From Artificial Intelligence to eXplainable Artificial Intelligence in Industry 4.0: A survey on What, How, and Where

비즈니스인포메틱스학과
2022144203 변재현

1. XAI의 연구배경

연구배경

- 산업에서 데이터에 따라 처리를 하고 의사결정을 하는 것은 생산성 향상에 기여한다.

- 비즈니스 프로세스에 인지적 통찰력을 더해줄 수도 있다

- AI에 기반한 시스템이 전문가들에 의해 쓰여지려면 그들을 설득 해야 한다

2. 논 문 의 방 법

본 연구는 4차 산업에서의 AI, XAI 기반 방법 및 응용에 대한 포괄적인 조사를 제공하는 것을 목표로 한다.

추출한 1000개 이상의 출판된 글을 사용했다.

데이터 세트는 “4차 산업 ", " IoT ", " 인공지능 ", " 설명 가능한 인공지능“키워드를 사용하여 추출하였다.

3. 4 차 산 업 혁 명 이 란

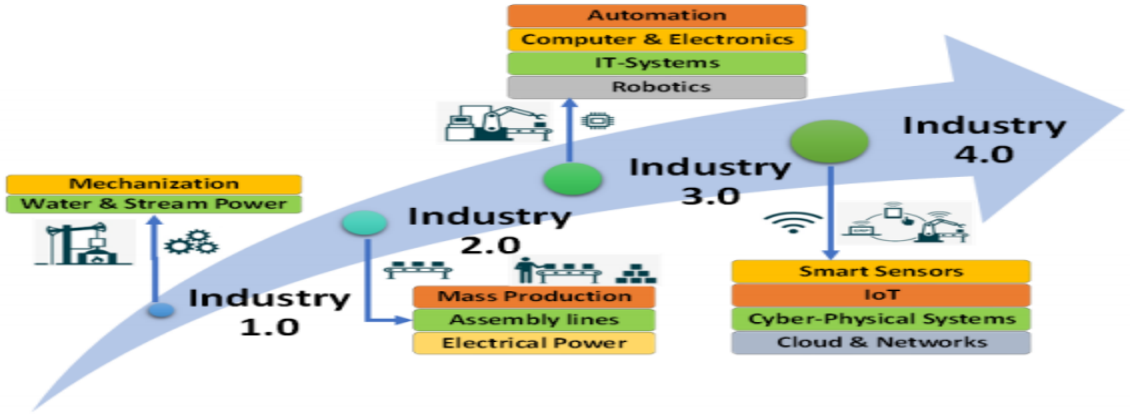


Fig. 1: A schematic illustration of the four industrial revolutions.

제조 및 여러 산업에서 자동화 및 디지털 데이터 전송이 발전된 방향으로 해석된게 4차산업 혁명의 큰 부분이다.

4. 4차 산업 혁명에서 AI의 역할

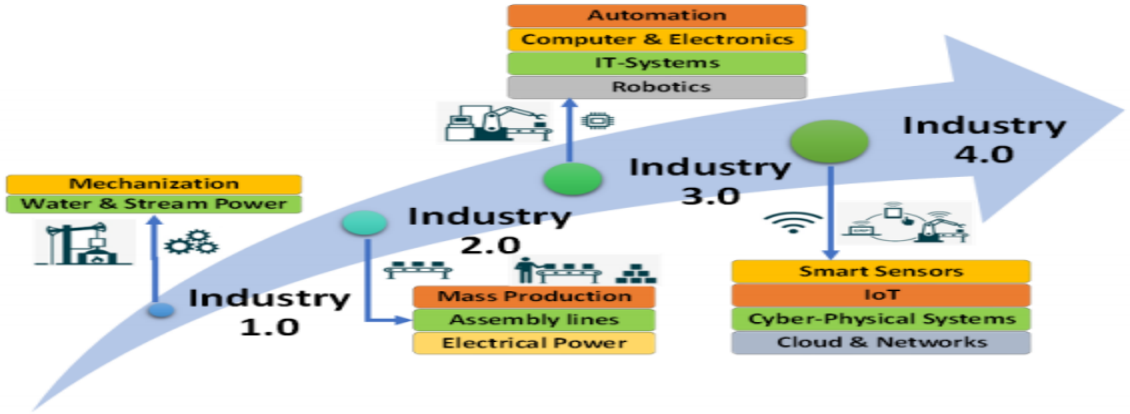


Fig. 1: A schematic illustration of the four industrial revolutions.

4차 산업 혁명에서 AI는 강력한 수준의 자동화로 탁월한 운영 성능과 생산성을 제공한다.

5 산업에서 사용되는 AI 기술들

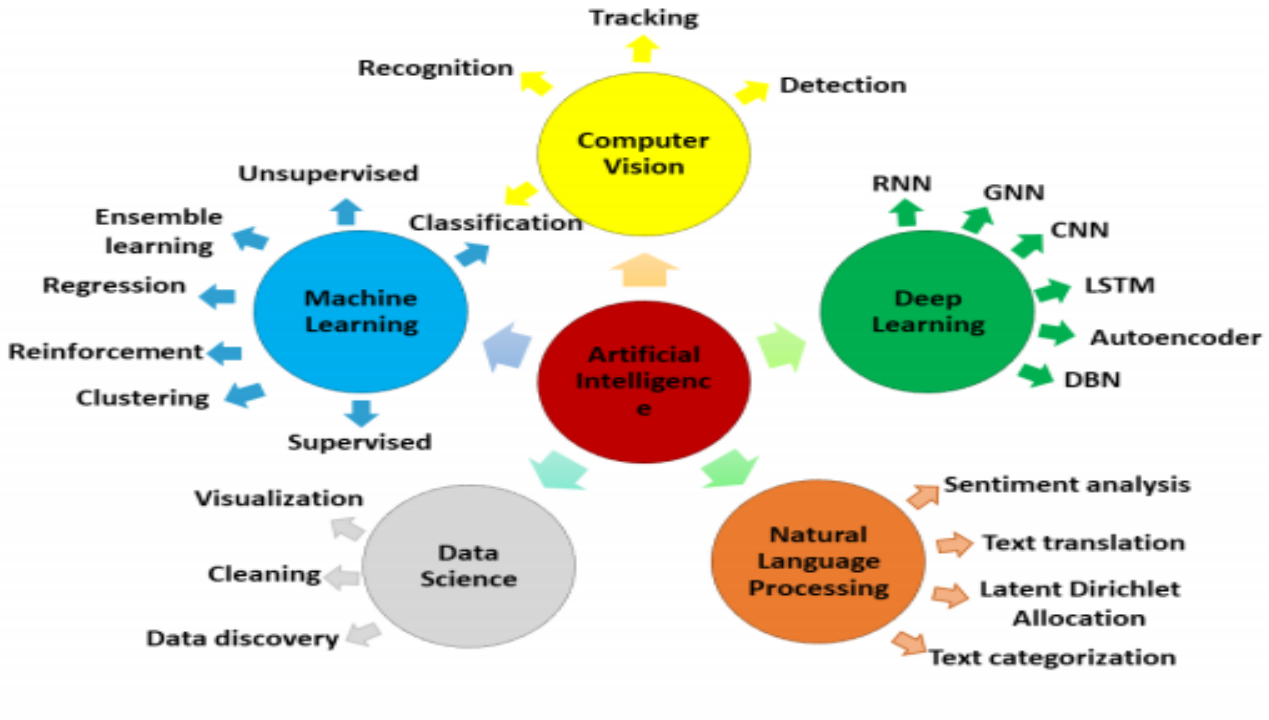


Fig. 3: The main AI-related areas in Industry 4.0.

인공지능은 소프트웨어, 시스템, 기계 및 장치가 자신의 경험을 감지, 지각, 개발, 이해 및 학습하거나 인간 활동을 확대할 수 있도록 하는 여러 기능들을 통합시킨다.

머신러닝(ML)은 인공지능의 기본 원리 중 하나이고
데이터 세트나 명령어를 통해 학습합니다.
ML 기반 방법은 학습을 통해 시스템 결과를 자동으로
학습하고 개선합니다.

**머신러닝은 주로 비지도, 지도, 강화 접근 방식을
사용한다.**

7. Machine Learning의 종류

SVM
패턴 인식, 자료 분석을 위한 지도 학습 모델이며, 주로 분류와 회귀 분석을 위해 사용한다.
DWT(discrete wavelet transform)
주어진 신호를 여러 집합으로 분해하여 변환하는 것
신호 코딩에 사용되며, 데이터 압축을 위해 이산 신호를 더 중복된 형태로 표현하기 위해 사용된다.
Fast Fourier Transform, Principal Component Analysis
엔지니어가 가진 주파수(외부적 요인으로 생긴 주파수)와 각 주파수에서의 진폭을 결정하는 데 도움이 된다
Gaussian Mixture Models
가우스 혼합 모델은 모든 데이터 점이 알 수 없는 모수를 가진 유한 수의 가우스 분포의 혼합물에서 생성된다고 가정하는 확률론적 모델이다.
K- Nearest Neighbors (KNN)
패턴 인식에서, k-최근접 이웃 알고리즘은 분류나 회귀에 사용되는 비모수 방식이다. 두 경우 모두 입력이 특징 공간 내 k개의 가장 가까운 훈련 데이터로 구성되어 있다.
Random forest
랜덤 포레스트는 분류, 회귀 분석 등에 사용되는 앙상블 학습 방법의 일종으로, 훈련 과정에서 구성한 다수의 결정 트리로부터 부류 함으로써 동작한다
Gradient Boosting
일반적으로 의사 결정 트리인 약한 예측 모델의 앙상블 형태의 예측 모델이다.

Machine Learning의 종류

Multi-Layer Perceptrons
입력 노드를 제외하고, 각 노드는 비선형 활성화 함수를 사용하는 뉴런이다.

context-aware intrusion detection system
기존의 세 가지 이상 감지 방법으로 제기된 문제를 해결하고, 보다 정확한 기준선 모델을 생성하여 감지 정확도를 향상시킨다.

Autoencoder
Auto Encoder란 입력 데이터를 압축시켜 압축시킨 데이터로 축소한 후 다시 확장하여 결과 데이터를 입력 데이터와 동일하도록 만드는 모델

Linear Regression
종속 변수 y와 한 개 이상의 독립 변수 x와의 선형 상관 관계를 모델링하는 회귀분석 기법이다.

Clustering
주어진 데이터들의 특성을 고려해 데이터 집단을 정의하고 데이터 집단의 대표할 수 있는 대표점을 찾는 것으로 데이터 마이닝의 한 방법이다

K-means
주어진 데이터를 k개의 클러스터로 묶는 알고리즘으로, 각 클러스터와 거리 차이의 분산을 최소화하는 방식으로 동작한다

DBScan(Density-based spatial clustering of applications with noise)
밀도 기반의 클러스터링은 점이 세밀하게 몰려 있어서 밀도가 높은 부분을 클러스터링 하는 방식이다.

Deep Learning

Deep Learning (DL)은 머신러닝의 일부분이다.
인간이 지식을 얻는 방법을 모방한 것이다.
시스템이나 기계가 층을 통해 정보를 처리하고, 분류, 해석, 결과를 예측하도록 지시한다.

Neural Networks 인간의 신경망과 같은 원리로 작동되며 여러 개의 알고리즘이 변수들간의 상관관계에 관한 정보를 얻으면서 데이터를 학습된다.

Convolutional Neural Networks (CNN) CNN은 이미지를 날것(raw input) 그대로 받음으로써 공간적/지역적 정보를 유지한 채 특성(feature)들의 계층을 빌드업합니다. CNN은 이미지 의 부분적 특징을 보며 한 픽셀과 주변 픽셀들의 연관성을 살린다.

