ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
https://www.jiisonline.org
J Intell Inform Syst 2023 September: 29(3): 229~247 http://dx.doi.org/10.13088/jiis.2023.29.3.229

이상탐지 알고리즘 성능 비교: 이상치 유형과 데이터 속성 관점에서

김재웅
국민대학교 비즈니스IT전문대학원
(ung315@kookmin.ac.kr)

정승렬
국민대학교 비즈니스IT전문대학원
(srjeong@kookmin.ac.kr)

김남규
국민대학교 비즈니스IT전문대학원
(ngkim@kookmin.ac.kr)

여러 분야에서 이상탐지의 중요성이 강조됨에 따라, 다양한 데이터 유형과 이상치 유형에 대한 이상탐지 알고리즘이 개발되고 있다. 하지만 이상탐지 알고리즘의 성능은 주로 공개 데이터 세트에 대해 측정될 뿐 특정 유형의 이상치에서 나타나는 각 알고리즘의 성능은 확인되지 않고 있으므로, 분석 상황에 맞는 적절한 이상탐지 알고리즘 선택에 어려움이 있다. 이에 본 논문에서는 이상치의 유형과 다양한 데이터 속성을 먼저 파악하여, 이를 기반으로 적절한 이상탐지 알고리즘 선택에 도움을 줄 수 있는 방안을 제시하고자 한다. 구체적으로 본 연구에서는 지역, 전역, 종속성, 그리고 군집화의 총 4가지 이상치 유형에 대해 이상탐지 알고리즘의 성능을 비교하고, 추가 분석을 통해 라벨 수준, 데이터 개수, 그리고 차원 수가 성능에 미치는 영향을 확인한다. 실험 결과 이상치 유형에 따라 가장 우수한 성능을 나타내는 알고리즘이 다르게 나타나며, 이상치 유형에 대한 정보가 없는 경우에도 안정적인 성능을 보여주는 알고리즘을 확인했다. 또한 비지도 학습 기반 이상탐지 알고리즘의 성능이 지도 학습 및 준지도 학습 알고리즘의 성능보다 낮게 나타나는 유형을 확인하였다. 마지막으로 데이터 개수가 상대적으로 적거나 많을 때 대부분 알고리즘들의 성능이 이상치 유형에 더 강하게 영향을 받으며, 상대적으로 고차원일 경우 지역, 전역 이상치에서는 우수한 성능을 보였지만 군집화 이상치 유형에서 낮은 성능을 나타냄을 확인하였다.

주제어 : 이상탐지, 이상치 유형, 성능 비교, 알고리즘 선택

논문접수일 : 2023년 8월 26일 논문수정일 : 2023년 9월 7일 게재확정일 : 2023년 9월 8일
원고유형 : Regular Track 교신저자 : 김남규

1. 서론

이상탐지(Anomaly detection)는 기계 학습 및 신경망을 기반으로 하여 정상과 이상치를 구분하는 기술이다. 이상치는 Anomaly 혹은 Outlier를 의미하는데, 전자는 구분하는 것 자체에 목적이 있을 때, 이와 반대로 후자는 구분 후 해당 데이터를 제거하는 것을 목적으로 할 때 주로 사용된다. 이때 이상치는 탐지하기 어려운 특정 패턴을 갖거나 기준이 명확하지 않고, 정상 데이터에 비해 수가 적다는 특징을 갖는다. 이러한 특징으로 인해 이상치를 정의하고 라벨링(Labeling)하는 과정이 더욱 중요하게 여겨지고 있으며, 이상치의 라벨링 여부에 따라 사용 가능한 이상탐지 알고리즘은 비지도, 지도, 그리고 준지도 학습 방식으로 다수 고안되어 왔다.

이상탐지는 현재 다양한 분야에서 활용되고 있으며, 대상이 되는 데이터의 종류도 매우 다양하다(Choi & Kim, 2019; Hwang, 2022; Lee & Kim, 2022). 그중 테이블 데이터는 일반적으로 쉽게 접할 수 있는 형식으로, 테이블 데이터의 이상치는 전역, 지역, 종속성, 그리고 군집화 이상치의

4가지 유형으로 구분된다. 이때 앞서 말한 이상탐지 알고리즘의 성능은 종종 다양한 유형의 이상치로 구성된 공개 데이터 세트를 통해 평가되는데, 이로 인해 특정 유형의 이상치에 대한 각 알고리즘의 성능은 확인이 어렵다는 한계가 있다(Steinbuss & Böhm, 2021). 물론 매 분석마다 모든 이상탐지 알고리즘을 각각 적용해 본 후 가장 성능이 우수하게 나타난 이상탐지 알고리즘을 선택하는 방법도 고려해 볼 수 있지만, 이러한 접근은 투입 시간과 자원을 고려할 때 현실적으로 가능한 대안이라고 보기 어렵다. 이상적으로는 각 이상치 유형에 대해 우수한 성능을 보이는 이상탐지 알고리즘이 알려져 있다면, 분석 대상 데이터의 이상치 유형을 사전에 확인한 후 그 결과에 따라 적합한 이상탐지 알고리즘을 선택하는 것이 바람직하다고 할 수 있다.

물론 이상탐지 알고리즘들의 성능 비교 및 분석을 진행한 기존의 이상탐지 벤치마크 결과를 참고할 수 있다. 하지만 기존의 분석은 라벨 정보가 없어도 진행 가능한 비지도 학습 기반 알고리즘만 비교하거나, 지역과 전역 이상치 유형만을 비교하는 등의 한계를 갖고 있다. 최근 이러한 한계를 극복하기 위해 4가지 이상치 유형과 3가지 지도 학습 기반 알고리즘의 성능을 다양한 관점에서 분석한 이상탐지 벤치마크가 수행되었지만, 데이터 개수와 차원수 등 데이터의 속성에 따른 심층적인 분석이 이루어지지 않았다는 아쉬움이 있다.

이에 본 논문에서는 4가지 유형의 이상치를 갖는 데이터 세트를 생성하고, 이에 대해 다양한 이상탐지 알고리즘을 적용하여 각 이상치 유형에 대한 알고리즘들의 성능을 비교한다. 이를 통해 얻은 결과를 분석하여 각 이상치 유형에서 좋은 성능을 나타내는 알고리즘을 탐색하고, 추가적으로 라벨 수준, 데이터 개수, 그리고 차원수 관점에서의 분석을 진행한다. 즉 4가지 이상치 유형에 대해 다양한 관점에 따른 알고리즘의 성능을 확인하고자 한다.

본 논문의 구성은 다음과 같다. 우선 다음 장인 2장에서는 본 연구와 관련된 기존의 연구 성과를 요약하고, 3장에서는 본 연구에서 다루고 있는 연구 문제와 모형을 소개한다. 실험의 과정 및 결과는 4장에서 소개하며, 마지막 장인 5장에서는 연구의 기여와 한계를 요약한다.

