합도를 갖는 염색체를 다음 세대의 부모 염색체로 선택한다.

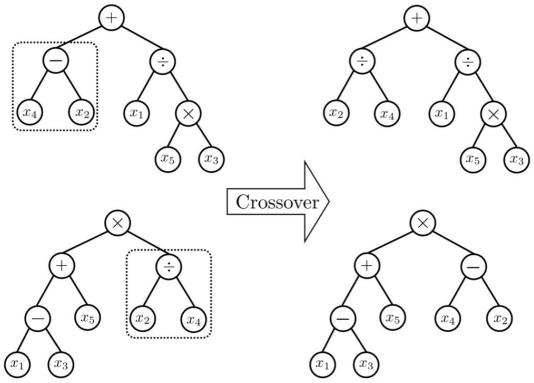

그림 2. 진화 기반 자동 프로그래밍 교배 과정

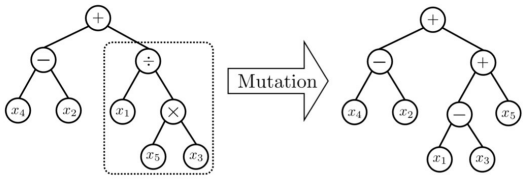

그림 3. 진화 기반 자동 프로그래밍 돌연변이 과정

다음 연산자인 교배(crossover)는 두 부모 염색체의 부분 염색체(partial chromosome)를 교환하여 보다 우수한 형질을 가진 자식 염색체를 생성하는 역할을 한다. 즉, 교배 기법은 최적 해의 존재 가능성이 높은 영역의 해-표면을 탐색하도록 하는 방법으로 본 논문에서는 유전자 위치에 의존성이 없는 교배 방식을 활용한다. 즉, 선택된 두 염색체의 유전자의 위치에 관계없이 동일한 노드 (대립 형질)를 포함하고 있는 유전자의 위치 쌍(locus pair)에 대한 집합을 형성한다. 그리고, 제안 알고리즘의 교배 동작은 그림 2에 예시되어 있다[10,12].

마지막 유전 연산자인 돌연변이(mutation)는 염색체의 유전자를 다른 대립 형질로 돌연변이 시켜서 개체군의 유전적 다양성을 유지하도록 한다. 해당 연산의 목적은 유전자 알고리즘이 준 최적해로 수렴하는 것을 방지하고, 최적 해로 수렴하도록 모든 해-표면을 탐색하는 것이다. 그러나 많은 수의 염색체를 돌연변이 시킬 경우 최적의 해를 향해 수렴중인 개체군 해-표면상에 무작위적으로 분산될 수 있기 때문에 수렴 및 최적 성능이 열악해질 수 있다[8,12].

그리고, 돌연변이 동작은 유전자 알고리즘의 수렴 성능을 약간 감소시킬 수는 있지만, 기존 개체군으로 탐색할 수 없는 영역의 해-표면을 탐색할 수 있는 변형된 형태의 염색체를 생성하므로 준 최적해로의 수렴 가능성을 완화 시킨다. 개체군의 유전적 다양성을 유지하는 형태와 동일한 의미를 지닌 돌연변이 기법은 새로운 부분적 입력 변수 관계를 생성하여 대처한다. 그림 2에서 돌연변이 동작 과정을 묘사한다.

다음으로 그림 4에서는 유전 프로그래밍을 활용한 제조 빅데이터

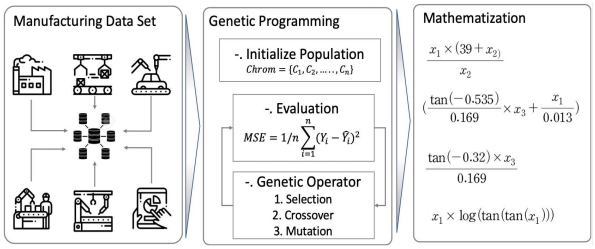

그림 4. 유전 프로그래밍 활용 제조 빅데이터 분석 방법

분석 과정을 설명한다. 먼저 수집 데이터 검증 및 전처리를 통하여 분석 최적의 데이터로 변환한다. 다음으로 유전 프로그래밍 활용 최적의 입력 변수 조합을 도출하기 위하여 트리 구조를 활용하여 해를 표현한다. 해당 트리 구조에서 루트 및 중간 노드에는 정의되어진 수학적 기호로 단말 노드는 주어진 입력 변수를 활용하여 해를 부호화한다. 이러한 표현된 해들은 출력 변수와 차이가 최소화 되도록 유전 연산자 과정 반복을 통하여 최적의 변수들의 관계 수식을 도출한다. 해당 결과 기반 제조 원리 기반 해석이 불필요하여 바로 현장 적용 가능하다. 그리고 최적의 입력 변수 조합 수식을 활용하여 기존 해석 불가한 제조 원리 해석도 가능하다.

