K-인공지능(AI) 제조데이터 경진대회 보고서

프로젝트명	오류 사례가 적은 현장에서의 오류 가능성 사전 진단
팀명	HaDa
내용요약	제조 현장에서, 설비의 규모와 가짓 수가 많을수록 설비의 정밀 점검 및 관리가 어렵고, 설비 오작동을 예측할 수 있는 수치가 누적되기까지의 보고되는 정보량이 많기 때문에, 비정상적 수치 또는 오작동 가능성의 조사에 많은 예산과 자원이 소모됩니다. 따라서 사전에 설비의 오작동 가능성과 발생 원인등을 조기에 진단하고, 보고하여 정비를 용이하게 할 수 있는 AI 솔루션이 필요하다고 생각하였고 해당 과제에 적합하다고 판단되는 모델을 선정하였습니다. 주어진 데이터의 편향적 분포에 따라 정확한 예측을 위해 무관한 특성을 제거, 특정 분포의 범위로 재정의하고, 앙상블 기법의 랜덤 포레스트, 앙상블 부스팅 알고리즘의 그래디언트 부스팅 그리고 직접 정의한 얕은 신경망 구조를 통한 도메인{데이터 셋} 최적화 학습기를 탐색 및 분석, 개발하였습니다. 제작된 AI 모델이 타 분야의 이상 수치 검측에 있어서 본 도메인과 유사하게, 이상 사례가 극히 적지만, 민감하게 작용되어 조기에 검측 및 예방해야하는 다양한 현장에 있어서 유용하게 사용될 수 있을 것이라 생각합니다.

상기 본인(팀)은 위의 내용과 같이 K-인공지능(AI) 제조데이터 경진대회 결과 보고서를 제출합니다.

2021 년 12 월 2 일

팀장 : (직인)

팀원 : (직인)

팀원 : (직인)

한국과학기술원장 귀중

□ 문제정의

○ 대규모 대량 설비의 정밀 점검과 실시간 관측이, 시설 내 관리적 측면의 한계로 어려움.

- 설비가 많고 규모가 클수록 시설 차원에서의 정비가 어렵기 때문에, 수치를 바탕으로 조기에 설비의 오작동 가능성을 알려주는 AI 솔루션이 있다면, 정비와 예방에 있어 도움이 될 것 이라 생각하였습니다.

□ 제조데이터 정의 및 처리과정

○ 오작동의 진단과 원인의 규명을 위한 데이터의 처리

- 오작동과 원인의 조기 진단을 위한 특성의 정의

* PART NAME, EQUIP_CD, EQUIP_NAME : 설비 파트, 장비 번호, 장비 모델명 또한 관리 또는 노후화 문제로 연관성이 있을 것이라 판단하여, 숫자 코드로 변환해 훈련 데이터로 활용.
* Reason의 None, 가스, 미성형, 초기허용불량을 각각 0, 1, 2, 3 의 숫자 코드로 변환
* PassOrFail은 Reason 과 종속관계에 있는 변수지만, Reason 의 예측을 위한 상관관계를 지원할 수 있어 이진 코드로 변환 및 사용하였습니다.
* _id, TimeStamp, PART_FACT_PLAN_DATE 의 경우 설비별 이상으로 적용시키기에 관계도가 유효하지 않다고 판단되어 제외하였습니다.
* 코드 적용된 데이터 셋을 RobustScaler를 활용해 이상치를 포함하여 데이터가 분포를 따르도록 재 구성하여, 보다 정확한 분류 스케일링을 적용하였습니다.

□ 분석모델 개발

○ Random Forest, Gradient Boost, Shallow Nerual Network

- Random Forest Classifier : RFC

* 랜덤 포레스트 분류기는 Bagging 과정을 거쳐 데이터 셋 내의 중복이 허용되는 부분집합을 이용해 모델을 훈련시킴으로서 해당 주어진 데이터 셋처럼 값의 분포가 적을 때 예측성능을 높이기에 적합하므로 선택하여 실험해보았습니다.