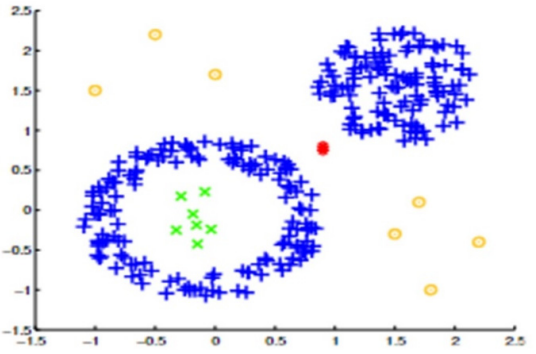
2. 관련 연구

2.1. 이상치의 정의와 유형

이상탐지는 대다수의 데이터 개체들과 크게 다른 값을 갖는 개체를 식별하는 것을 목표로 한다. 즉 이상치란 정상 데이터들의 관측 결과와 현저히 달라서, 이들이 다른 메커니즘에 의해 생성되었을 가능성을 의심케 하는 경우를 의미한다. 이상탐지는 사이버 보안, 금융, 그리고 제조 등 다양한 분야에서 사용되고 있으며, 이상탐지에 사용하는 데이터의 종류와 이상치 유형 역시 다양하다. 대표적으로 테이블 데이터, 시계열 특성 데이터, 그리고 그래프 특성 데이터 등에 대해 이상탐지가 이루어지고 있다. 특히 테이블 데이터는 주변에서 쉽게 찾아볼 수 있는 데이터 유형으로, 테이블 데이터의 이상치는 전역 이상치(Global anomalies), 지역 이상치(Local anomalies), 종속성 이상치(Dependency anomalies), 그리고 클러스터 이상치(Clustered anomalies)로 총 4가지 유형으로 구분된다.

우선 가장 일반적인 이상치 유형인 전역 이상치는 전체 데이터의 분포를 살폈을 때 이웃 밀도가 낮은 포인트를 의미한다. 한편 제한된 영역에서의 이웃 밀도가 낮은 데이터는 지역적으로 이웃 밀도가 낮은 데이터 포인트라 할 수 있으며, 이것을

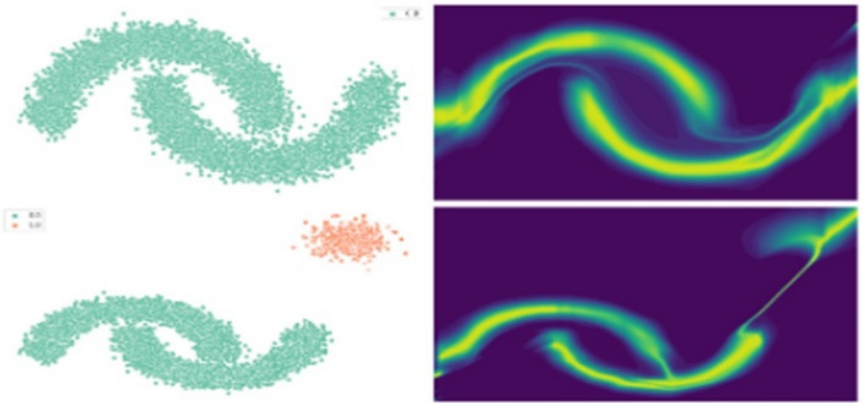
지역 이상치라고 한다. 이러한 측면에서 지역 이상치는 전역 이상치의 일반화된 개념으로 간주되기도 한다. 구체적으로 지역 이상치는 지역 이웃과 비교했을 때 상이성이 큰 개체를 일컬으며, 전역 이상치와 지역 이상치의 차이는 <그림 1>의 예를 통해 직관적으로 이해 가능하다.



〈그림 1〉 지역 이상치 포함 데이터 분포 예 (Huang et al., 2012).

<그림 1>에서 파란색 플러스 기호로 표시된 개체는 정상 데이터를 나타내고 있다. 이러한 정상 개체들은 일반적인 패턴을 따르고 있으며, 데이터 분포의 핵심 부분을 형성한다. 반대로 노란색 원과 빨간색 원, 그리고 초록색 X는 이상치를 나타낸다. 이들 이상치 중 빨간색과 초록색으로 표시된 이상치는 정상 개체로 형성되는 주변 데이터의 패턴을 따르지 않고 지역적으로 특이한 패턴을 나타낸다. 노란색 원 역시 정상 데이터와는 상이한 패턴을 보이지만, 특정 지역이 아닌 전역에 걸쳐 분포한다는 특징이 있다. 따라서 노란색 원은 전역 이상치로, 빨간색 원과 초록색 X는 지역 이상치라고 볼 수 있다.

다음으로 종속성 이상치는 일반적인 데이터가 갖는 의존성 구조를 따르지 않는 데이터 포인트를 의미한다(Martinez & Mata, 2016). 즉 종속성 이상치는 일반적인 데이터와는 독립적으로 동작하며, 특정 속성들 간의 관계가 일반적인 데이터와 다른 형태를 나타낸다는 특징이 있다. 마지막으로 군집화 이상치는 그룹 이상치라는 명칭으로도 널리 알려져 있다. 해당 이상치는 정상 개체들과 멀리 떨어져 있거나 희박하게 분포하며, 정상 데이터와 인접하면서도 밀집되어 나타난다는 특징을 갖는다(Liu et al., 2010). <그림 2>에서 좌측 상단은 정상 데이터만으로 구성된 데이터 세트의



〈그림 2〉 군집화 이상치의 분포 예 (Liu et al., 2022)

분포를, 좌측 하단은 정상 데이터와 이상치가 혼재한 데이터 세트의 분포를 시각화하여 나타내고 있다. 좌측 하단 그래프 중 붉은색으로 표시된 군집화 이상치들은 정상 개체들과 떨어진 지역에서 별도의 군집을 형성하고 있음을 알 수 있다.

2.2. 이상탐지 알고리즘

기존에 제안된 이상탐지 알고리즘의 대표적 예는 <표 1>과 같으며, 본 절에서는 각 알고리즘의 특징을 학습 방법에 따라 나누어 간략하게 살펴본다.

2.2.1. 비지도 학습 기반 이상탐지 알고리즘

이상탐지 알고리즘은 다른 분류 알고리즘에 비해 특정 패턴의 경우 식별이 어렵고, 정상 데이터가 이상 데이터에 비해 훨씬 많으며, 이상치에 대한 기준이 명확하지 않다는 어려움이 있다. 이상탐지 분야에서는 각 데이터에 대한 정상 혹은 이상의 정보를 라벨이라 하며, 라벨의 유무에 따라 지도 학습, 준지도 학습, 그리고 비지도 학습 기반의 이상탐지 알고리즘이 사용된다.

비지도 학습 기반 이상탐지(Unsupervised anomaly detection) 알고리즘은 데이터의 정상 혹은 이상에 대한 라벨 정보가 주어지지 않은 상태에서, 대부

〈표 1〉 이상탐지 알고리즘

학습 방법	알고리즘 명	참고 문헌
비지도 학습	LOF (Local Outlier Factor)	Breunig et al., 2000
	CBLOF (Clustering Based Local Outlier Factor)	He et al., 2003
	COF (Connectivity-Based Outlier Factor)	Tang et al., 2002
	COPOD (Copula Based Outlier Detector)	Li et al., 2020
	KNN (K-Nearest Neighbors)	Ramaswamy et al., 2000
	LODA (Lightweight on-line detector of anomalies)	Pevný, 2016
	PCA (Principal Component Analysis)	Shyu et al., 2003
	SOD (Subspace Outlier Detection)	Kriegel et al., 2009
	ECOD (Empirical-Cumulative-distribution-based Outlier Detection)	Li et al., 2022
	IForest (Isolation Forest)	Liu et al., 2008
지도 학습	SVM (Support Vector Machine)	Cortes & Vapnik, 1995
	RF (Random Forest)	Breiman, 2001
	XGBoost (eXtreme Gradient Boosting)	Chen & Guestrin, 2016
	LightGBM (Highly Efficient Gradient Boosting Decision Tree)	Ke et al., 2017
	CatBoost (Categorical Boosting)	Dorogush et al., 2018
준지도 학습	GANomaly (Semi-Supervised Anomaly Detection via Adversarial Training)	Akcay et al., 2019
	DeepSAD (Deep Semi-supervised Anomaly Detection)	Ruff et al., 2019
	REPEN (REPresentations for a random nEarest Neighbor distance-based method)	Pang et al., 2018
	FEAWAD (Feature Encoding With Autoencoders for Weakly Supervised Anomaly Detection)	Zong et al., 2018
	XGBOD (Extreme Gradient Boosting Outlier Detection)	Zhao & Hryniewicki, 2018

분의 데이터가 정상일 것으로 가정하고 학습을 진행한다. 이러한 접근을 따르는 알고리즘은 일반적으로 라벨 정보를 사용하는 지도 학습에 비해 성능이 높지 않다는 한계를 갖지만, 이상치가 정상 데이터 대비 발생 빈도가 훨씬 적은 데이터 불균형 상황에서도 사용 가능하다는 장점이 있다.