다음 장에서는 실제 제조 데이터 3종을 활용하여 기계 학습 알고리즘 기반 데이터 분석 방법과 제안 유전 프로그래밍 활용 분석 방법에 대한 성능을 비교 검증한다.

Ⅳ. 실험 결과

본 장에서는 총 3가지 실제 제조 데이터 셋을 활용하여 제안 유전 프로그래밍 활용 분석 방법과 대표적인 기계 학습 알고리즘 분석 방법들과 성능을 비교 분석한다. 성능 평가 방법은 예측치와 실측치와의 차이를 평균 제급근 잔차(mean squared error) 활용 평가한다. 모델들의 성능 검증을 위하여 전체 데이터 중에서 80% 데이터를 학습용으로 나머지 20% 데이터는 테스트로 활용하여 분석 방법에 대한 성능을 검증한다. 그리고 동일 조건 100회 반복 실험을 하여 반복 실험에 대한 평균값, 표준 편차 및 사분위수 활용하여 도출한 결과에 대한 신뢰를 확보하고자 한다. 먼저 유전 프로그래밍에서 개체군의 크기는 $1000 \times \sqrt{Dim}$, 세대수는 $10 \times \sqrt{Dim}$, 교배 확률 0.7, 그리고 돌연변이 확률 0.1로 설정하였으며, 해의 수식 표현을 위하여 $\{+, -, \times, \div, \log, \tan\}$ 함수 셋을 정의한다. 그리고 성능 비교 검증을 위하여 활용한 Lasso, Ridge 및 Elastic Net 알고리즘의 경우 penality term 제어를 위하여 최적의 정규화 방법을 사용하였고 SVM은 선형 커널 및 Z-score (평균 = 0, 표준편차 = 1) 활용 데이터 표준화하였다. 마지막으로 RF 알고리즘에서는 각 노드에서 고려될 변수의 개수를 $\sqrt{\frac{Dim}{3}}$ 로 몇 개의 나무를 생성할지 설정하는 인자를

500으로 정의하였다.

1. Auto MPG Data Set 검증 결과

먼저 활용한 Auto MPG 데이터는 1983년 American Statistical Association Exposition에서 최초로 사용한 데이터로 3가지 이산형 변수(cylinder, horsepower, model_year)와 3가지 연속형 변수(displacement, weight, acceleration)의 총 6가지 입력 변수로 사이클 연료 소비량(mpg 실린더)을 출력 변수로 구성된다[5].