- Gradient Boost Classifier : GBC

* 그래디언트 부스트 분류기 역시 앙상블 기법에 해당하는 모델로서, 데이터 셋의 값의 분포가 불균형하고, 적은 비율을 가지고 있을 때 이상치에 대한 학습량이 적어 모델의 성능이 잘 나오지 못하는 한계를 극복하기 유리한 특성이 있습니다. 해당 알고리즘은 약한 학습 모델들의 오답에 가중치를 부여해, 보다 정답에 가까워지도록 반복해서 약한 학습 모델들을 진보, 결합시켜 강한 학습기를 만들어 내는데 그 목표가 있기 때문입니다. Gradient Boost 라는 이름에서 알 수 있듯, Gradient Descent 기법을 예측에 사용하여 손실의 전역적 최소 구간에 이를 수 있도록 손실율을 최소화하는 방법을 통해 분류기들의 예측 오류 구간을 보정하여 보다 강한 분류기를 만들어 낼 수 있는 장점이 있어 사용하였습니다.

- Shallow Nerual Network : SNN

* 4계층의 활성화 함수 LeakyReLU 와 분류기 softmax 로 구성된 얕은 구조의 신경망을 구성하였습니다. 편향된 데이터의 정확한 예측을 위해 신경망 훈련 과정에 클래스 별 가중치를 다르게 적용하여, 저빈도 데이터의 상관관계 예측을 위해 가중치를 데이터의 이상치에 민감하게 작용함으로서 더 정확한 분류를 할 수 있도록 구성하였습니다.

□ 분석결과 및 시사점

○ Gradient Boost Classifier 소수 비율의 클래스 예측 가능

- Accuracy : GBC 100%, RFC 100%, SNN 99% 정확도는 매우 높음

* 정확도는 위 4개의 분류를 해내는 데 옳은 척도로 쓰일 수 없었습니다. 직관적으로 볼 때 우리 domain { 데이터 셋 } 의 주제에 관해 데이터의 존재 비율은 원인 가스가 0.4%, 초기허용불량이 0.2%, 미성형이 0.2%, 정상이 99.2로 이 타 이상 수치를 모두 파악하지 못하더라도 옳다고 판단만 한다면 정확도가 99.2% 로 매우 높게 계산될 수 밖에 없습니다. 데이터의 편향이 존재하므로 정확도만을 가지고는 모델을 평가 할 수 없었습니다.

- Macro F1: GBC 86% > RFC 77% > SNN 71%

* Macro F1 수치는 재현율, 정밀도를 모두 고려해 각 분류 클래스에 대한 결과의 평균을 산출해내기 때문에 해당 domain처럼 편향이 일어난 경우 유용하게 사용되는 평가 척도기 때문에 사용하였습니다.

해당 분석에서 구성된 GBC, SNN과 RFC 모두 초기허용불량, { 3번 클래스 } 에 대한 domain 의 이상치를 예측하는 데에 높은 수치를 보여주었습니다..

다수를 차지하는 가스 { 1번 클래스 } 의 예측에는 SNN이 취약한 모습을 보였습니다. 선형의 결합으로 비선형적 분류의 형태를 나타내는 SNN의 경계만으로는 클래스를 분류하는 정확한 경계를 나타내기 힘들었던 것으로 보입니다.

미성형 { 2번 클래스 } 에 대해선 SNN과 GBC 가 비슷한 성능을 보였지만 정확한 예측을 해내기엔 어려워 보였습니다, 이는 미성형의 원인이 수치적, 조건적 결합을 아우른 모든 경계에서 이상만으로 예측해 내기 어려운 클래스일 것으로 생각됩니다. 이는 오히려 해당 결과에 매우 연관도 높게 작용하는 특성이 적용되지 않았거나, 데이터의 분포를 재정의할때의 오류로 생각됩니다.

RFC 또한 우수한 성적을 보였지만, 편향된, 고차원의 데이터 분포에 강하지 않다는 한계를 명확히 보여주었습니다. 데이터 셋에서 선택된 특성이 25개이고 가짓수가 만개를 넘지 않기 때문에 성능개선이 이뤄지지 않을 가능성이 있는 한계를 보여주었습니다. 특히 1, 2 번인 가스와 미성형에 대해 약한 모습을 보였던 이유는 희소한 데이터의 예측이 어려운 RFC의 한계로 보입니다.