Recurrent Neural Networks (RNN) RNN은 은닉층의 노드에서 활성화 함수를 통해 나온 결과값을 출력층 방향으로도 보내면서, 다시 은닉층 노드의 다음 계산의 입력으로 보내는 특징을 갖고있습니다.

Neural Networks (GNN) 그래프 데이터 구조로 가장 잘 표현되는 데이터를 처리하기 위한 신경망의 한 종류입니다. 점이 주위 점들의 특징에 의해 정의된다는 것이다.

ANN
시냅스의 결합으로 네트워크를 형성한 인공뉴런(노드)이 학습을 통해 시냅스의 결합 세기를 변화시켜, 문제 해결 능력을 가지는 모델. 은닉층에서는 활성화함수를 사용하여 최적의 Weight와 Bias를 찾아내는 역할을 합니다

RNN
유닛간의 연결이 순환적 구조를 갖는 특징을 갖고 있고 신호의 길이가 한정되지 않은 동적 데이터를 처리한다.

Hidden Markov Model
시스템이 은닉된 상태와 관찰가능한 결과의 두 가지 요소로 이루어졌다고 보는 모델이다. 관찰 가능한 결과를 야기하는 직접적인 원인은 관측될 수 없는 은닉 상태들이고 도출된 결과 들만이 관찰 가능하다.

10. Deep Learning

Backpropagation neural networks
신경망의 결과값에서 미분을 통해 기울기(변화량)을 통해 weight이랑 bias값 을 갱신 시킨다.

Deep Belief Networks (DBN))
심층신뢰망은 입력층과 은닉층으로 구성되어 있는 RBM을 층층이 쌓아 올린 형태로 연결한 신경망이고 여러 번의 사전 학습을 통해 가중치를 어느 정도 보정하고, 역전파 및 피드포워드 알고리즘을 통해 최종 가중치를 계산한다. 이러한 특성은 학습 데이터의 양이 적을 때 굉장히 유용하게 사용됩니다

LSTM
장단기 메모리는 순환 신경망 기법의 하나로 셀, 입력 게이트, 출력 게이트, 망각 게이트를 이용해 기존 순환 신경망의 문제인 기울기 소멸 문제를 방지하도록 개발되었다

Multilayer Gated Recurrent Unit (MGRU)
GRU는 LSTM의 장기 의존성 문제에 대한 해결책을 유지하면서, 은닉 상태를 업데이트하는 계산을 줄였습니다. 다시 말해서, GRU는 성능은 LSTM과 유사하면서 복잡했던 LSTM의 구조를 간단화 시켰습니다.

Adaptive Kernel Spectral Clustering
구조가 손상되지 않은 단계에서 초기화 및 보정되며 보정 프로세스는 조기 손상을 감지하고 잘못된 경보 수를 최소화하는 데 매우 중요합니다.

Natural Language Processing (NLP)
전달되는 인간 언어를 학습, 지각 및 이해하는 시스템 또는 기계의 능력이다.
ML, DL 기반 기술을 사용하여 인간 언어에서 인사이트를 도출한다.

-감성분석
-Latent Dirichlet Allocation(문서의 텍스트를 특정 주제로 분류한다)
-Text categorization

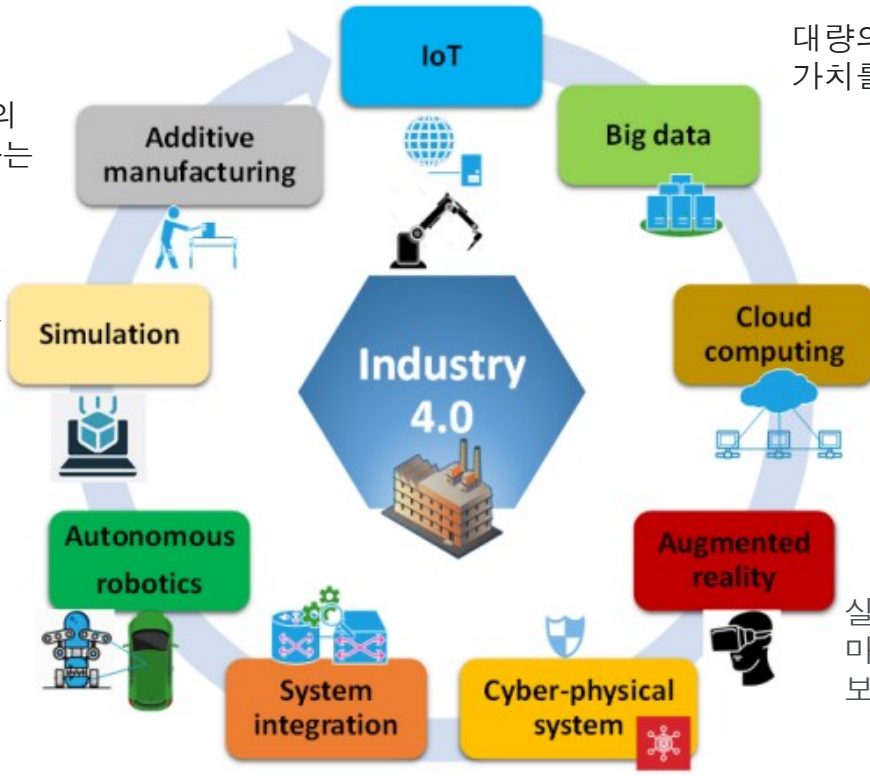
컴퓨터 비전: 컴퓨터가 디지털 비디오와 이미지를 '보고, 해석하고' 이해하여 시각적으로 결론을 내리고 실제 문제를 해결하는 데 활용할 수 있도록 합니다.

많은 사업들이 **fault detections**
금융권에서는 **fraud prevention**
프린트 산업에서는 **deep learning-based automated Quality Control**

4 차 산 업 혁 명 의 기 술 들

다른 사물과 데이터를 송수신할 수 있는 센서와 소프트웨어,
기타 기술을 장착하고 서로 연결된 물체와 기기

3D프린팅 기술을 제조업에 접목한 것으로써, 제조산업의 원자재, 부품, 제품을 생산하는 것을 말 합니다.



대량의 정형데이터 비정형의 데이터로 가치를 추출하고 결과를 분석하는 기술

실제로 실행하기 어려운 실험을 간단히 행하는 모의실험

인터넷을 통해 서버, 스토리지, 데이터베이스, 네트워킹, 소프트웨어, 분석, 인텔리전스 등의 컴퓨팅 서비스를 제공

인간의 통제에 의지하지 않고 행동하는 로봇

실제공간에서 가상의 사물이나 정보를 합성하여 마치 원래의 환경에 존재하는 사물처럼 보이도록 하는 컴퓨터 그래픽 기법이다

하드웨어·소프트웨어·네트워크 등 유형의 제품과 컨설팅·시스템 설계 및 유지보수 등 무형 서비스 기술을 통합

우리가 살아 가는 물리 세계와 사이버 세계와의 융합을 추구하는 새로운 패러다임

Technology	Application	Technology Description	Main Contribution Summary
IoT [1]	Manufacturing	Networking of smart physical objects (sensors, devices, machines, cameras, vehicles, buildings). Allows exchange and collection of data, communication, and collaboration of objects.	Presented a brief summary of Industry 4.0 and associated technologies.
Big data [2]	Business	Selection and evaluation of extensively available data sets. Applying a set of methods to clean, record the data. Present observations during data processing with various variety, higher velocities in greater volumes of data.	Investigate the operational and necessary impacts of Big data. Presented systematic analysis and case study conclusions. Studied the applications and highlight future directions.
Cloud Computing [3]	Industry 4.0.	System for establishing online storage functions (data applications). Models, and programs in a virtual server.	Explored emerging IT trends: IoT, Big data, and Cloud Computing. Investigates their industrial implementation.
Augmented reality [4]	Industry	Collection of HCI methods can insert virtual objects. Collaborate in the physical environment.	Presented overview of the importance of Augmented reality.
Cyber-physical system [5]	Industry 4.0	Set of advanced technologies, link the processes of physical resources and computational capacities, control physical systems, while designing a virtual model.	Reviewed current research trends of cyber-physical systems, their applications in industries and identifies challenges.
System Integration [6]	Manufacturing	Establishment of a standard data network system. Allows various organizations and departments to be integrated and linked, where a smooth collaboration and computerized value chains are feasibly formed.	Review important aspects of additive manufacturing: new improvements in process development and material science. Analyze modern science and technological trends) and highlight its possible applications.
Autonomous robotics [7]	Automation & Robotic	Instrument and machinery that automate operational processes, consisting of collaborative robotics, enables machines and humans to operate and interact in a distributed learning environment.	Review on the progress of robotic and mechanization technology for industrial applications.
Simulation [8]	Manufacturing	Technologies that illustrate real-life data like products, systems, and humans in the real world, intending for interpretation and affordability of the system, design, experimentation, and live development of the processes.	Analyze ongoing research trends of Industry 4.0, highlights important design systems and technology aims, distinguish its architectural layout, and direct strategic road maps.
Additive manufacturing [6]	Manufacturing	Process of combining objects in subsequent layers to create objects using data of 3D model and 'unlock' system choices to accomplish high potential for mass customization.	Focus on the principle concept of Industry 4.0. Identify research gaps between modern systems and Industry 4.0 requirements.

어떠한 기법들이 어떤 기술로써 어떤 산업에 쓰이는가

Broad Area	Application	Technology	Methods/ Models	Main Contribution
ML [10]	Design, Manufacturing Operation	Simulation, Cyber physical system, IoT, Big data analytics	Generic ML algorithm	Data analysis and processing for design improvements of ships manufacturing.
ML [11]	Tracing tracking in production	Cloud computing	Ontology-based Interpretable Model	Product-in-use assessment in Industry 4.0.
ML [22]	Automotive industry	IoT, Cyber physhopsical system	Not Specified	Customized and advanced services Industry 4.0
ML DL [12]	Manufacturing Industry	System integration, IoT	SVM, DWT Transformation, Fast Fourier Transform, FPCA, Gaussian Mixture Models, GRNN, Hidden Markov Model, KNN, RNN, ANNs, DBN	Overview of descriptive, predictive, and prescriptive prognostic maintenance and analytics in Industry 4.0.
ML [14]	Assembly systems	IoT, Cloud computing , Additive Manufacturing	Autonomous decision using ML	Explore the impact of Industry 4.0 strategic, tactical, operational levels of assembly lines.
ML [15]	Production line failure	IoT, Simulation	Random Forest, Gradient Boosting, MLP, ANN	Presented case study, a real-time early detection and product failure system for Industry 4.0.
ML [16]	Manufacturing Process	IoT, Cyber physical system,	Clustering, Autoencoder, Linear Regression, K-mean, Random Forest, and DBScan	Presented a context-aware intrusion detection system for Industry 4.0.
ML CV [23]	Tracking human	IoT	Feature and Blob based method	Presented a features-based person tracker model for industrial environment.
ML CV [24]	Manufacturing Industry	Big data, IoT	SVM, Decision trees, Random forests, Logistic regression, KNN	Identification and classification of materials in the context of Industry 4.0.
ML [25]	Maintenance Industry	IoT, Cyber physical system	K-means, Gaussian mixture	Presented a real-world implementation cycle, for knowledge discovery using machining learning.
NLP [26]	Human Resource Management	IoT	Text Mining	Presented a tool to automatically determine Industry 4.0 impact on human resource management.
NLP [27]	Supply chain	IoT	Latent Dirichlet Allocation	Multi-tier supply chain in Industry 4.0.
NLP [28]	Industry	IoT	Neural Network	Utilized attention mechanisms, presented a NN based model for the translation of natural language used in SQL database system.
DL [17]	Manufacturing	IoT, Cloud computing	CNN	Presented robust real time product inspection system based on fog computing and DL.
CV DL [29]	Printing Industry	IoT, Cloud computing	DNN	Presented a deep learning based automated Quality Control system for Printing Industry 4.0.
CV DL [30]	Manufacturing	IoT, Cloud computing, Simulation	DBN-DL	Developed and IoT sensor based fault detection system based on deep learning for Industry 4.0.
DL [19]	Fault Detection	IoT	DNN CNN, (LiftingNet)	Presented a DL network for fault classification and to adaptively train features from noisy data without previous knowledge.
DL [18]	Manufacturing	IoT	CNN, Autoencoders	Presented a DL based method for early fault identification in time-varying conditions situations.
ML DL [20]	Manufacturing	IoT	MGRU, LSTM Multilayer LSTM and SVM.	Presented a MGRU method for spur gear fault diagnosis. Evaluation of the methods classification accuracy is made using LSTM, MLSTM, and SVM.
DL [21]	Predictive Maintenance	IoT	AKSC, LSTM-RNN	Presented a novel method using DL, to classify anomaly behaviors from multiple degradation features.
DL [31]	Human Detection	IoT	Faster-RCNN, SSD, YOLOv3	Detecting human in complex industrial environment using different deep learning methods.
ML [32]	Various Industries	Not Specified	Interpretability and explainability ML	Real-world application of explainability methods used for search and recommendation systems. (lending, leasing, sales, and fraud discovery)
DL [33]	DFKI-Smart-Lego-Factory	IoT	Global and Local explanations for process outcome predictions using the applied deep neural network.	Presented a utilization case for DL method for prediction. Illustrated the insights of decision-making and the purpose of the intelligence in Industry 4.0.
ML [34]	Anomaly Detection	Big data, IoT	Explainable Feature importance Isolation forest algorithm.	Introduced a method for determining a feature importance in Anomaly identification in Industry 4.0.
ML [35]	Condition-based monitoring, predictive maintenance,	IoT, Cyber physical system, Cloud computing	Explainable ML quantitative association rule mining.	Explainable and Predictive model for IoT applications in Industry 4.0