대표적인 비지도 방식 이상탐지 알고리즘으로는 데이터 포인트 간 이웃 공간 수를 활용하는 LOF(Breunig et al., 2000), 데이터가 속한 군집의 크기 및 해당 군집과 가장 가까운 군집 사이의 거리를 고려하는 CBLOF(He et al., 2003), 낮은 밀도와 고립성을 구별하기 위한 연결성 기반의 COF(Tang et al., 2002), 경험적인 COPula를 사용하는 고차원 데이터에서 이상적이고 계산 부하가 적은 것으로 알려진 COPOD(Li et al., 2020), 인접 이웃 기반 KNN(Ramaswamy et al., 2000), 약한 감지기들의 앙상블을 활용한 LODA(Pevný, 2016), 다차원 데이터의 차원 축소 기술을 이상치 식별에 사용한 PCA(Shyu et al., 2003), 적합한 부분 공간을 선정하여 이상 정도를 측정하는 SOD(Kriegel et al., 2009), 경험적 누적 분포 함수를 기반으로 하는 ECOD(Li et al., 2022), 그리고 무작위로 선택한 특성의 최대값과 최소값 사이에서 임의의 분할 값을 선택하여 관측치를 분리하는 IForest(Liu et al., 2008) 등이 있다.

2.2.2. 지도 학습 기반 이상탐지 알고리즘

지도 학습 기반 이상탐지(Supervised anomaly detection) 알고리즘은 정상 데이터와 이상치 데이터의 구분에 대한 라벨 정보를 학습에 사용하는 방법이다. 하지만 이상치는 라벨 정보가 알려진 경우가 드물며, 라벨링 작업 중 모든 유형의 이상치를 포착하는 것이 어렵다는 한계가 있다. 따라서 지도 학습 기반 알고리즘은 알려지지 않은 유형의 이상치 탐지에 적용이 어렵다는 한계가 있다(Aggarwal & Aggarwal, 2017).

대표적인 지도 학습 기반 이상탐지 알고리즘으로는 초평면을 이용한 비 모수적 분류 모형인 SVM(Cortes & Vapnik, 1995), 다양한 하위 샘플에 여러 의사 결정 트리 분류기를 적합 시키는 방식의 RF(Breiman, 2001), 앙상블 학습 방법 중 하나인 Gradient Boosting 기반의 XGBoost(Chen & Guestrin, 2016), 메모리 사용량이 적고 데이터 병렬 처리를 효율적으로 다루는 LightGBM(Ke et al., 2017), 그리고 범주형 변수 처리에 초점을 둔 CatBoost(Prokhorenkova et al., 2018) 등이 있다.

2.2.3. 준지도 학습 기반 이상탐지 알고리즘

준지도 학습 기반 이상탐지(Semi-supervised anomaly detection) 알고리즘은 부분적으로 라벨이 지정된 데이터를 사용하여 감지 성능을 개선하고, 라벨이 없는 데이터를 활용하여 표현 학습을 촉진한다. 즉 지도 학습을 부분적으로 활용하여 감지 성능을 향상시키면서, 비지도 학습을 부분적으로 활용하여 알려지지 않은 유형의 이상치도 탐지하는 것을 목표로 한다. 일반적으로 준지도 학습은 약한 지도(Weakly supervised learning)에서의 불완전한 라벨 학습을 의미하며(Zhou, 2018), 구체적인 예로 준지도 학습 기반 이상탐지 알고리즘인 GANomaly는 정상 데이터만을 학습한 후 훈련 과정에서 이미 학습한 정상 표현과 다른 특성을 갖는 이상치를 식별하는 방식으로 동작한다.

대표적인 준지도 학습 기반 이상탐지 알고리즘으로는 앞서 소개한 조건부 적대적 생성 신경망을 활용한 GANomaly(Akcay et al., 2019), DeepSVDD 모델을 개선한 end-to-end 방법론인 DeepSAD

(Ruff et al., 2019), 표현 학습과 이상치 탐지를 통합하여 가장 적합한 저 차원 표현을 학습하는 랭킹 모델 기반의 REPEN(Pang et al., 2018), 심층 오토인코더와 DAGMM의 네트워크 아키텍처를 사용한 FEAWAD(Zong et al., 2018), 지도 및 비지도 기계 학습 방법의 강점을 결합한 하이브리드 접근 방식인 XGBOD(Zhao & Hryniewicki, 2018) 등이 있다.

2.3. 이상탐지 알고리즘 벤치마크

이상탐지 알고리즘의 선택과 설계를 위해, 공정한 평가를 체계적으로 수행할 수 있도록 지원하는 이상탐지 벤치마크들이 다수 공개되었다. 각 벤치마크들은 사용하는 이상탐지 알고리즘 및 데이터 세트, 그리고 비교하는 관점이 서로 다르다. 우선 Emmott et al. (2015)은 여러 차원에서 주요 구성 요소에 따라 다양한 합성 이상탐지 데이터셋을 생성하고, 19개의 공개 데이터 세트를 8가지의 비지도 이상탐지 방법으로 평가하였다. 또한 Campos et al. (2016)은 23개의 데이터 세트에 대해 12가지 비지도 이상탐지 알고리즘을 비교하여, 벤치마크 데이터 세트의 특성과 이상탐지 벤치마크 세트로서의 적합성을 제시했다. Domingues et al. (2018)은 15개의 공개 데이터 세트에 대해 14가지의 비지도 이상탐지 방법을 테스트하고 확장성, 견고성, 그리고 메모리 사용량을 비교하였다.

한편 Soenen et al. (2021)은 16개의 데이터 세트에 대해 6가지의 알고리즘을 적용하여 하이퍼파라미터(Hyperparameter) 설정을 위한 작은 검증 세트를 사용할 것을 제안하였으며, Steinbuss & Böhm (2021)은 현실적인 합성 데이터 생성을 위한 일반적인 과정을 제안하였다. Ruff et al. (2021)은 딥러닝 학습 기반의 이상탐지 방법과 고전적인 방법을 통합적으로 검토하고, 3개의 데이터 세트에 대해 9가지의 알고리즘의 실증적 평가를 수행하였다. 마지막으로 Han et al. (2022)은 정상 데이터를 기준으로 4개의 이상치 데이터 유형을 갖는 데이터 세트를 생성하고, 57개의 데이터 세트에 대해 14가지의 비지도 학습 기반 알고리즘, 7개의 준지도 학습 기반 알고리즘, 그리고 9개의 지도 학습 기반 알고리즘인 총 30가지의 이상탐지 알고리즘을 비교하였다. 또한 해당 연구는 3개의 데이터 오염 설정에서의 성능과 견고성, 그리고 안정성을 평가하고 이를 오픈소스로 제공한 바 있다.