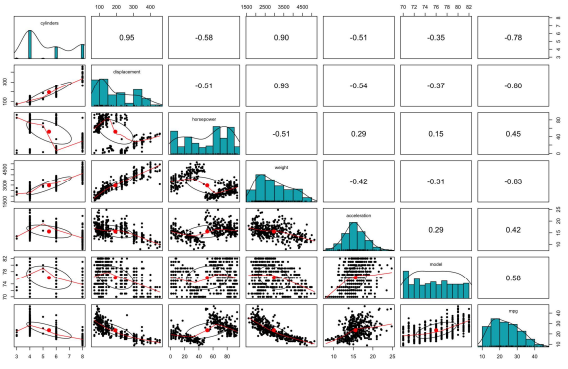

그림 5. Auto MPG Data Set 데이터 분포 및 상관관계

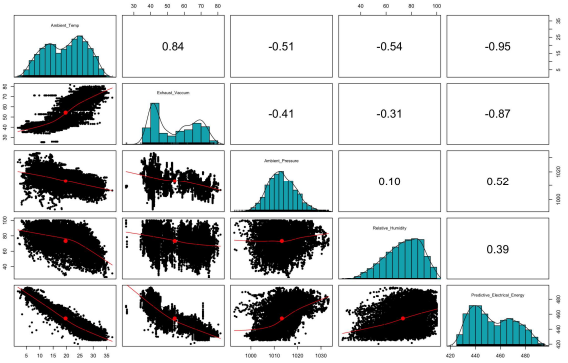

그림 6. Combined Cycle Plant 데이터 분포 및 상관관계

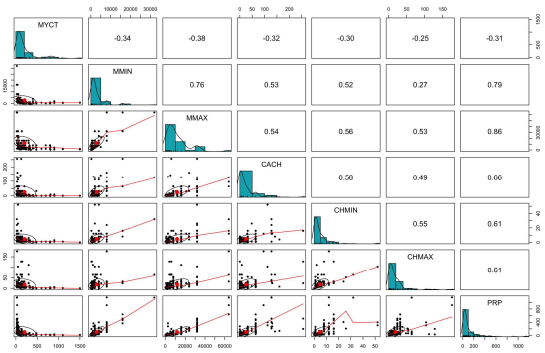

그림 7. Combined Cycle Plant 데이터 분포 및 상관관계

주어진 데이터는 총 398개 행으로 구성되었지만, 데이터 결측이 존재하여 이를 제거하는 데이터 전처리 방법을 적용을 통하여 총 391개의 모수만 활용하여 성능을 검증한다. 그림 5를 통하여 입력 및 출력 변수 분포 및 변수들 간의 상관관계 정도를 확인 가능하다. 여기서 출력 변수 mpg는 모든 입력 변수와 양/음 상관관계를 가지며 입력 변수들 간에도 상관관계가 존재하는 한다. 표 1를 통하여 해당 데이터에 대한 성능 검증 결과 유전 프로그래밍 활용 분석 방법과 black box 기반 기계 학습 기법(random forest 및 svm)에서 우수한 성능을 보였지만, 회귀분석 알고리즘들에서는 낮은 정합성을 보였다. 이를 통하여 복잡한 관계가 존재하는 실제 데이터에서는 회귀 분석 기법의 성능의 한계를 확인 가능하다. 그리고 표 2에서는 각 알고리즘을 활용 입력 및 출력 변수에 관계에 대한 도출되어지는 수식 확인이 가능하다.

2. Combined Cycle Plant 데이터 셋 검증 결과

두 번째 사용 데이터는 2006년 ~ 2011년 동안 복합 사이클 발전소 공장의 시간당 전기 에너지 출력(PE)과 시간 단위 수집되는 평균 주변 온도(AT), 주변 압력(AP), 상대 습도(RH) 및 배출 진공(V) 총 5가지 변수와 9568개의 행으로 구성된다[5]. 그림 6을 통하여 입력 및 출력 변수 분포 확인과 변수들 간의 상관정도 확인이 가능하다. 여기서 에너지 출력 변수는 평균 주변 온도 및 배출 진공와 -0.95 및 -0.87의 강한 음의 상관 관계를 주변압력 및 상대습도와는 0.52 및 0.39의 양의 상관 관계를 가진다. 그리고 시간별 평균 주변 온도 변수는 나머지 입력 변수들과 양 또는 음의 다양한 상관 관계를 가진다. 표 3를 통하여 해당 데이터 기반 성능 비교 결과 유전 프로그래밍 기반 분석 알고리즘 대비 black box 기반 기계 학습 기법 (random forest 및 svm)에서 가장 우수한 성능을 보였다. 다음으로 선형 / Lasso / Elastic net 회귀분석 기법에서는 높은 정합성을 보였다. 그리고 표 4에서는 각 알고리즘에서 도출되어지는 입력 및 출력 변수 관계에 대한 수식 확인이 가능하다.