XAI ML [36]	Manufacturing System Operations	IoT, Cyber physical system,	Visualization and Explainable ML.	Presented visual features of real-time predictive analytics and multivariate time series method to highlight possible errors, warnings, and malicious intrusions attacks.
XAI ML DL [37]	Anomalous behavior prediction	IoT	Explainable interpretable regularize logistic regression, Feature extraction (shape based, direct, CNN, RNN, and selection (Kolmogorov–Smirnov (KS))	Introduced Interpretable Anomaly Prediction model for Industry 4.0. Identify anomalies in the current data and also be able to predict it probability in the future data.
XAI [38]	Smart Industries	IoT, Cyber physical system	Integration of meta learning and AI based approaches	Implement integration of XAI methods and Deep meta-learning paradigms for Cyber physical systems.

article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2022.3146552, I
Transactions on Industrial Informatics

TABLE II: Summary of various AI and XAI based methods used in various industrial applications.

Broad Area	Application	Technology	Methods/ Models	Main Contribution
XAI ML [39]	Business	IoT	Visualization and ML (Shapley values, XGBoost predictive classification algorithm.)	Proposed an XAI model that can be applied to justify why a customer buys or leaves non-life insurance coverage.
XAI ML [40]	Industrial machinery	IoT	Visualization and ML techniques local and global explanations, Random forest, ELI5 and LIME.	Implementation and explanations of a residual life estimator design based on machine learning employed to industrial data.
XAI [41]	Manufacturing	IoT	Visualization Nonlinear modeling with SHAP values data-driven decision model. [41]	Proposed a data-driven decision model to improve process quality in manufacturing by combining nonlinear modeling and SHAP values from the field of explainable AI.
XAI DL [42]	Manufacturing	IoT	XAI methods Smoothed Integrated Gradients, Guided Gradient Class Activation Mapping DeepSHAP CNN classifier.	Presented a DL based classification method for fiber layup fault identification in the automatic composite manufacturing.
XAI DL [43]	Predictive Process	IoT	Local post-hoc explanation DL, Surrogate (Decision tree)	Introduced a new local post-hoc explanation method for monitoring problems in the predictive process.
XAI [44]	Fault Detection Diagnosis	IoT	Unsupervised classification, SHAP and Local Depth-based Feature Importance Isolation Forest (Local-DIFFI).	Presented a fault diagnosis and anomaly detection techniques in rotating machinery to interpret black-box models.
XAI DL [45]	Manufacturing	IoT	Synthetic Minority Oversampling Technique (SMOTE), Random forest and association Rule Mining, Lightweight on-line detector of anomalies, Minimum Covariance Determinant PCA, t statistic.	Presented an XAI-based method for fault analysis and penetration harvesting for steel plates manufacturing.

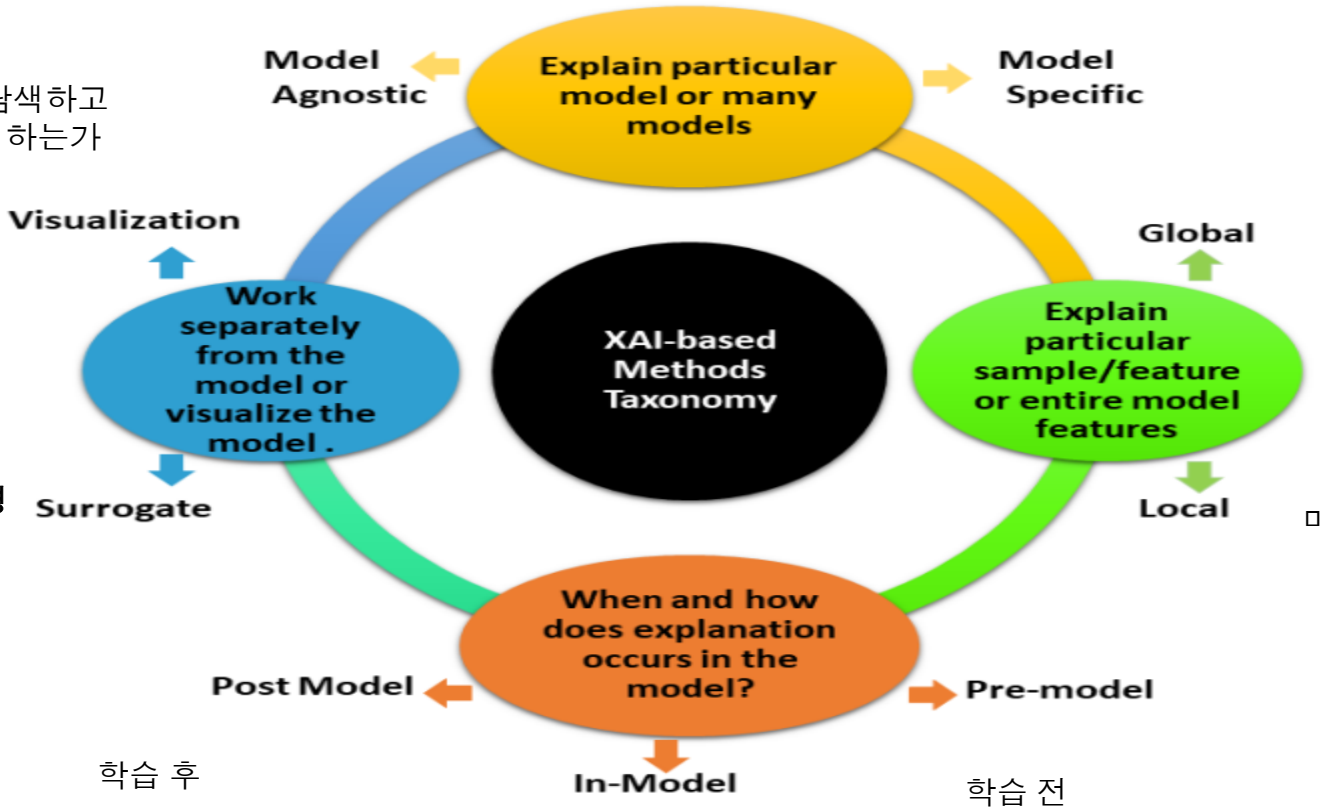
XAI 방법의 분류 체계

학습에 사용되는 모델과 설명에 사용되는 모델을 분리

설명에 있어서 특정 모델을 사용

,내부의 숨겨진 패턴을 탐색하고
분석하기 위해 시각화를 하는가

모델의 거시적 설명에
초점을 맞춥니다.



유사 모델을 만들어 설명

미시적 설명

학습 후

학습 전

XAI 시스템은 예측 유지보수 시스템을 생산 방식에 적용시킴으로써 제품의 품질을 향상시키는 데 도움이 될 수 있는 기술의 자동 학습을 향상시킬 수 있습니다.

로봇 또는 코봇으로 시각적 조사를 대체를 통한 효율적 품질 검사의 수행

XAI의 기술 구현은 미래의 소비자들을 위한 제품과 설명의 전형적인 잠재력을 드러내는 것에 관한 것이다. IoT는 4차 산업에서 상당한 역할을 수행하며 패턴과 행동을 인식하고 관련 산업에 실시간 데이터를 제공합니다

19. Challenges

1. AI 기반 시스템, 방법 및 알고리즘이 전력 소모가 커 효과적으로 작동하려면 점점 더 많은 양의 코어와 GPU가 필요합니다

2. AI 기반 방법의 알려지지 않은 특성이다: ML 및 DL 모델이 출력을 예측하는 성능이 좋다고 하더라도 특성을 모른다면 **인간 수준의 정확한 설명에 가까워질 수 없다.**

3.또한 AI 방법론은 종종 하이퍼파라미터 최적화, 미세 조정, 거대한 데이터 세트, 강력한 컴퓨팅 기능, 데이터에 대한 지속적인 교육을 필요로 한다.

4.사이버 보안

5. 데이터의 편향된 성질은 실시간 응용 프로그램에서 ai 성능에 영향을 준다
이를 방지하기 위해 인간수준의 설명이 필요하다

6. XAI를 얼마나 믿어야 하는가에 대한 분석

7. 너무 복잡해서 비논리적인 것을 간파해야 한다.

A Survey on Explainable Artificial Intelligence Techniques and Challenges

비즈니스인포메틱스학과
2022144203 변재현

연구배경

- 2026년에 299.64 billion Dollars

- AI 기반 시스템은 여러 과제들에 대해 전문가들을 능가한다

- 하지만 복잡하고 투명성이 부족하다

광범위한 XAI 기술의 해석이 어느 정도까지 이해 가능한지
어떠한 부분이 시스템에 통찰력을 제공하는 데 기여하는지
논의한다.

3. AI의 설명정도를 측정 하는 측면들

Accountability: 파트너, 사용자, 그리고 시스템을 사용하는 다른 사람들에게 의사 결정의 정당화의 정도

Responsibility: AI 예상치 못한 결과와 실수에 대해 반응하는 정도

Transparency: 용어, 형식, 언어로 된 의사 결정에 대해 고객의 이해할 수 있는 정도

Fidelity 의사결정 과정에 시스템의 설명이 최대한 반영되는 정도

Bias AI 시스템이 불충분한 데이터 수집 또는 모델 구축의 결과 또는 인간성 편견의 결과로 시스템에 대한 편향된 관점을 학습하지 않는 정도

Causality 데이터에서 훈련된 모델이 적절한 통찰력을 제공할 수 있는 정도

Fairness 시스템이 설명된 후 AI 시스템이 내린 결정이 공정하다고 말할 수 있는 정도

Safety 시스템의 결정에 의존할 수 있는 정도

XAI 기반 접근법:

1. Intrinsically interpretable: 직관적으로 해석이 가능한 접근법. 다른 과정 없이 해석이 가능하다

Ex, linear Regression, KNN, Baysian Models, Logistic Regression, Decision Tree, Decision Rule, GLM, GAM

2.Agnostic technique: 독립적 모델로 해석
훈련받은 모델을 블랙박스로 취급할 수 있고 이는 내부에 대한 해석이 필수가 아니라는 것을 의미한다.