3. 연구 문제 및 모형

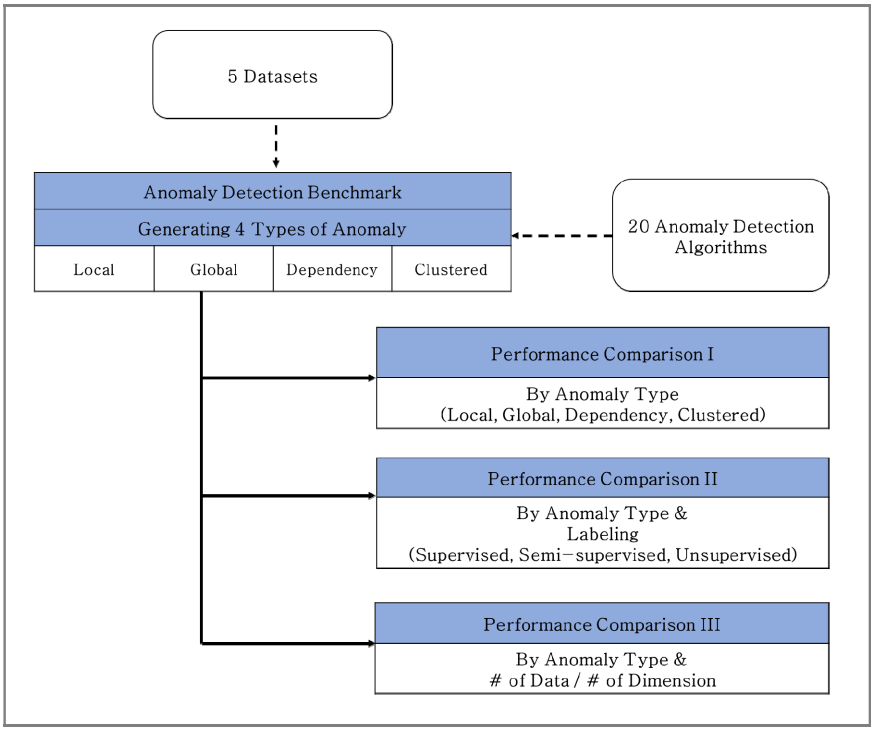
일반적으로 이상탐지 알고리즘은 공개 데이터 세트에 대한 실험을 통해 성능을 평가한다. 하지만 공개 데이터 세트는 일반적으로 다양한 유형의 이상치를 동시에 포함하고 있기 때문에, 이러한 실험을 통해서는 특정 유형의 이상치에 대한 이상탐지 알고리즘의 상대적 성능을 비교하기 이해하기 어렵다는 한계가 있다(Steinbuss & Böhm, 2021). 또한 기존에 수행된 다양한 이상탐지 벤치마크는 대부분 4가지 이상치 유형을 종합적으로 비교하기보다는 지역 이상치와 전역 이상치를 비교하는 데 초점을 맞추었다. 또한 기존의 벤치마크는 데이터 비지도 학습 기반 알고리즘을 주로 사용하였는데, 이는 비지도 알고리즘의 경우 데이터의 라벨링에 소요되는 추가 비용이 없다는 장점을 갖기 때문이다. 하지만 현실에서는 도메인 전문가가 식별한 몇 가지 이상 징후 또는 능동 학습과 같은 human-in-the-loop 기술을 사용해서 라벨이 지정된 데이터 세트를 이상탐지에 활용하는 경우도 존재하므로, 지도 학습과 준지도 학습을 아울러

성능을 비교하는 연구가 수행될 필요가 있다.

Han et al. (2022)은 이러한 요구를 충족시키기 위해, 공개 데이터 세트의 정상 데이터를 사용하여 생성 모델을 구축하고 4가지 이상치 유형의 현실적인 이상치를 생성하는 방법을 사용하여 실험을 수행했다. 또한 지도 학습 방법의 라벨 수준에 따른 성능을 비교하였으며, 이를 통해 어떤 비지도 학습 방법도 통계적으로 다른 방법보다 우수하지 않다는 점을 재차 확인했다. 또한 알고리즘들의 성능이 이상치 유형별로 다르다는 것과, 이상치 유형에 대한 사전 지식이 부분적인 라벨 데이터 사용보다 더 큰 영향을 줄 수 있다는 점도 확인했다. 하지만 해당 연구는 포괄적인 비교를 진행하는 과정에서 데이터 개수와 차원 수가 상이한 데이터 세트를 사용하였기 때문에, 이상치 유형 외에 성능에 영향을 미치는 요인을 더욱 엄밀하게 분석하지는 못했다. 한편 Domingues et al. (2018)은 다양한 데이터 개수와 차원에서의 알고리즘 성능 비교를 수행했지만, 여러 이상치 유형과 다양한 지도 학습 알고리즘의 비교는 다루지 않았다는 한계를 갖는다.

이에 본 논문은 4가지 이상치 유형에 대해 라벨 수준, 데이터 개수, 그리고 데이터 차원 수 측면에서 다양한 이상탐지 알고리즘들의 성능을 확인하는 것을 연구 목표로 설정하였다. 구체적으로 1) 이상치 유형별 알고리즘들의 성능 차이, 2) 이상치 유형별 라벨 수준에 따른 알고리즘들의 성능 차이, 3) 이상치 유형별 데이터 개수와 차원 수에 따른 알고리즘들의 성능 차이를 분석하고자 한다.

본 연구의 모형은 <그림 3>으로 요약된다. 우선 이상치 유형별 알고리즘들의 성능 비교를 위



〈그림 3〉 전체 분석 모형

해, 4가지 이상치 유형을 갖고 있는 데이터 세트를 생성한다. 기본적으로 이렇게 생성된 4가지 데이터 세트에 20가지 이상탐지 알고리즘을 적용하여 성능을 분석한다. 첫 분석은 이상치 유형에 따른 알고리즘들의 성능 비교이다. 두 번째 분석은 4가지 이상치 유형의 데이터 세트들에 대해 라벨 수준(1%, 5%, 10%, 25%)와 지도, 준지도, 비지도 학습 기반 접근법의 관계를 분석한다. 마지막으로 4가지 이상치 유형에서 데이터 개수와 차원 수가 성능에 미치는 영향을 분석한다.

4. 실험

4.1. 실험 설계

4.1.1. 실험 데이터

본 연구에서는 실험을 위해 의료 분야 데이터 세트인 Cardio와 WBC(Ayres-de-Campos et al., 2000; Mangasarian et al., 1995), 이미지 처리 데이터 세트인 InternetAds(Campos et al., 2016), 문서 처리 데이터 세트인 PageBlocks(Malerba et al., 1996), 그리고 기부에 관한 사회학 분야 데이터 세트인 donors(Pang et al., 2019)의 총 5개의 데이터 세트를 사용한다. 이들은 이상탐지 벤치마크에서 자주 사용되는 데이터 세트들로, 모두 테이블 형식으로 구성되어 있다. 이들 데이터 세트들의 데이터 개수와 차원 수, 이상치 수, 그리고 이상치 비율은 <표 2>와 같다.