3. CPU Performance Data Set 검증 결과

마지막으로 CPU 성능 검증 데이터 기반으로 분석 알고리즘에 대한 성능 비교한다. 총 209가지 측정 모수에서 estimated relative performance(ERP)을 종속변수로 machine cycle time(MYCT), minimum main memory(MMIN), maximum main memory(MMAX), cache memory(CACH), minimum channels in units(CHMIN), maximum channels in units(CHMAX)를 6가지 정수형 입력 변수로 데이터가 구성된다[5]. 그림 6을 통하여 입력 및 출력 변수 분포 확인 결과 모든

변수에서 평균이 왼쪽으로 치우쳐 있다. 변수들 간의 상관계수를 통하여 상관관계가 확인결과 예측변수(ERP)와 입력 변수 MMIN, MMAX, CACH, CHMIN, CHMAX간에는 각각 0.82, 0.90, 0.65, 0.61, 0.59의 양의 상관관계를 가진다. 그리고 MYCT 변수만 다른 모든 변수들과 음의 상관관계를 가지는 반면 나머지 변수들은 모두 양의 상관 관계를 가진다. 표 5에서는 해당 데이터 기반 성능을 비교한 결과 보여준다. 여기서 유전 프로그래밍 기반 분석 기법이 가장 우수한 성능을 black box 기반 기계 학습 기법 random forest가 다음으로 우수한 성능을 보였다. 다음으로 회귀분석 기법들이 우수한 성능을 가진다. 그러나 SVM의 경우 가장 낮은 정합성 성능을 보였다. 이를 통하여 정수형 데이터의 경우 SVM 활용을 통하여 정합성 확보함에 있어 한계가 있는 것으로 분석된다. 그리고 표 6에서는 각 알고리즘을 통하여 도출되는 입력 및 출력 변수 관계 표현 결과 확인이 가능하다.

실제 제조 데이터 기반 대표적인 기계학습 기법의 성능 검증 결과 데이터 관계 복잡도에 따라 성능의 차이를 검증하였다. 일반적으로 black box 기반 기계학습 기법인 random forest의 경우 우수한 성능을 보였다. 하지만 분석 결과에 도출에 대한 설명력이 부재하여 제조 분야 활용을 위한 제조 공정 기반 해석과정이 추가로 필요하다. 그리고 선형적 관계가 있는 데이터에서는 회귀 분석 기법의 성능이 우수하지만 분석 모델링이 선형 수식으로 도출되어 설명력이 단조로운 단점이 있다. 그러나 해당 기법의 경우 우수 또는 유사한 정합성을 가지면서 분석 결과에 대한 선형 또는 비선형의 결과 해석력을 보여 준다.

V. 결 론

본 연구에서는 입력 변수에 대한 출력을 예측하는 black-box 기반 기계 학습 알고리즘이 가지고 있는 분석 결과에 대한 해석이 불가한 한계점을 극복을 위하여 유전 프로그래밍 기반 분석 방법을 제안한다. 기존 기계 학습 알고리즘의 경우 높은 예측 정합성 가지지만, 특히 제조업에서는 제조 공정 원리 기인 해석을 통하여 결과의 근거, 도출 타당성에 대한 검증이 중요하다. 제안 분석 방법의 경우 입력 및 출력 변수 관계 수식화를 통하여 제조 원리 기반 직관적 해석이 용이하다. 그리고 변수간의 수식화된 결과를 활용하여 기존에 해석이 불가한 제조 원리 도출도 가능하다. 실제 제조 데이터 기반 성능 검증 결과 우수한 성능을 보였다. 추후 설명력 및 정합성이 향상을 위한 추가 연구를 진행하고자 한다.

표 1. Auto MPG Data Set 검증 결과

	LR	Lasso	Ridge	Elastic	PLS	RF	SVM	GP
Mean	12.15	12.09	12.56	12.16	18.8	8.38	8.91	**7.35**
Std	1.17	1.20	1.32	1.20	1.75	1.23	1.53	**1.25**
Min	9.80	9.68	9.60	9.93	14.37	5.76	5.79	**5.18**
Q1	11.26	11.19	11.53	11.18	17.69	7.56	7.83	**6.52**
Q2	12.04	11.96	12.46	12.08	18.60	8.34	8.87	**6.93**
Q3	12.90	12.85	13.43	12.97	19.83	9.11	9.73	**7.92**
Max	15.36	15.70	17.24	15.80	23.63	**11.37**	13.96	11.57