3.Instance Based Explanation: 모델을 데이터셋의 인스탠스(한 열)을 가지고 해석.

4. Propagation Based Methods:피쳐들의 변화에 따른 결과의 변화로 해석

1) Linear Regression: Linear Regression은 회귀 계수가 명확한 의미를 가질 때 해석할 수 있는 것으로 간주된다. 이 기법의 예측은 회귀계수의 가중치 합으로 한다.

가중 합계는 시스템의 투명성을 제공한다

6. intrinsically interpretable techniques-Logistic Regression

로지스틱 회귀는 선형 회귀의 확장으로 분류 문제에 대한 해결책을 제공한다.

직선 관계 대신 [0,1] 사이의 출력을 압축하여 클래스의 출력 가능성을 예측하고 무게는 기울기 방향의 해석을 나타냅니다.

7. intrinsically interpretable techniques-Decision Tree Model

트리 기반 모델은 시스템 feature 세트의 컷오프 값을 기반으로 하여 데이터를 반복적으로 분리하여 작동한다.

그래서 이 모델은 선형 회귀 분석과 로지스틱 회귀 분석이 실패하는 데이터의 상호작용적 특징을 예측하는 데 적합하다.

트리모델에서 사소한 입력 변동은 결과에 상당한 영향을 미칠 수 있습니다.

intrinsically interpretable techniques-Decision Rule

간단한 if-else 문장은 시스템의 의사결정 규칙 역할을 한다.

터미널을 사용하여 트리로 바꿀 수 있으며 명확성과 설명 측면에서 뛰어나다.

intrinsically interpretable techniques-Generalized Linear Models

특징의 가중 합계를 보존하고 가우스 이외의 분포를 허용한다.

GLM은 잠재적으로 비선형 함수를 통해 워크아웃 분포의 예측 평균을 가중 합계와 연관시킨다.

intrinsically interpretable techniques-Generalized Linear Rules

GAM은 간단한 가중치 요구사항을 완화하고 대신 결과가 각 속성의 임의 함수의 조합으로 특징지어질 수 있다고 가정한다.

GAM은 선형 모델의 가정 일부를 해독하는 데 유용하나 선형 모형 확장은 더 복잡하고 해석하기 어려운 모형을 만듭니다.

Agnostic Teqniques

Global Model-Agnostic

Partial Dependence Plot(PDP)
Accumulated Local Effect(ALE)
Feature Interaction
Permutation Feature Importance
Global Surrogate
Prototypes and Criticisms

Local Model-Agnostic

Individual Conditional Expectation(ICE)
Local Surrogate
Shapley Values
SHAP(Shapley Additive explanations)

1. Model Flexibility-해석 방법은 랜덤 포레스트 및 심층 신경망과 같은 모든 머신 러닝 모델과 함께 작동할 수 있다.
2. Explanation Flexibility=경우에 따라 선형 공식을 갖는 것이 유용할 수도 그래픽의 특징이 중요할 수도 있다.
3. Representation Flexibility=설명 중인 모델과 다른 특징 표현을 사용할 수 있어야 한다.

Partial Dependence Plot(Partial Dependence Plots): PDP는 목표와 특징 사이의 관계가 선형인지, 단조로운지 또는 더 복잡한지 여부를 나타낼 수 있습니다.

$$\hat{f}_S(x_S) = E_{X_C}\left[\hat{f}(x_S, X_C)\right] = \int \hat{f}(x_S, X_C) d\mathbb{P}(X_C)$$

Xs=부분 종속 함수를 표시해야 하는 특성
xc= 기계학습. 모델인 f(hat)에 쓰인 다른 특징들
F(hat) random variables

The partial function $\hat{f}_S$ is estimated by calculating averages in the training data, also known as Monte Carlo method:

$$\hat{f}_S(x_S) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x_S, x_C^{(i)})$$

x(i)CxC(i) 관심 없는 특성값

인과관계 반영된 해석 가능
부분 의존성으로 계산되는 특징은 다른 특징들과 상관관계가 없다고 가정한다.

13. Global Model Agnostic ALE

Accumulated Local Effects(ALE) Plot
평균적으로 특징이 기계 학습 모델의 예측에 어떻게 영향을 미치는지 설명한다.

장점:PDP보다 더 빨리 더 공정성 있게 계산한다.
ALE는 중점을 0으로 하여 해석력이 깔끔하다.

단점:간격이 커짐에 따라 그래프가 불안정 할 수 있다.
구간 간 미치는 영향에 대한 해석은 어렵다.
PDP에 비해서 직관적 해석이 어렵고 복잡하다

Global Model Agnostic Feature Interaction

Feature Interaction

예측 모형에서 특징이 서로 상호작용하는 경우 한 특징의 효과가 다른 특징 값에 따라 달라지기 때문에 예측효과의 단순 합으로 표현할 수 없다.

Aristole's The whole is greater than the sum of its parts

$$PD_{jk}(x_j, x_k) = PD_j(x_j) + PD_k(x_k)$$

1. 두 특성들이 상호작용하지 않으면 부분적 변수를 분리한다

PDjk(xj,xk)는 PDj(Xj)와 PDk(Xk)는 하나의 특징에대한 종속 함수

2. 관측된 부분적 종속 변수와 상호 작용이 없는 분해 함수 사이의 차이를 측정한다.
3. 부분 종속 변수의 분산의 계산
4. 상호작용으로 설명되는 분산의 양은 상호작용 강도 통계량으로 사용됩니다.

15. Global Model Agnostic Feature Importance

Feature Importance: 특징과 실제 결과 사이의 관계를 끊기 위해 모델의 예측 오차의 증가를 측정하여 특징의 값을 변경 시키면서 타 특징의 관련성을 증가시킨다.

1.특징값을 변경한 후 모델의 예측 오차의 증가를 측정한다.
2.오차가 커진다면 변경된 특징이 중요한 특징이다.
3.만약 오차가 달라지지 않는다면 변경된 특징은 중요하지 않은 특징이다.

장점
다른 특징 들과의 상호작용을 고려한다
모델 재훈련을 필요로 하지 않는다

단점
Label이 있어야 에러값을 특정 할 수 있다.

Global Model Agnostic Global Surrogate

Global Surrogate는 블랙박스 모델과 유사하게 예측하도록 설계된 모델이며 해석을 가능하게 한 모델이다.

Black Box Model:내부 작동에 대한 어떠한 정보도 드러내지 않고 유용한 정보를 생산하는 장치, 시스템 또는 물체이다.

Black Box모델과 대체 모델의 Rsquare로 모델의 완성도를 따져본다.

Rsquare의 좋은 기준이 애매하다.
전체 데이터셋의 유사도는 확인 가능하지만 각 변수의 유서도를 확인이 힘들다.

17. Global Model Agnostic Prototypes and Criticism

프로토타입은 모든 데이터를 대표하는 데이터 인스턴스다.
Criticism은 프로토타입 집합으로 잘 표현되지 않는 데이터 인스턴스이고 프로토타입과 함께 인사이트를 제공한다.

프로토타입과 Criticism은 데이터를 설명하기 위해 머신 러닝 모델과 독립적으로 사용될 수 있지만 해석 가능한 모델을 만들거나 해석 가능한 블랙박스 모델을 만드는 데 사용될 수도 있다.

MMD-critic은 데이터의 분포와 선택된 프로토타입의 분포를 비교한다.

MMD-critic은 두 분포 사이의 불일치를 최소화하는 프로토타입을 선택한다. 특히 다른 데이터 클러스터 에서 점을 선택하는 경우 밀도가 높은 영역의 데이터 점이 좋은 프로토타입이 된다. 프로토타입에 의해 잘 설명되지 않은 지역의 데이터 포인트가 비평으로 선택됩니다.

18. Local Model-Agnostic

Individual Conditional Expectation (ICE) Plot:
The ICE Plots

LIME(Local interpretable model-agnostic explanations)

Shapley 값

SHAP

Individual Conditional Expectation (ICE) Plot: The ICE Plots

특성이 변경될 때 인스턴스의 예측이 어떻게 변경되는지 보여준다

PDP는 ICE Plots Lines들의 평균 값이다

장점: 직관적이다, 여러종류의 관계들에 대해 설명력을 얻는다

단점:개별 인스턴스에서만 좋다

Local surrogate model

로컬 대리 모델은 블랙박스 머신 러닝 모델의 개별 예측을 설명하는 데 사용되는 해석 가능한 모델이다.

Local surrogate model의 구체적인 구현 LIME(Local interpretable model-agnostic explanations)

LIME: 다양한 데이터를 기계 학습 모델에 대입할때 예측의 변화에 대해서 확인하는 것.

$$\text{explanation}(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

Model g minimizes Loss L

X의 설명모델은 이며 설명이 원 모델 f에 대한 에측이 근접한 정도를 계산하여 로스를 최소화 한다.

옴(g)는 모델의 복잡도(낮아야 한다)

장점 **tabular data, text and images** 사용 가능

.

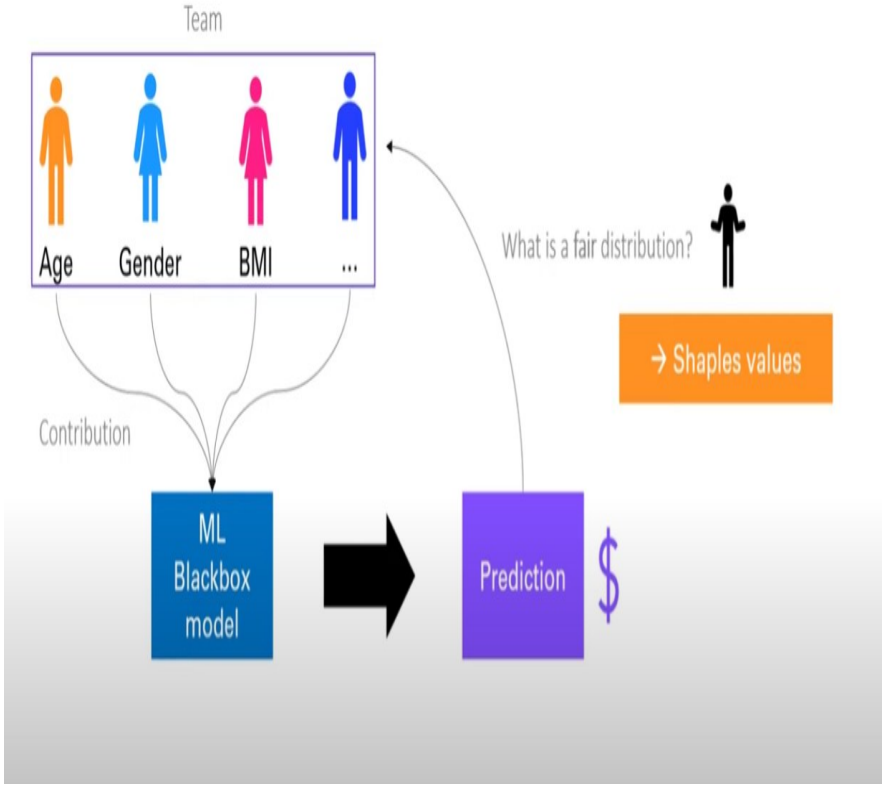
단점:각 응용 프로그램에 대해 서로 다른 커널 설정을 사용해 보고 설명이 적절한지 직접 확인해야 합니다.