4.1.2. 이상치 유형별 데이터 생성

실험을 위한 이상치 생성 방법은 다음과 같다. 실험 데이터 세트의 정상 샘플을 사용하여 가우시안 혼합 모델 (GMM)과 같은 생성 모델을 구축하고, 기존의 이상치는 분석에서 제외한다. 이후 생성 모델을 조정하고 기존의 데이터 분포를 가정하여 합성 이상치를 생성한다. 이러한 방식으로 특정 유형의 이상치를 생성할 수 있으며, 이를 통해 정상 샘플과 4가지 유형의 이상치를 갖는 데이터 세트를 구축한다. 이때 각 이상치 유형별 구체적 생성 방법은 아래와 같이 정의된다 (Han et al., 2022).

지역 이상치는 GMM을 사용해 합성 정상 샘플을 생성하고, 스케일링 매개변수 α=5 로 공분산 행렬 $\hat{\Sigma} = \alpha\hat{\Sigma}$을 조정한다. 다음으로 전역 이상치는 입력 특징 k의 최솟값과 최댓값으로 정의한 경계를 기준으로 스케일링 된 균등 분포 $\text{Unif}\left(\alpha \cdot min(\mathbf{X^k}), \alpha \cdot max(\mathbf{X^k})\right)$에서 이상치를 생성한다. 종속성 이상치는 Vine Copula(Aas et al.,

〈표 2〉 실험 데이터 세트 요약

Data	데이터 개수	차원 수	이상치 개수	이상치 비율(%)	분야
cardio	1,831	21	176	9.61	Healthcare
PageBlocks	5,393	10	510	9.46	Document
internetAds	1,966	1,555	368	18.72	Image
donors	619,326	10	36,710	5.93	Sociology
WBC	223	9	10	4.48	Healthcare

2009)을 사용하여 데이터의 종속성을 모델링한다. 모델링 된 종속성을 제거하여 독립성으로 만들고, 커널 밀도 추정(Kernel Density Estimation; KDE) (Hastie et al., 2009)을 사용하여 특성의 확률 밀도 함수를 추정하고 이를 통해 정상 샘플을 생성한다. 군집화 이상치는 정상 데이터의 평균 피처 벡터를 α=5로 스케일링 하고, 이후 스케일링 된 GMM을 활용하여 생성한다. 하지만 종속성 이상치의 경우 차원이 많은 데이터에 적합시키는데 계산 비용이 많이 발생한다는 어려움이 있다(Steinbuss & Böhm, 2021). 따라서 본 실험에서는 종속성 이상치는 하나의 데이터 세트인 cardio에 대해서만 실험을 수행한다.

4.1.3. 실험 환경과 하이퍼파라미터

1,000개 미만의 작은 데이터 세트의 경우 샘플 크기를 1,000개로 리샘플(resample) 하고, 10,000개를 초과하는 큰 데이터 세트의 경우 부분 집합을 사용하여 계산 비용을 줄인다. 이 과정에서 계층화 추출법 (Stratified sampling)을 적용하여 이상치의 비율을 일정하게 유지한다. 데이터는 70%의 데이터를 train 데이터로, 나머지 30%는 test 데이터로 분리하여 사용한다. 비지도 학습 기반 이상탐지 알고리즘의 경우 주로 새로운 데이터에 대한 예측을 제외하고 입력 데이터에 대해서만 이상 점수를 출력하는 방식으로 설계되었다. 이에 실험을 위해 모든 알고리즘의 예측을 수행하는 추론 설정으로 적용하여 새로운 데이터의 예측을 가능하게 하였다. 이러한 방법은 이상탐지 라이브러리인 PyOD(Zhao et al., 2019), TODS(Lai et al., 2021), 그리고 PyGOD(Liu et al., 2022) 등에서도 일반적으로 사용된다. 각 알고리즘의 하이퍼파라미터는 공정한 비교를 위해 해당 알고리즘의 기본 하이퍼파라미터 설정을 그대로 사용한다.

4.1.4. 평가 지표

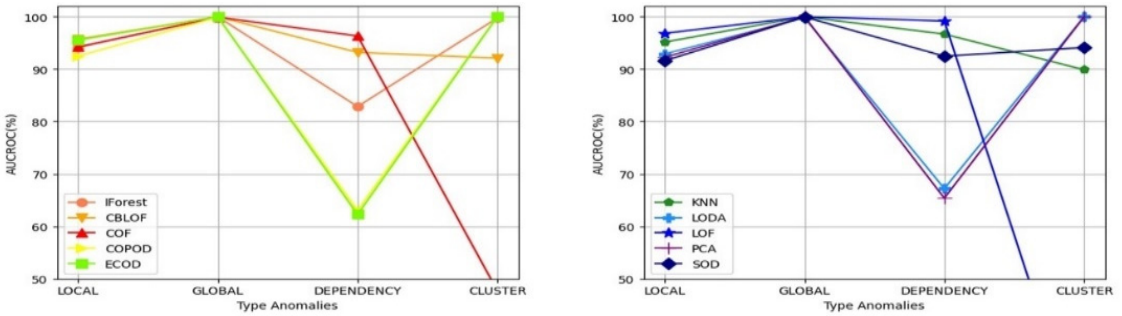
평가를 위해 정밀도(Precision), 재현율(Recall), F1-Score, AUCROC(Area Under Receiver Operating Characteristic Curve), 그리고 AUCPR(Area Under Precision-Recall Curve) 등 다양한 평가 지표가 사용되고 있다. 특히 AUCPR은 정밀도-재현율 곡선의 밑면적을, 그리고 AUCROC는 True Positive Rate과 False Positive Rate의 관계를 나타낸 그래프인 ROC Curve의 밑면적을 계산한 값으로 이상탐지 모델의 종합적인 성능 평가를 위해 사용된다(Boyd et al., 2013). 이들 여러 지표들 중 본 실험에서는 AUCROC를 핵심 평가 지표로 활용한다.