표 2. Auto MPG Data Set 수식 결과 예제

	분석 결과
LR	$mpg = -11.24 - 0.01 \times weight + 0.72 \times Modelyear$
Lasso	$mpg = -11.30 + 0.01 \times horsepwr - 0.01 \times weight + 0.01 \times acceleration + 0.71 \times Modelyear$
Ridge	$mpg = -9.32 - 0.36 \times cylinders - 0.01 \times displacement + 0.01 \times horsepwr - 0.01 \times acceleration + 0.66 \times Modelyear$
Elastic	$mpg = -11.55 + 0.01 \times horsepwr - 0.01 \times weight + 0.02 \times acceleration + 0.71 \times \mathrm{mod}elyear$
GP	$mpg = \log(\frac{cylinders \times Modelyear}{horsepwr}) + \frac{Modelyear}{accleration} + \frac{Modelyear^2}{horsepwr \times cylinders}$

표 3. Combined Cycle Power Plant Data Set 검증 결과

	LR	Lasso	Ridge	Elastic	PLS	RF	SVM	GP
Mean	20.87	20.88	24.02	20.88	60.20	**12.45**	16.34	24.34
Std	0.51	0.51	**0.50**	**0.50**	1.54	0.52	0.55	1.41
Min	19.72	19.72	22.59	19.73	56.05	**11.34**	15.04	21.00
Q1	20.49	20.50	23.66	20.51	59.15	**12.07**	15.95	23.23
Q2	20.86	20.88	24.09	20.88	60.08	**12.39**	16.29	24.32
Q3	21.25	21.26	24.37	21.26	61.14	**12.87**	16.76	25.35
Max	22.14	22.08	25.21	22.09	64.26	**13.64**	17.62	27.76

표 4. Combined Cycle Power Plant Data Set 수식 결과 예제

	분석 결과
LR	$PE = 455.14 - 1.98 \times AT - 0.23 \times V + 0.06 \times AP - 0.16 \times RH$
Lasso	$PE = 454.87 - 1.96 \times AT - 0.24 \times V + 0.06 \times AP - 0.15 \times RH$
Ridge	$PE = 276.44 - 1.40 \times AT - 0.40 \times V + 0.23 \times AP - 0.05 \times RH$
Elastic	$PE = 448.99 - 1.95 \times AT - 0.24 \times V + 0.07 \times AP - 0.15 \times RH$
GP	$PE = 1 + \frac{2V}{AT} - 2 \times \tan(\log(RH)) + \log(RH) + \frac{0.934}{V} + V - \frac{AP + RH}{\log(AP) \times (-0.305)}$

표 5. CPU Performance Data Set 검증 결과

	LR	Lasso	Ridge	Elastic	PLS	RF	SVM	GP
Mean	6062.30	6042.90	5925.79	6054.67	7120.61	5535.26	12558.54	**4690.94**
Std	2913.99	**2863.57**	2894.36	2920.57	2590.53	5345.29	8895.90	4087.61
Min	1976.25	1848.51	1623.37	1904.49	2978.80	854.78	1481.23	**900.34**
Q1	3911.82	3968.56	3920.86	3925.28	5671.00	1917.05	4419.79	**2238.31**
Q2	5090.59	5161.14	5115.76	5174.76	6660.71	3745.71	11151.02	**3488.37**
Q3	8630.48	8312.11	7368.07	8294.33	7701.59	7027.19	17593.98	**5507.6**
Max	**15135.69**	15282.50	15983.83	15277.86	21302.31	26638.17	45920.76	25718.17

표 6. CPU Performance Data Set 수식 결과 예제

	분석 결과
LR	$ERP = -59.82 + 0.05 \times MYCT + 0.01 \times MMIN + 0.01 \times MMAX + 0.59 \times CACH + 1.41 \times CHMAX$
Lasso	$ERP = -58.04 + 0.05 \times MYCT + 0.01 \times MMIN + 0.01 \times MMAX + 0.58 \times CACH + 1.39 \times CHMAX$
Ridge	$ERP = -49.73 + 0.04 \times MYCT + 0.01 \times MMIN + 0.01 \times MMAX + 0.58 \times CACH + 0.91 \times CHMIN + 1.33 \times CHMAX$
Elastic	$ERP = -54.80 + 0.04 \times MYCT + 0.01 \times MMIN + 0.01 \times MMAX + 0.57 \times CACH + 0.03 \times CHMIN + 1.37 \times CHMAX$
GP	$ERP = \frac{MMIN}{MYCT}\tan(0.402) + \log(MMIN) + CACH + \log\left(\frac{MMIN \times CHMAX}{MYCT}\tan(0.310)\right) + \log\left\{\tan\left\{\frac{MMIN \times CACH}{MYCT^2 \times \tan\left(\log\left(\frac{VMMIN \times x5}{MYCT}\tan^2(0.310) + CHMAX + CHMIN + 2CACH^2 - MMAX + 0.688\right)\right)}\right.\right.$