21. Model Agnostic-Local Surrogate-Shapley

Shapley **값:**게임 기반 시스템 예측 방법이다
각 피처들의 값이 게임에서 "플레이어 " 로 가정
예측을 " 보상 " 이라고 가정한다.
플레이어들이 보상을 공정하게 나누는 방법(예측 기여도를 피처들이 공정하게 나누는 방법)

실제 예측과 평균 예측 사이의 차이에 대한 특징 값의 기여가 추정된 섀플리 값이다

장점: 피쳐들의 기여도를 공정하게 배분하고 대조적 설명이 가능하다.(셰플리값을 전체 데이터 값의 평균이랑 비교하는 것이 아니라 일부랑 비교 가능하다)
단점: 항상 모든 피쳐들을 사용해야 한다.



Model Agnostic-Local Surrogate-SHAP

SHAP의 목적은 최적화된 셰플리값을 통해 예측에 대한 각 특성의 기여도를 계산하여 관측치 x의 예측값을 설명하는 것이다.

LIME과 Shapley value를 활용하여
추정 접근법인 **Kernel SHAP, 트리기반 Tree SHAP**

장점: 트리기반 SHAP모델은 계산이 빠르고 글로벌 모델을 해석하기 적합하다

커널SHAP은 피쳐들 간의 상호작용을 무시한다(랜덤 피쳐로 변하기 때문)

Propagation Based Methods

Forward Propagation Based Methods: 입력 계층에서 출력 계층으로 순서대로 신경망의 중간 변수(출력 포함)를 계산하고 저장하는 것을 말한다.

Back-Propagation Based Methods: 이 접근 방식은 입력 기능을 사용하여 결과의 미분을 취하는 것을 기반으로 한다.

24. Instance-Based Explanations

Instance-Based 접근 방식은 시스템을 해석하지만, 대규모 설정의 경우 계산적으로 비용이 많이 들 수 있으며 설명 생성 시간이 충분하지 않다.

Break Down
전체 결과에 대한 영향을 기반으로 특성을 식별한다.

Adversarial
기계 학습 모델이 잘못된 예측을 하게 하는 인스택스 피쳐의 작은 변화이다.

Feature Selection
이 프레임워크에서 피처 셀렉터는 선택한 피처와 반응 변수에 대한 정보를 최대화 한다.

Counter Factual
A counterfactual explanation 만약 X가 없었다면 Y가 없었을 것이다

Adversarial Example

Adversarial Example: 기계 학습 모델이 잘못된 예측을 하게 하는 인스택스 피쳐의 작은 변화이다.

Adversarial example과 조작되어야할 인스턴스 사이의 거리를 최소화하면서 예측을 원하는 예측에 도달한다.

$$loss(\hat{f}(x+r), l) + c \cdot |r|$$

X는 벡터화된 픽셀의 이미지, r 은 adversarial 이미지를 만들때 픽셀의 변화 (x+r 새이미지 생성), l는 원하는 결과값, c 는 이미지와 예측값의 거리의 균형을 맞춥니다.

Counter Factual

A counterfactual explanation 만약 X가 없었다면 Y가 없었을 것이다

Advantages
머신러닝 없이도 사용 가능하다
데이터나 모형에 액세스할 필요가 없이 API와 같은 모델의 예측 함수에만 접근이 되면 된다.
실행하기 비교적 간단하다

Disadvantages
하나의 인스탠스에 여러해석이 가능할 수 있다.

Challenges

1.기존 모델을 설명하기 위해 또 다른 블랙박스 모델을 만들고 있는건 아닌지 확인해야 한다

2. 장기적으로 보면 단순 설명만 제공하는 것이 아니라 XAI 기반 시스템을 배치하기 위한 표준 프레임워크를 개발해야 한다.

3. 표준 프레임워크는 대부분의 애플리케이션에 적용 가능하도록 공동 속성을 넣어 유연성과 적응성을 제공해야 한다.

감사합니다

A Survey of Data-Driven and Knowledge-Aware eXplainable AI

비즈니스인포메틱스학과
2022144203 변재현

연구배경

- 모델의 메커니즘중에 설명이 불가한 부분이 있다.

- 블랙박스 모델의 예측은 편향된 데이터로 인해 잘못 될 수도 있다

- ML기법들은 복잡하고 투명성이 부족하다

2. Phases of life cycle of XAI

1. 방법- 블랙박스 모델을 설명하기 위한 보다 우수한 방법의 개발

2. 평가-설명의 효과성 평가에 대한 방법 설계 또는 메트릭 정의

3 적용– 실제 상황에서의 XAI 적용

데이터기반 eXplanation AI

사전 지식 등 외부 정보 없이 순수하게 데이터로부터 블랙박스 모델을 설명을 생성하는 방법을 말한다. 설명가능한 데이터를 선택하는 것에서 시작한다.

블랙박스 모델을 설명한다는 것

입력과 출력 사이의 설명 불가능한 속성들이 있는 블랙박스 모델을 설명하는 것을 목표로 한다.

해석의 범위에 따라서 Global**과** Local**로 나뉜다**.

Data Driven XAI의 글로벌 해석의 3가지 방법

글로벌 방법은 모델의 논리와 모든 예측에 대한 완전한 추론을 이해하는 것을 목표로 한다.

모델 추출 방법

- 블랙박스 모델과 유사한 모델로 추측 후 설명

특징 기반 방법

- 기능의 중요성 또는 관련성을 추정하여 설명

투명한 모델 설계

- 해석 가능성을 향상시키기 위해 블랙박스 모델 자체를 수정 혹은 재설계

모델 추출의 기본 아이디어는 해석 가능한 것을 배우는 것입니다.

블랙박스 모델의 훈련 데이터 세트 또는 동일한 분포를 가진 새로운 데이터 세트 선택

선택한 데이터 세트 x에 대해 블랙박스 모델의 추론을 수행 후 예측

해석 가능한 모델 유형의 선택과 모델의 구성

데이터 세트 x에서 해석 가능한 모델을 블랙박스 모델의 예측에 맞도록 훈련한다.

*문제점
모델 복잡성의 단순화 과정에서 설명이 가능 기존의 모델과 정확도 차이가 있다.

Feature Importance

각 트리의 MSE는 피처의 변경 값과 원래 값으로 각각 한 번씩 계산된다

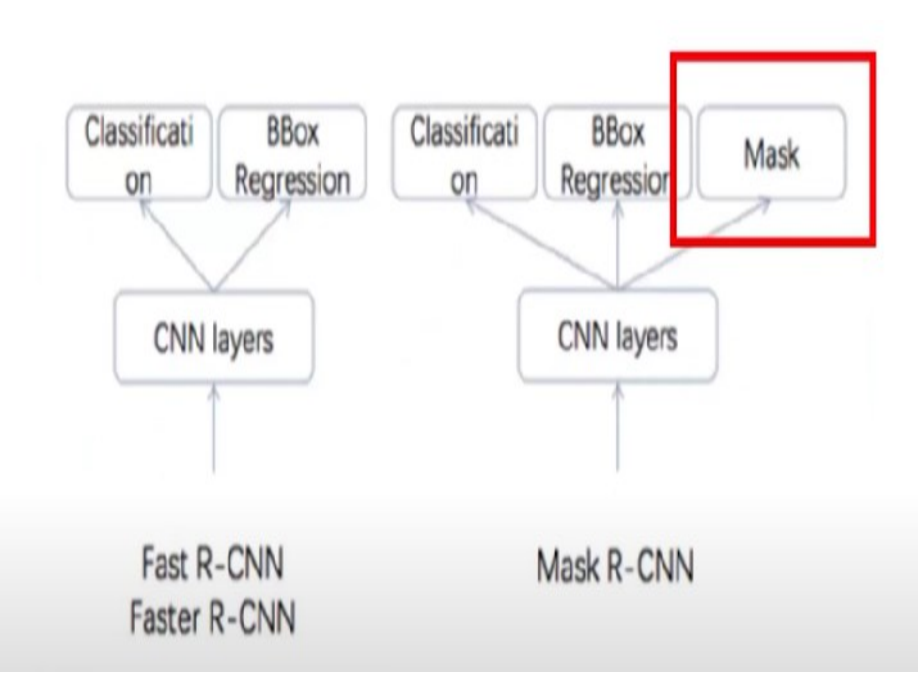
트리에서 이 두 값의 평균 차이가 feature importance다.

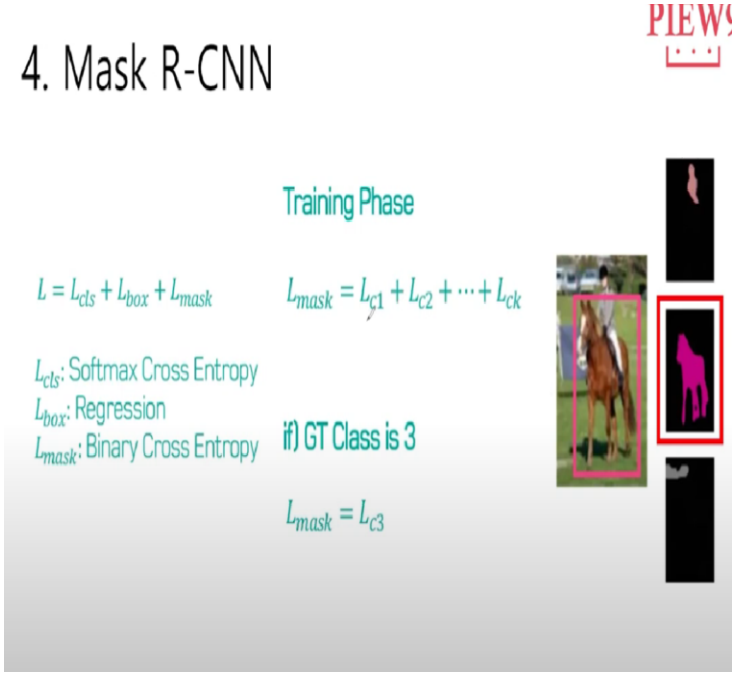
피쳐의 변화된 값으로 인해 성능이 저하될 경우가 중요한 피쳐다

$$PFI_S = \mathbb{E}(P(\hat{f}(\tilde{X}_S, X_C), Y)) - \mathbb{E}(P(\hat{f}(\tilde{X}), Y)),$$

XS: 값이 변경된 복제 모델
XC: xS(값이 변경된 복제모델에)에 없는 피쳐
P(f,Y): f에 대한 성능

Transparent Model Design





특정 부분을 입력하면 마스크필터는 단일 부분을 출력한다.
특정부분과 단일부분은 높은 관련성이 있다.

Global Model은 부분적으로 보지 않기 때문에 단일 관찰에 대한 설명을 생성할 때 신뢰도가 부족하다.

9. Local Approximation

주어진 인스턴스와 관련된 요소를 사용하여
훈련함으로써 복잡한 모델의 로컬 동작을
시뮬레이션한다.

LIME은 Local Approximation을 생성하기 위해 보편적 패턴 사용

$$\xi(x) = \arg \min_{g \in G} L(f, g, \pi_x) + \Omega(g), \qquad (6)$$

f 블랙 박스 모델
L local approximation과 블랙 박스 모델의 차이로 측정된 손실 함수

단점: Local영역의 행동은 복잡할 수 있고 객관적 값을 갖기 위해서는 방대한 매개 변수가 필요하기 때문에 해석할 수 없는 프록시 모델로 이어질수도 있다.