4.2. 실험 결과

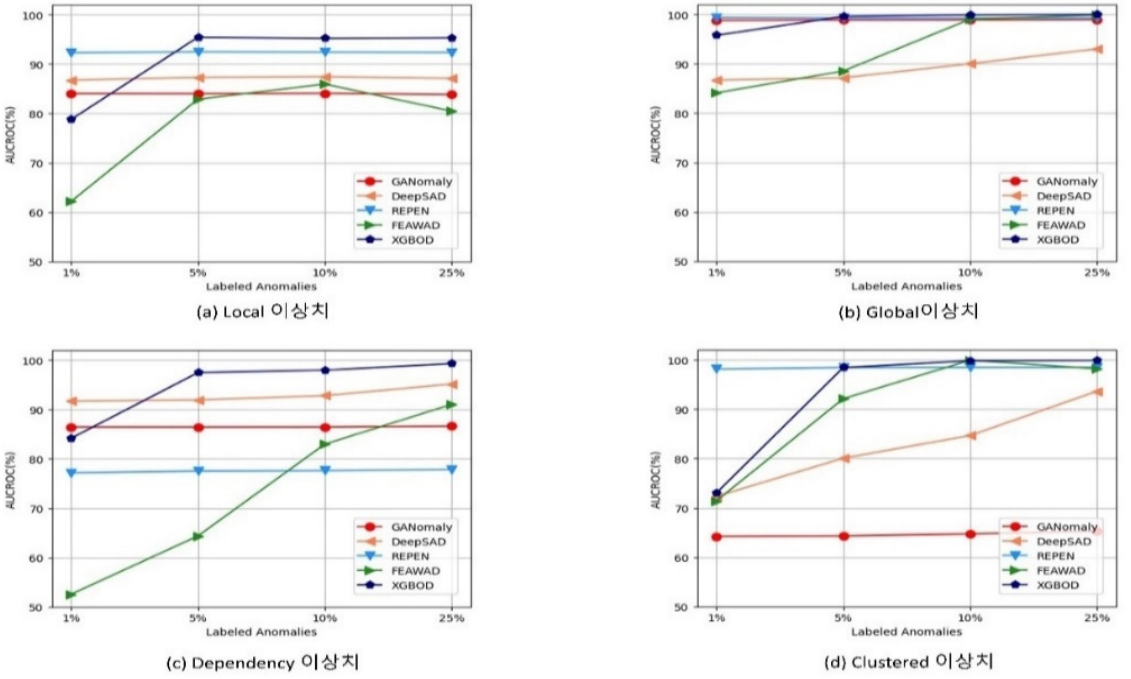
4.2.1. 이상치 유형별 알고리즘 성능 비교

첫번째 실험은 cardio 데이터 세트로부터 생성한 이상치 합성 데이터 세트에서, 알고리즘들의 AUCROC 값을 지도 학습 방법별로 나누어서 살펴보는 것이다. 실험 결과 비지도 학습의 경우 지역 이상치와 전역 이상치에서는 대부분 AUCROC가 90% 이상으로 높게 나타났으며, 종속성 이상치 유형에서는 LOF, KNN, COF, CBLOF가, 군집화 이상치에서는 IForest, ECOD, CBLOF가 높은 성능을 보였다(그림 4).

준지도 학습의 경우 지역 이상치와 전역 이상치에서는 라벨 수준 1%에서 REPNE, 라벨수준 5% 이상에서 XGBOD가, 종속성 이상치에서는 라벨 수준 1%에서 DeepSAD, 라벨 수준 5% 이상

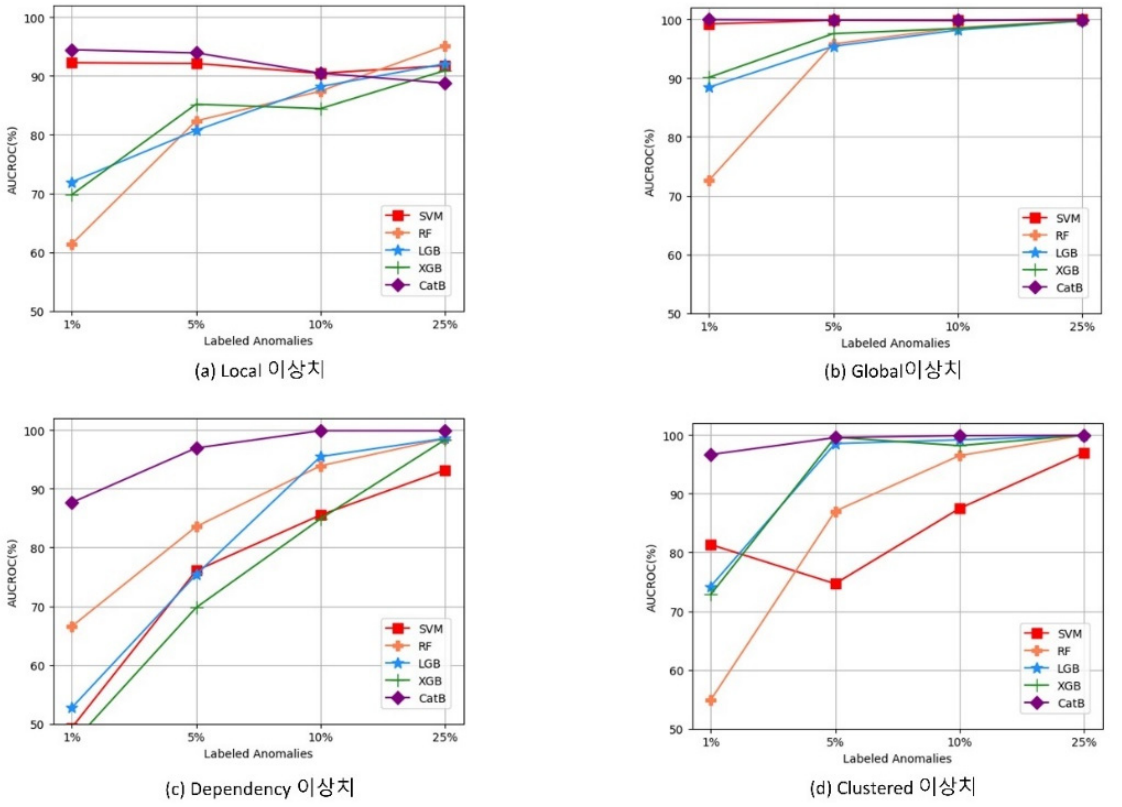


〈그림 4〉 이상치 유형별 비지도 학습 알고리즘의 AUCROC(%)



〈그림 5〉 이상치 유형별 준지도 학습 알고리즘의 AUCROC (%)

에서 XGBOD가, 군집화 이상치에서는 REPEN가 우수한 성능을 나타내는 것을 확인하였다(그림 5). 또한 지도 학습의 경우 모든 이상치 유형에서 CatB가 상대적으로 우수한 성능을 보였으며(그림 6), 라벨 수준에 상관없이 5가지 알고리즘 중 가장 좋은 성능을 나타냈다. 또한 모든 이상치 유형에서 각 지도 학습별로 CBLOF, XGBOD, CatB가 안정적인 성능을 나타냈다. 해당 결과를 정리한 <표 3>에 따르면, 이상치 유형에 대한 사전 정보가 없는 경우 CBLOF, XGBOD, 그리고 CatB을 사용하는 것을 우선적으로 고려할 수 있다.



〈그림 6〉 이상치 유형별 지도 학습 알고리즘의 AUCROC (%)

〈표 3〉 이상치 유형별 알고리즘 성능 요약

학습 방법	이상치 유형별 성능 우수 알고리즘				전반적 성능 우수 알고리즘
	지역	전역	종속성	군집화	
비지도	LOF	모두	LOF	IForest	CBLOF
준지도	REPEN, XGBOD	REPEN, XGBOD	DeepSAD, XGBOD	REPEN	XGBOD
지도	CatB	CatB	CatB	CatB	CatB

4.2.2. 라벨 수준별 알고리즘 성능 비교

다음 실험은 모든 데이터 세트에서 이상치 유형별 알고리즘들의 평균 AUCROC 값을 산출하여 비교하는 것이다. 지역 이상치의 경우 LOF가, 전역 이상치에서는 SOD, KNN가 좋은 성능을 보였으며, 비지도학습 방법의 성능이 상대적으로 우수하게 나타남을 확인하였다. 한편 종속성 이상치에서는 라벨 비율 10%에서 CatB, LOF, 그리고 XGBOD의 순으로 우수한 성능을 보였으며,

군집화 이상치에서는 라벨 5%, 10%에서 XGBOD, XGB, 그리고 CatB의 순으로 우수한 성능을 보였다.

또한 지역, 전역, 이상치에서는 비지도 학습 방법보다 우수한 지도 학습 방법이 없는 것으로 나타났지만, 종속성 이상치의 일부 라벨 수준과 군집화 이상치에서는 비지도 학습 방법보다 우수한 성능을 보이는 지도 학습 방법이 있음을 <표 4>와 같이 확인하였다. 즉 비지도 학습 방법이 항상 가장 좋은 대안인 것은 아니며, 모든 유형의 이상치에서 항상 가장 우수한 성능을 보이는 알고리즘은 없음을 재차 확인하였다.