REFERENCES

[1] 이훈혜, "제조업 경쟁력 강화를 위한 빅데이터 활용 방안,"*산업연구원*, 2013년 12월

[2] 곽기호, 이하목, "제조업 빅데이터 활용 동향 분석과 시사점," *주간기술동향*, 제1762호, 2016년 9월

[3] 과기부처합동, "데이터·AI경제 활성화 계획," *과학기술정보통신부*, 2019년 1월

[4] A. Hashmi, and T. Ahmad, "Big data mining: Tools & algorithms," *International Journal of Recent Contributions from Engineering*, vol. 4, no. 1, pp. 36 - 40, Oct. 2016.

[5] D. Dua, and C. Graff, UCI Machine Learning Repository Irvine, CA: University of California, School of Information and Computer Science(2019). http://archive.ics.uci.edu/ml (accessed July 8, 2020).

[6] K. Arun, and L. Jabasheela, "Big Data: Review, Classification and Analysis Survey," *International Journal of Innovative Research in Information Security*, vol. 1, no. 3, pp. 17 - 23, Sept. 2014.

[7] B. David, R.B. David, and R.M. Ralph, "An Overview of Genetic Algorithms: Part 1, Fundamentals," *University Computing*, vol. 15, no. 2, pp.58-69, 1993.

[8] B. L. William, "Genetic Programming + Data Structures = Automatic Programming," *Kluwer Academic Publishers,* 1998.

[9] Tom Mitchell, “Machine Learning,” *McGRAW-HILL International Editions, Computer Science Series*, 1997.
[10] W. Banzhaf, P. Nordin, R.E. Keller, and F.D. Francone, “Genetic Programming: An Introduction on the Automatic Evolution of Computer Programs and Its Applications,” *Morgan Kaufmann Publishers, Inc. San Francisco, California*, July 1997.
[11] C. W. Ahn., “Advances in Evolutionary Algorithms : Theory,” *Design and Practice. Berlin, Germany: Springer*, 2006.
[12] P. A. Whigham, “Inductive bias and genetic programming,” *First International Conference on Genetic Algorithms in Engineering Systems : Innovations and Applications, Sheffield, UK*, pp. 461 - 466, Oct. 1995.
[13] I. H. Witten and E. Frank, “Data Mining: Practical Machine Learning Tools and Techniques,” *San Francisco, CA, USA: Morgan Kaufmann Publishers*, 2005.
[14] 손민규, “데이터 분석을 떠받치는 수학 : 통계로 배우는 데이터 과학의 기술,” 2010년 6월
[15] 오정원, 김행곤, 김일태, “머신러닝 적용 과일 수확 시기 예측시스템 설계 및 구현,” 스마트미디어저널, 제8권, 제1호, 74-81쪽, 2019년 3월
[16] 박동주, 김병우, 정연선, 안창욱, “Deep Neural Network 기반 프로야수 일일 관중 수 예측: 광주-기아 챔피언스 필드를 중심으로,” 스마트미디어저널, 제7권, 제1호, 16-23쪽, 2018년 3월

저 자 소 개

오상헌 (정회원)
2011년 광주과학기술원 정보통신공학과 박사 졸업.
2017년 ~ 현재 한국 방송통신대학교 바이오-정보 통계 학과 대학원
2018년 ~ 현재 SK Hynix P&T DT 수석 연구원.

<주관심분야: 메타-진화 알고리즘, 기계 학습 기반 제조 빅데이터 분석>

안창욱 (정회원)
2005년 광주과학기술원 정보통신공학과 박사 졸업.
2005년 ~ 2007년 삼성종합기술원 전문연구원.
2008년 ~ 2017년 성균관대학교 컴퓨터공학과 부교수.
2017년 ~ 현재 광주과학기술원 전기전자컴퓨터공학과/AI대학원 교수.
<주관심분야: 메타-진화 알고리즘, 퀀텀 기계 학습, 진화 강화 학습, AI 음악>