Forward Propagation

- 인풋을 하나씩 추가하면서 결과의 변화를 확인한다

Backward Propagation

- 기울기가 큰 피쳐들이 모델 예측의 변화에 더 많은 영향을 미친다.
- 블랙박스 함수의 폭이 큰 분산으로 인한 노이즈가 많고 무의미한 결과가 나온다.

Instance-Based Methods

Prototypes & Criticisms

Counterfactual

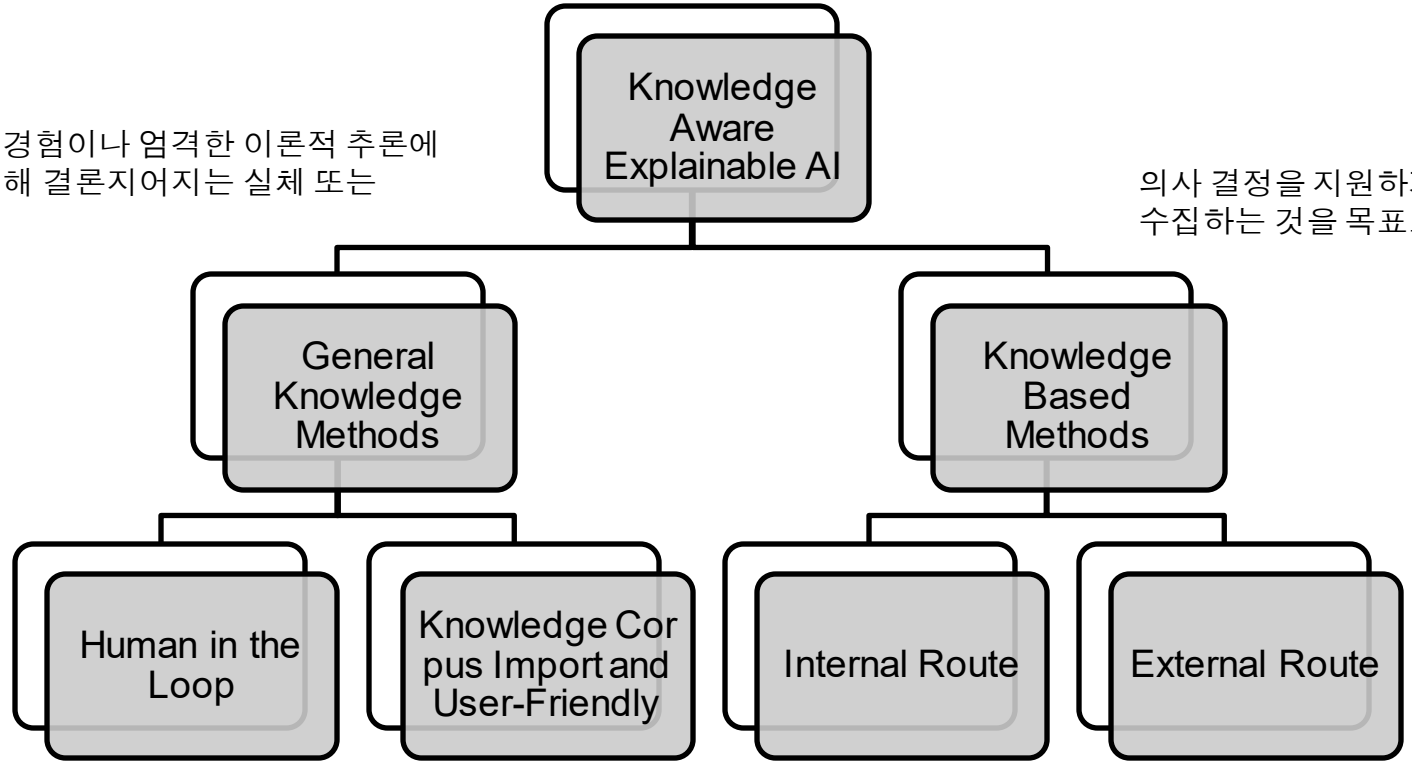
Instance-Based Methods

- 인스턴스 기반 설명 방법은 블랙박스 모델의 인사이트에 대한 데이터 세트의 특정 인스턴스를 선택하거나 생성한다.

- 이 방법은 이미지와 같은 데이터 항목이 인간적으로 이해될 때 효과적이다. 인스턴스 기반 설명은 훈련 데이터가 복잡한

13. KNOWLEDGE-AWARE EXPLAINABLE AI

인간들의 삶의 경험이나 엄격한 이론적 추론에 따라 인간에 의해 결론지어지는 실체 또는 관계들



의사 결정을 지원하기 위해 인간 전문가의 지식을 수집하는 것을 목표로 한다.

Human In The Loop

- 일반적으로 AI는 학습된 모델로 예측을 하고 중간에 사람이 개입되기가 어렵다. 원하는 결과가 나오지 않는다면 새로운 데이터를 써야 한다.
- 휴먼인더루프는 전문가가 중간중간 AI의 학습 데이터를 조정 할 수 있다.

Human In The Knowledge Corpus Import and User-Friendly

- 인간의 참여가 XAI 시스템에 외부 지식을 투영하고 지식 있는 설명을 생성할 수 있지만 절차를 표준화하고 자동화하기는 어렵다. 게다가, 특히 관련 개념들이 많이 있을 때, 사람은 관련 지식을 열거하기 어렵다

- 이러한 문제를 해결하기 위해 도메인 지식 또는 Corpus(언어자료)를 XAI 시스템으로 가지고 온다

Internal Route

- 지식 그래프의 관계, 그리고 규칙과 같은 기능을 사용하여 항목 간의 유사성 측정에 쓰이며 권장 시스템에 주로 적용된다.

External Route

- 추론에 지식을 더해 설명한다. 추론은 논리적 설명을 지식은 설명을 다듬는 역할을 한다.
- 블랙박스 모델의 입출력 동작을 설명하고 규칙을 설명한다.

Computational

- 전문가들에게 직관적 해석을 한다
- 합리적인 공식들에 의해서 해석되기 때문에 인간의 참여 없이도 해석 방법의 발전이 가능하다

Cognitive

- AI의 해석에 대한 인간의 이해를 측정한다.

Metrics

Computational

Post-Explanation Performance

- 조정 전과 조정 후의 모형 성능을 비교하여 설명을 평가하여 직관적인 설명

Faithfulness and Fidelity

- AI의 설명에서 중요한 특징이나 개념이 진정으로 관련 있는 정도를 feature importance의 특징들을 빼거나 더해 결과를 비교

Robustness.

- 견고성은 불안정한 데이터로 인한 노이즈와, adversarial attack(약간만 달라지면 AI가 못 알아보는 문제)에 를 극복하면서 해석할 수 있는 정도

Cognitive

Mental Model

- 사람이 시스템의 메커니즘을 알는 정도

Satisfaction

- 사용자가 AI시스템의 설명을 알아듣는 정도

Trust and Reliance

- 사용자의 믿음에 따라 AI의 설명을 받아들이는 정도

감사합니다

Predictive case-based feature importance and interaction

비즈니스인포메틱스학과
2022144203 변재현

연구배경

- 특징은 주요성은 XAI에서 중요한 부분이다

- 중요성과 상호작용의 실질적 의미를 이해하는 것은 어렵다

- 특징중요도와 상호작용의 실질적 의미를 이해하는 것이 필요하다.

2. Feature Interaction

Anova

- . 각 특징의 p-값을 계산하여 feature interaction측정. 긴 계산 시간 소요(Hastie and Tibshirani)

Pairwise

- 카이검정제곱을 통해 쌍방향 feature interaction.(Loh,Lou)

Grove Based Method

- 통계적 상호작용 탐지(Sorokina et al.)

Random Forest

- Features들이 Decision Tree와 같은 경로에 위치한다면 상호작용 하는 것으로 간주한다

부분 종속성 기반 기능 중요도 및 상호 작용 측정

- 부분 종속성의 분산으로 한 Feature Importance를측정
- 분산이 크면 상호작용하고 0이면 상호작용하지 않음(Green Well)

3. Feature Selection, importance

Feature Selection

Feature Importance

Permutation based Feature Importance

변수가 주어진 모형의 예측 결과에 영향을 미치는 정도를 측정합니다.
상호 작용끼리 합쳐지는 경향이 있다.

상호작용이 모형의 성능에
어느정도 영향을 미치는지 알 수 없다

Feature값을 랜덤하게 섞어서
중요성을 측정한다. 다른 특징들과의
상호작용을 고려하지 않는다
모델의 예측성능을 바꾼다

Feature Importance=Feature Power+Feature interaction

$$FP(F_U) = (n(G1) + n(G2) + n(G3) + n(G4))/NI$$

G1: F_x contribution (for correct prediction)
G2: $F_{x(-)}$ contribution (for correct prediction)
G3: Common contribution of F_x and $F_{x(-)}$ (for correct prediction)
G4: Cooperative contribution of F_x and $F_{x(-)}$ (for correct prediction)

(옳은 예측을 통한 특징의 기여도+공통 기여도+상호작용을 통한 기여도)/피쳐의 수

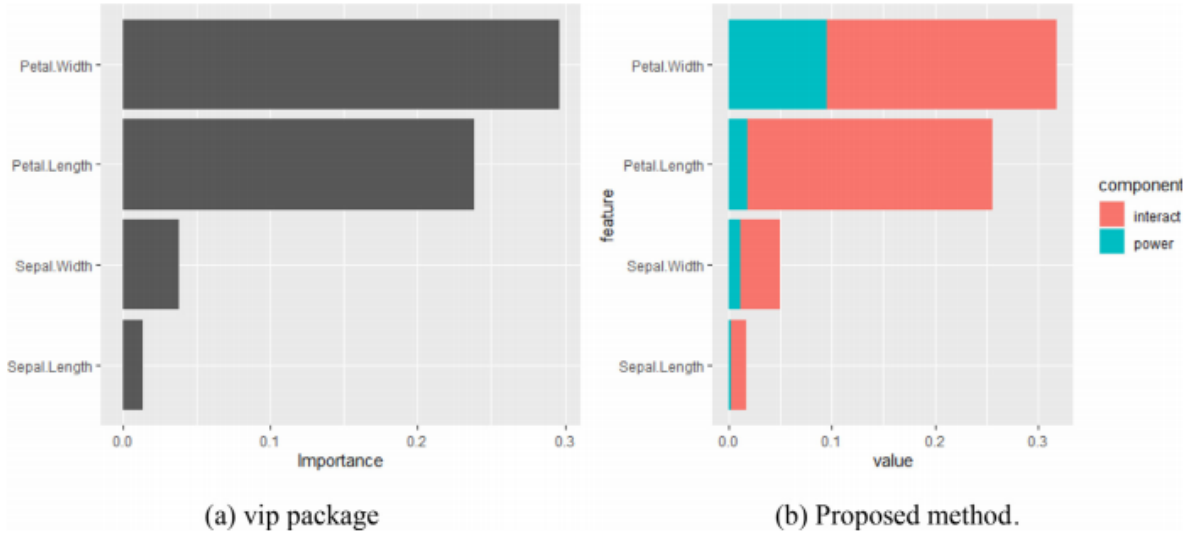
5. Comparing feature importance methods(Feature Imp Chart)

Table 2
Decomposition of predictive cases and feature importance.

Feature	PR_me[1]	PR_other[2]	CC[3]	Int[4]	Imp[5]
Sepal.Length	0.002	0.643	0.314	0.015	0.017
Sepal.Width	0.011	0.610	0.314	0.038	0.049
Petal.Length	0.018	0.337	0.381	0.238	0.255
Petal.Width	0.095	0.287	0.369	0.223	0.318

[1] Feature power of given feature.
[2] Feature power of other features.
[3] Common contribution of given feature and others.
[4] Feature interaction between given feature and others.
[5] Feature importance (PR_me + Interact).