4.2.3. 데이터 개수와 차원 수에 따른 성능 분석

다음 실험은 데이터의 개수에 따른 성능을 분석하는 것으로, 차원 수가 유사하고 데이터 개수가 서로 다른 donors(600,000개)와 WBC(223개) 데이터 세트의 이상치 유형별 AUCROC를 산출하여 분석하였다. 실험 결과 지역 이상치 유형에서 데이터 개수가 상대적으로 많은 경우에 SVM, IForest, ECOD, KNN, LODA, 그리고 PCA 알고리즘이 더 우수한 성능을 나타냄을 보였으며, 전역 이상치 유형에서 데이터 개수가 상대적으로 많은 경우에 GANomaly, SVM 알고리즘을 제외한 모든 알고리즘의 성능이 더 우수한 것으로 나타났다. 한편 군집화 이상치의 경우 데이터 개수가 상대적으로 많은 경우 GANomaly, DeepSAD, XGBOD, IForest, CBLOF 알고리즘이 더 우수한 성능을 나타냈다.

마지막 실험은 차원 수에 따른 성능을 분석하는 것으로, 데이터 개수가 비슷하며 차원 수가 서로 다른 데이터 세트인 PageBlocks(10개)와 internetAds(1,555개)에 대해 이상치 유형별 알고리즘의 AUCROC를 비교한다. 실험 결과 지역 이상치에서 차원 수가 상대적으로 많은 경우 FEAWAD를 제외한 모든 알고리즘이 AUCROC 90%를 넘으며 성능이 더 우수함을 확인하였다. 또한 전역 이상치 유형에서도 차원 수가 상대적으로 많은 경우 라벨 수준 5% 이하에서 FEAWAD를 제외한 모든 알고리즘이 AUCROC 90%를 넘으며, DeepSAD, REPEN, 그리고 FEAWAD를 제외한 알고리즘의 성능이 더 우수하게 나타남을 확인하였다. 반면 군집화 이상치에서 차원 수가 많은 경우, SVM을 제외한 다른 알고리즘들의 성능은 더 우수하지 않은 것으로 나타났다.

이상의 결과를 <표 5>와 같이 종합했을 때, 데이터의 크기의 영향이 알고리즘의 성능에 미치는 직접적인 영향은 명확히 확인되지 않았다. 하지만 데이터 차원 수의 경우, 지역, 전역 이상치에서 대부분의 알고리즘이 적은 차원에 비해 많은 차원에서 우수한 성능을 보이는 명확한 현상

〈표 4〉 라벨 수준별 알고리즘 성능 요약

라벨 수준	이상치 유형별 우수한 성능을 보이는 알고리즘			
	지역	전역	종속성	군집화
1%	LOF	SOD	LOF	CatB
5%	LOF	SOD	CatB	XGBOD
10%	LOF	SOD	CatB	XGBOD
25%	LOF	SOD	CatB	XGBOD

〈표 5〉 데이터 개수에 따른 알고리즘 성능 요약

이상치 유형	데이터 개수에 따른 성능 우수 알고리즘		데이터 개수 증가에 따라 성능이 향상되는 알고리즘 비율(%)
	데이터 少	데이터 多	
지역	LOF, CatB	LOF, COF	30%
전역	KNN, CBLOF	KNN, CBLOF	90%
군집화	FEAWAD, CatB, PCA	REPEN, CBLOF	25%

〈표 6〉 차원 수에 따른 알고리즘 성능 요약

이상치 유형	차원 수에 따른 성능 우수 알고리즘		차원 수 증가에 따라 성능이 향상되는 알고리즘 비율(%)
	차원 수 小	차원 수 大	
지역	LOF	LOF, GANomaly, XGBOD, CatB 등	95%
전역	KNN	KNN, GANomaly, SVM, CatB 등	85%
군집화	FEAWAD	SVM	5%

을 확인하였으며, 군집화 이상치에서는 SVM을 제외한 모든 알고리즘들이 많은 차원에 비해 적은 차원에서 오히려 우수한 성능을 보임을 확인하였다<표 6>.

5. 결론

본 연구는 테이블 데이터가 갖는 4가지 이상치 유형에 대해, 다양한 이상탐지 알고리즘의 성능을 라벨 수준, 데이터 개수, 그리고 차원 수 관점에서 분석하였다. 결과적으로 모든 이상치 유형에서 가장 좋은 성능을 보이는 알고리즘은 없음을 재차 확인하였다. 또한 비지도 학습 방법은 지역, 전역, 종속성 이상치에 대해 높은 성능을 보이지만, 군집화 이상치에 대해서는 지도 학습 방법에 비해 낮은 성능을 보임을 확인하였다. 데이터의 개수와 차원 수의 경우, 지역, 전역, 군집화 이상치에서는 성능에 큰 영향을 주지 않는 것으로 나타났다. 한편 차원 수의 경우 지역, 전역 이상치에서는 상대적으로 많은 차원이 사용된 경우 성능이 좋았지만, 군집화 이상치에서는 SVM을 제외한 모든 알고리즘이 그 반대의 결과를 보였다. 이러한 결과를 종합하면 지역, 종속 이상치의 경우 LOF 알고리즘이 모든 경우에서 가장 뛰어난 성능을 보이는 것으로 평가할 수 있다. 전역 이상치에서는 알고리즘들의 성능 차이가 크지 않아, 학습 시간이나 예측 시간 등의 추가적인 기준으로 알고리즘을 선택할 수 있다. 군집화 이상치에서는 CatB, XGBOD, 그리고 XGB 등의 알고리즘을 가장 우선 고려할 수 있지만, 군집화 이상치가 차원 수에 많은 영향을 받음을 감안하면 SVM 알고리즘의 사용도 고려해 볼 수 있다.

본 연구에서는 테이블 데이터의 이상탐지를 위한 알고리즘 선택에 앞서 이상치 유형을 파악하는 것이 중요하다는 것을 강조하고, 비지도 학습 방법이 이상탐지에서 항상 유리한 것은 아니라는 점을 확인하였다. 또한 데이터 차원 수가

많을수록 성능 향상에 도움이 되지만, 군집화 이상치 유형에서는 많은 수의 차원이 오히려 이상탐지를 어렵게 할 수 있음을 확인하였다.

본 연구에서는 공정한 테이블 데이터의 이상치 유형별 비교를 위해, 기존 이상탐지 벤치마크에서 널리 사용되던 공개 데이터 세트들 중 일부를 실험에 사용했다. 이로 인해 데이터 개수와 차원 수에 대한 비교에서 일부 한계가 있었고, 개별 알고리즘의 하이퍼파라미터를 최적화하는 과정을 다루지 않았으므로, 향후 이 과정을 포함하는 실험을 통해 더욱 엄밀한 성능 평가가 이루어질 필요가 있다. 또한 최근 직접 알고리즘을 개발하기 보다는 허깅 페이스 등을 통해 다양한 모델을 선택해서 이용할 수 있게 됨에 따라 분석 문제 혹은 데이터 특성에 따른 모델 추천이 더욱 중요해진 점을 감안하여, 향후 허깅 페이스 등과 같은 플랫폼에서 이러한 알고리즘 성능 비교 결과를 활용할 수 있는 방안에 대해서도 추후 연구에서 폭 넓은 논의가 이루어져야 한다.

참고문헌(References)

Aggarwal, C. C., & Aggarwal, C. C. (2017). An introduction to outlier analysis (pp. 1-34). Springer International Publishing.