GBC가 최종적으로 가장 높은 성적을 보여주었습니다, 데이터의 편향적 수치관계, 조건관계에 따른 이해 모든 측면에서, 약한 학습기들의 가중치를 할당한 결합으로 하여금, 가능한 모든 탐색 조합을 활용할 수 있었을 것이라 분석됩니다. 때문에 오작동의 발생율 자체는 적지만 적은 발생율에도 민감한 분야에서의 예측이 필요한 경우, 유용하게 사용될 수 있을 것으로 보입니다.

다루어진 데이터 셋 {Domain} 과 같이 예측하고자 하는 분류값의 편향이 심하고, 빈도가 매우 적은 경우 학습의 기회가 적게 적용되는 AI 학습기에게는 올바른 예측을 하기에 어려움이 있습니다. 해당 문제를 해결하기 위해서 적은 빈도의 데이터를 Oversampling, 복제하여 학습에 높은 빈도로 적용시킬 수 있는 방법이 존재하나, 가중치를 차등 적용한 SNN이 GBC에 비해 성능이 좋지 않았던 것으로 미뤄볼 때 뛰어난 성능이 관측되진 않았을 것이라 생각됩니다.

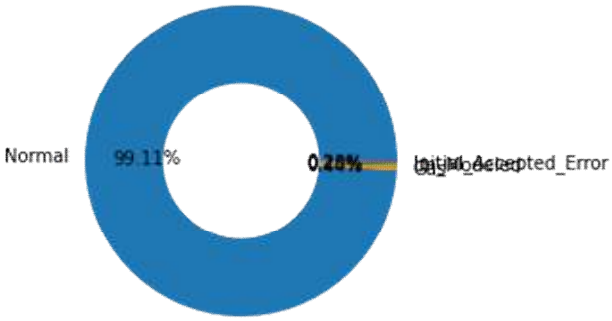

그림 1 : 실제 데이터의 원인 분포도

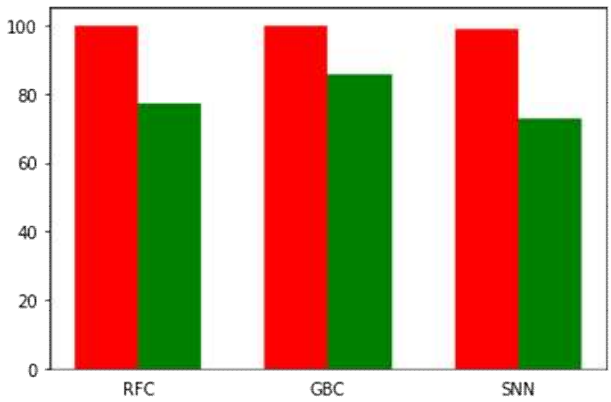

그림 2 : 모델별 정확도(적) f1 매크로평균 점수(녹)

□ 중소제조기업에 미치는 파급효과

○ Domain, 데이터의 편향, Boosting 알고리즘

- 데이터의 편향성

* Label 이 없더라도 Clustering 알고리즘을 통한 데이터 셋 내 특징들의 집합으로 원인, 이상여부, 품질 이상 등을 공통적으로 가지는 경우가 확인 된다면 Domain을 이해 할 수 있겠지만, 시스템 적으로 대규모, 다량의 설비의 정보를 실시간으로 다룰 수 없다면 기기의 정비와 품질관리에 막대한 예산이 소모 될 수 있습니다.

해당 도메인과 같이 예측이 어렵고 잘못된 예측에 민감한 분야에 대해선 위와 같은 앙상블 {Ensemble} 기법의 알고리즘을 활용해 도메인의 대해 정답에 집중된 결과를 산출하는 Boosting 알고리즘을 사용하여 보다 정확한 예측이 가능해 보입니다.

때문에 GBC와 같은 Boosting 알고리즘을 활용한다면, 해당 도메인과 같이 데이터 셋의 비율에 비해 오작동, 오류의 사례가 극히 적게 편향된 경우, 금융권의 사고, 제조현장에서의 오류, 정밀 의료기기의 오작동 등의 문제에서의 오작동의 가능성을 계산하는데 유리하게 사용될 수 있다고 생각됩니다. 특히 중소제조기업과 같이 대규모 설비를 다 갖추지 못한 상황에서 경제적으로 관리 및 점검이 가능하다는 점에서 유용하게 사용할 수 있다고 생각합니다.