제안된 방법에 따라 중요도를 Feature Power와 상호작용으로 나눔



(a) vip package (b) Proposed method.

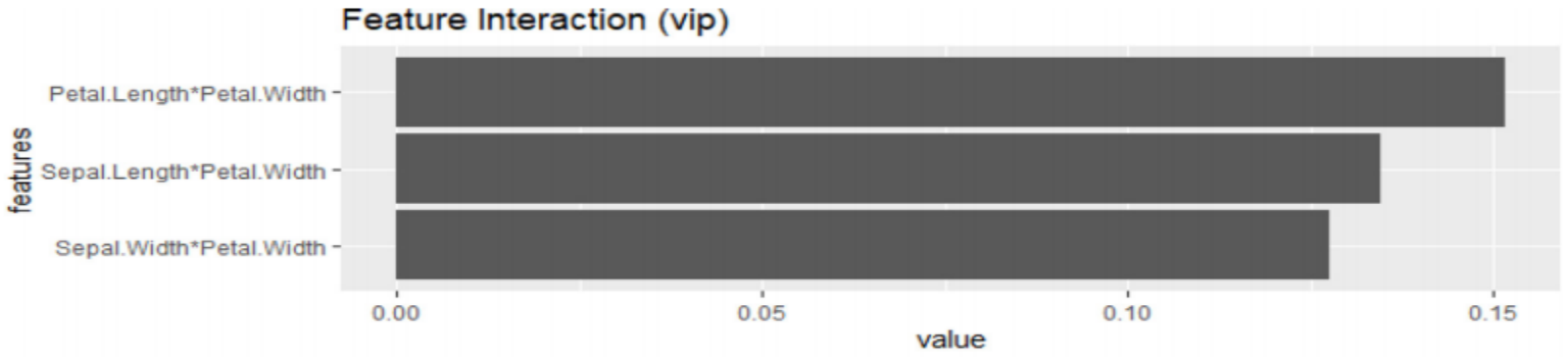
Fig. 3. Feature importance chart from vip package and proposed method.

vip Gini의 평균 감소에 따라 내림차순으로 가장 유의한 변수를 순서대로 제공
Gini가 낮다는 것은 데이터 불균형이 낮다는 것으로 해석됨

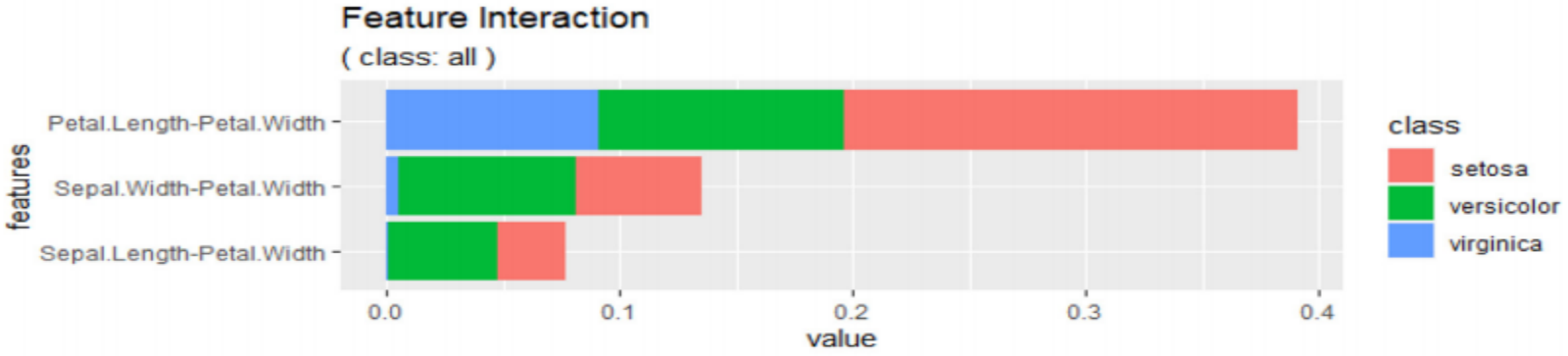
Iml: Feature importance, PDP, ICE, 셰플리 값등을 보여줌

제안된 방법론에서 feature Importance의 많은 부분이 변수들 간의 상호작용에 의해 많은 영향을 받는 것을 알 수 있다.

Feature interaction chart



(a) Result using *vip* package



(b) Results using the proposed method

제안된 방법은 클래스별로 구분된 특징 상호작용을 나타낸 반면, vip 패키지의 결과는 단일값의 상호작용만을 나타낸다.
클래스별 상호작용의 값을 알아보는데 유용하다

VIP, iml 제안된 성능의 차이

Pima 데이터셋의 feature importance를 Classification으로 구한거
Boston 데이터셋의 feature importance를 Regression으로 구한거

Table 3
Feature importance for pima dataset (classification).

Feature	Proposed	iml	vip
glucose	0.165	1.659	0.108
age	0.062	1.175	0.028
pedigree	0.039	1.135	0.021
pregnant	0.035	1.119	0.022
mass	0.034	1.127	0.022
insulin	0.033	1.095	0.020

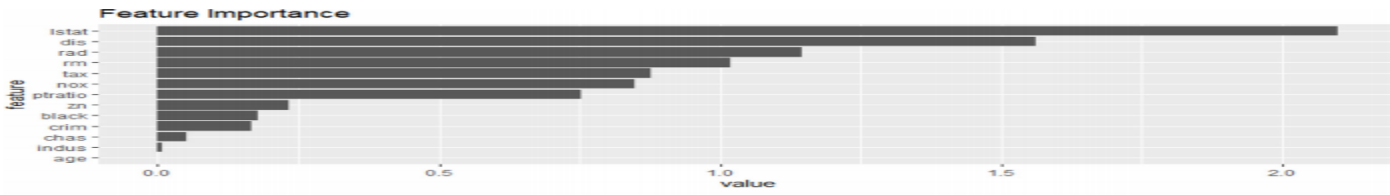
상관계수:**0.991, 0.992**

Table 4
Feature importance for Boston dataset (regression).

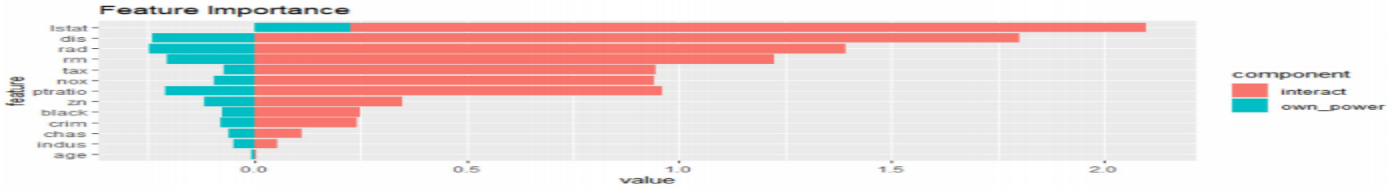
Feature	Proposed	iml	vip
lstat	2.098	1.659	2.12
dis	1.562	1.474	1.55
rad	1.145	1.335	1.13
rm	1.017	1.324	1.05
tax	0.876	1.269	0.902
nox	0.847	1.264	0.838
ptratio	0.752	1.236	0.789
zn	0.232	1.076	0.239
black	0.178	1.051	0.170
crim	0.166	1.052	0.017
chas	0.050	1.017	0.055
indus	0.006	1.001	0005
age	0.0	1.0	0.0

상관계수:**0.999, 0.999**

Prediction Analysis Chart



(a) Feature importance combining feature interaction and contribution.



(b) Feature importance splitting feature interaction and its own contribution

Fig. 9. Feature importance plot for Boston dataset.

보스턴 dataset의 Feature Importance를 특징의 기여도와 특징들간의 상호작용으로 나누었다.

Information Sciences 593 (2022) 15.

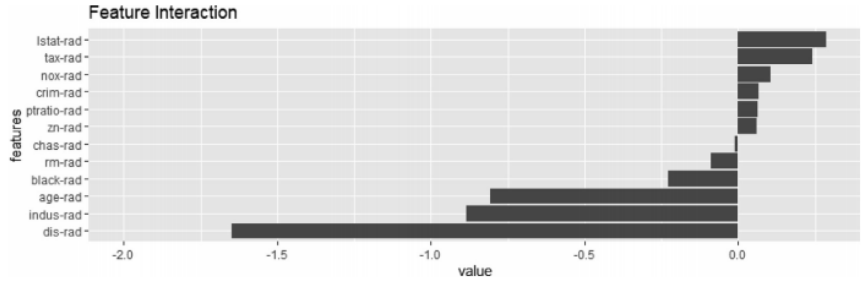
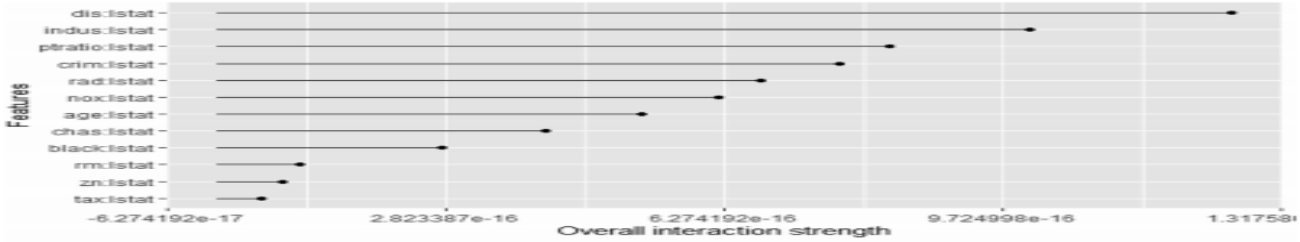


Fig. 10. Feature interaction plot for rad and other features in Boston dataset.

Rad라는 특징과 다른 특징들과의 상호작용 분석

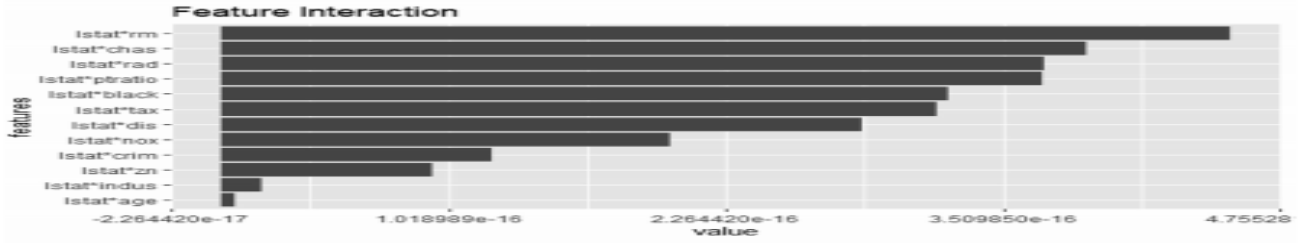
대부분의 상호작용이 음의 값을 가지고 있기 때문에 특징 상호작용이 특징의 중요성을 감소시킨 다는 것을 알 수 있다.

9. Interaction Type



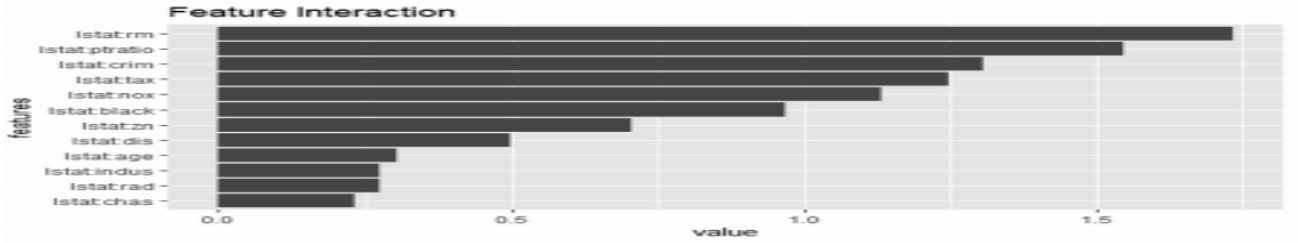
(a) Feature interaction graph using Friedman.

ids, indus, ptratio



(b) Feature interaction graph using Greenwell.

rm, chas, rad



(c) Feature interaction graph using proposed method

rm, pratio, crime}

Friedman, Greenwell 및 제안된 방법의 기능 lstat 과의 상호작용을 요약

방법들에 따라 상호작용에 미치는 값이 달라진다

1. Feature interaction Graph
2. Feature Contribution Chart
3. Prediction analysis Chart

Feature Interaction Graph

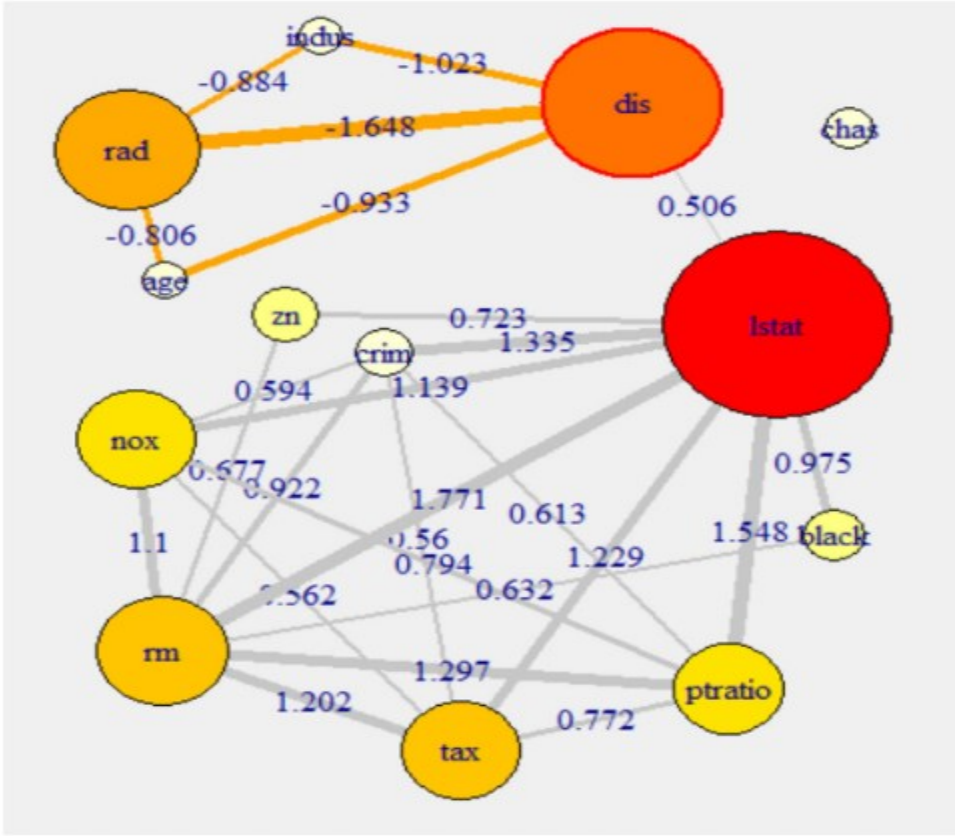


Fig. 12. Feature interaction graph for regression model created on the Boston dataset.

167

보스턴 데이터 의 리그레션 결과에 대한 상호작용 정도
'lstat'가 가장 중요한 특징이고 ' rad ' 와 ' dis ' 도 중요하나 서로 부정적으로 상호작용한다

상호작용으 라인을 생략한 임계값은 0.5다.
임계값:독립 변수 X가 어느 값이 되었을 때 종속변수 Y가 특이한 상태나 급격한 변화가 일어나는 경우

Feature Contribution Chart

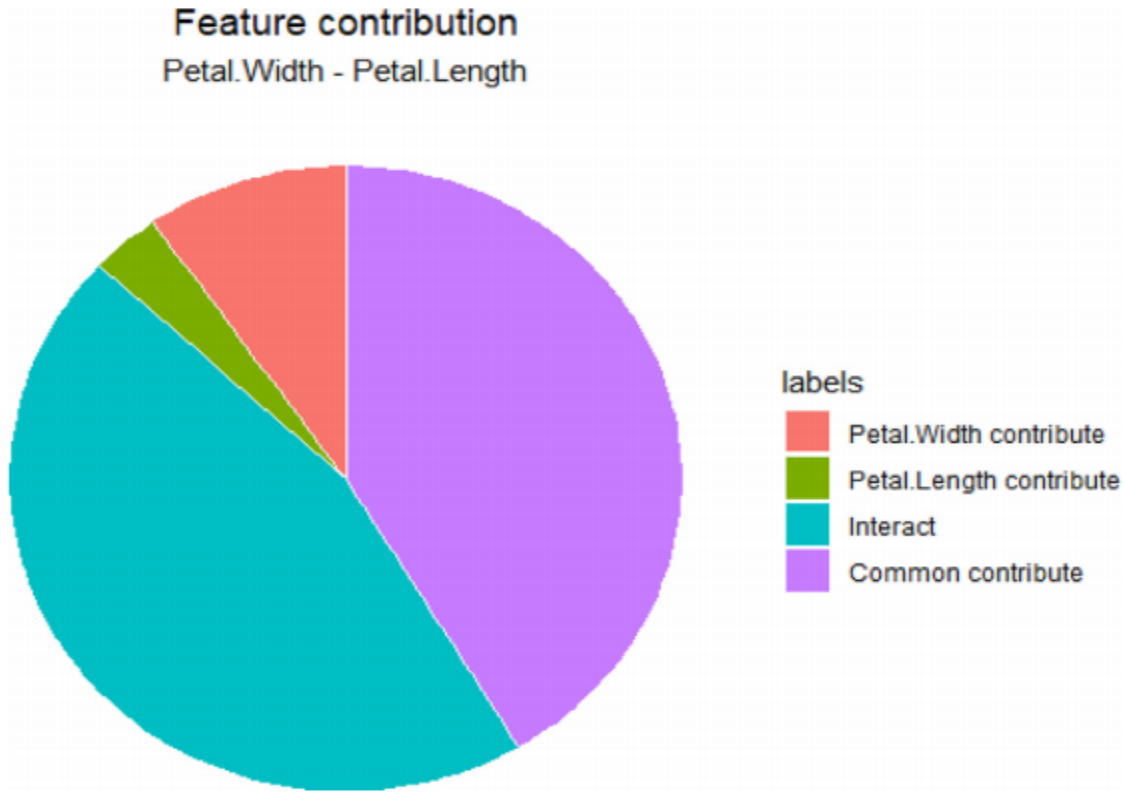


Fig. 13. An example of feature contribution chart.

Peta.Width가 Petal.Length보다 더 많이 기여하고 상호작용을 통한 기여도가 두 피쳐들의 합보다 크다.

Prediction Analysis Chart

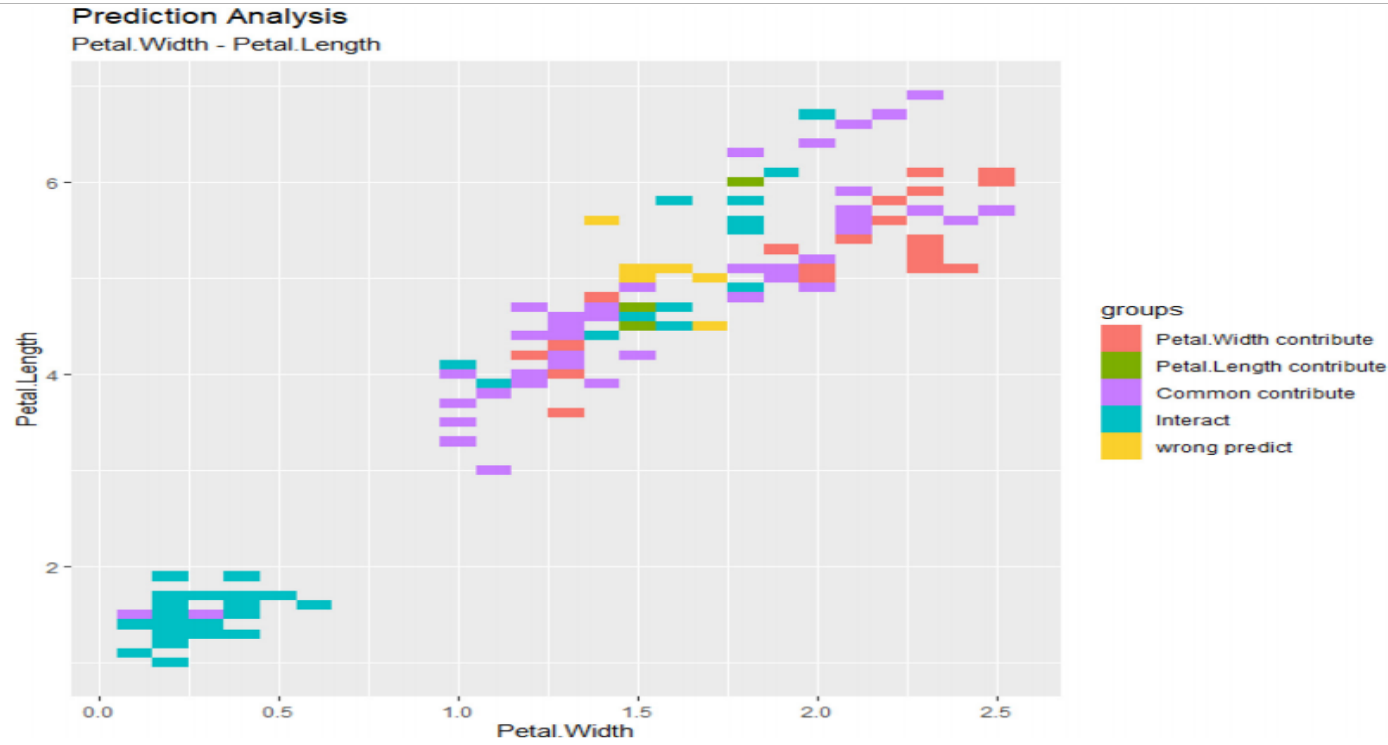


Fig. 14. Example of prediction analysis chart for classification model.

Classification 모델의 Prediction analysis chart

각 블록에서는 잘못 예측된 인스터스를 제외한 인스턴스를 나타낸다

차트의 주황색 블럭은 전의 모형에서는 정확하게 예측되지만 위 모형에서는 잘못 예측되었다.

차트를 통해 Petal.Width와 Petal Length의 기여도 및 상호작용을 알 수 있다.

제안된 방법론은 예측모델에서 특징의 기여도와 상호작용에 대한 더 세부적인 정보를 제공한다.

제안된 방법론은 직관적이고 이해하기 쉽다. 어려운 이론이나 수학 공식은 없다.

제안된 방법론은 모델에 구애받지 않는다. 데이터 세트만 있으면 모든 유형의 예측 모델에 적용될 수 있다.

대부분의 방법론들은 중요도와 상호작용 값의 상대적 으로만 의미를 찾을 수 있었으나 제안된 방법론은 중요도와 상호작용의 값을 정규화하지 않기 때문에 값 자체가 의미를 갖는다.

감사합니다