Akcay, S., Atapour-Abarghouei, A., & Breckon, T. P. (2019). Ganomaly: Semi-supervised anomaly detection via adversarial training. In Computer Vision – ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2 – 6, 2018, Revised Selected Papers, Part III 14 (pp. 622-637). Springer International Publishing.

Ayres-de-Campos, D., Bernardes, J., Garrido, A., Marques-de-Sa, J., & Pereira-Leite, L. (2000). SisPorto 2.0: a program for automated analysis of cardiotocograms. Journal of Maternal-Fetal Medicine, 9(5), 311-318.

Aas, K., Czado, C., Frigessi, A., & Bakken, H. (2009). Pair-copula constructions of multiple dependence. Insurance: Mathematics and economics, 44(2), 182-198.

Boyd, K., Eng, K. H., & Page, C. D. (2013). Area under the precision-recall curve: point estimates and confidence intervals. In Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13 (pp. 451-466). Springer Berlin Heidelberg.

Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000, May). LOF: identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD international conference on Management of data (pp. 93-104).

Campos, G. O., Zimek, A., Sander, J., Campello, R. J., Micenková, B., Schubert, E., ... & Houle, M. E. (2016). On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. Data mining and knowledge discovery, 30, 891-927.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Choi, N. W. & Kim, W. (2019). Anomaly Detection for User Action with Generative Adversarial Networks. 25(3), 43-62.

Cortes, C., & Vapnik, V. (1995). Support-vector

networks. Machine learning, 20, 273-297.

Domingues, R., Filippone, M., Michiardi, P., & Zouaoui, J. (2018). A comparative evaluation of outlier detection algorithms: Experiments and analyses. Pattern recognition, 74, 406-421.

Emmott, A., Das, S., Dietterich, T., Fern, A., & Wong, W. K. (2015). A meta-analysis of the anomaly detection problem. arXiv preprint arXiv:1503.01158.

Han, S., Hu, X., Huang, H., Jiang, M., & Zhao, Y. (2022). Adbench: Anomaly detection benchmark. Advances in Neural Information Processing Systems, 35, 32142-32159.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer.

He, Z., Xu, X., & Deng, S. (2003). Discovering cluster-based local outliers. Pattern recognition letters, 24(9-10), 1641-1650.

Huang, H., Qin, H., Yoo, S., & Yu, D. (2012, October). Local anomaly descriptor: a robust unsupervised algorithm for anomaly detection based on diffusion space. In Proceedings of the 21st ACM international conference on Information and knowledge management (pp. 405-414).

Hwang, C. (2022). Resolving data imbalance through differentiated anomaly data processing based on verification data. Journal of Intelligence and Information Systems, 28(4), 179-190.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30.

Kriegel, H. P., Kröger, P., Schubert, E., & Zimek, A. (2009). Outlier detection in axis-parallel subspaces of high dimensional data. In Advances in Knowledge Discovery and Data Mining: 13th Pacific-Asia Conference, PAKDD 2009 Bangkok, Thailand, April 27-30, 2009 Proceedings 13 (pp. 831-838). Springer Berlin Heidelberg.

Lai, K. H., Zha, D., Wang, G., Xu, J., Zhao, Y., Kumar, D., ... & Hu, X. (2021, May). Tods: An automated time series outlier detection system. In Proceedings of the aaai conference on artificial intelligence (Vol. 35, No. 18, pp. 16060-16062).

Lee, D. H. & Kim. N. (2022). Anomaly detection methodology based on multimodal deep learning. 28(2), 101-125.

Li, Z., Zhao, Y., Botta, N., Ionescu, C., & Hu, X. (2020, November). COPOD: copula-based outlier detection. In 2020 IEEE international conference on data mining (ICDM) (pp. 1118-1123). IEEE.

Li, Z., Zhao, Y., Hu, X., Botta, N., Ionescu, C., & Chen, G. (2022). Ecod: Unsupervised outlier detection using empirical cumulative distribution functions. IEEE Transactions on Knowledge and Data Engineering.

Liu, B., Tan, P. N., & Zhou, J. (2022, June). Unsupervised anomaly detection by robust density estimation. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 36, No. 4, pp. 4101-4108).

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008, December). Isolation forest. In 2008 eighth ieee international conference on data mining (pp. 413-422). IEEE.

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2010). On detecting clustered anomalies using SCiForest.

In Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part II 21 (pp. 274-290). Springer Berlin Heidelberg.

Liu, K., Dou, Y., Zhao, Y., Ding, X., Hu, X., Zhang, R., ... & Yu, P. S. (2022). Pygod: A python library for graph outlier detection. arXiv preprint arXiv:2204.12095.

Malerba, D., Esposito, F., & Semeraro, G. (1996). A further comparison of simplification methods for decision-tree induction. Learning From Data: Artificial Intelligence and Statistics V, 365-374.

Mangasarian, O. L., Street, W. N., & Wolberg, W. H. (1995). Breast cancer diagnosis and prognosis via linear programming. Operations research, 43(4), 570-577.

Martinez-Guerra, R., & Mata-Machuca, J. L. (2016). Fault detection and diagnosis in nonlinear systems. Springer International Pu.

Pang, G., Cao, L., Chen, L., & Liu, H. (2018, July). Learning representations of ultrahigh-dimensional data for random distance-based outlier detection. In Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 2041-2050).

Pang, G., Shen, C., & van den Hengel, A. (2019, July). Deep anomaly detection with deviation networks. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 353-362).

Pevný, T. (2016). Loda: Lightweight on-line detector of anomalies. Machine Learning, 102, 275-304.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. Advances in neural information processing systems, 31.

Ramaswamy, S., Rastogi, R., & Shim, K. (2000, May). Efficient algorithms for mining outliers from large data sets. In Proceedings of the 2000 ACM SIGMOD international conference on Management of data (pp. 427-438).

Ruff, L., Kauffmann, J. R., Vandermeulen, R. A., Montavon, G., Samek, W., Kloft, M., ... & Müller, K. R. (2021). A unifying review of deep and shallow anomaly detection. Proceedings of the IEEE, 109(5), 756-795.

Ruff, L., Vandermeulen, R. A., Görnitz, N., Binder, A., Müller, E., Müller, K. R., & Kloft, M. (2019). Deep semi-supervised anomaly detection. arXiv preprint arXiv:1906.02694.

Shyu, M. L., Chen, S. C., Sarinnapakorn, K., & Chang, L. (2003). A novel anomaly detection scheme based on principal component classifier. Miami Univ Coral Gables Fl Dept of Electrical and Computer Engineering.

Soenen, J., Van Wolputte, E., Perini, L., Vercruyssen, V., Meert, W., Davis, J., & Blockeel, H. (2021). The effect of hyperparameter tuning on the comparative evaluation of unsupervised anomaly detection methods. In Proceedings of the KDD'21 Workshop on Outlier Detection and Description (pp. 1-9). Outlier Detection and Description Organising Committee.

Steinbuss, G., & Böhm, K. (2021). Benchmarking unsupervised outlier detection with realistic synthetic data. ACM Transactions on Knowledge Discovery from Data (TKDD), 15(4), 1-20.

Tang, J., Chen, Z., Fu, A. W. C., & Cheung, D. W. (2002). Enhancing effectiveness of outlier detections for low density patterns. In Advances in Knowledge Discovery and Data Mining: 6th Pacific-Asia Conference, PAKDD 2002

Taipei, Taiwan, May 6–8, 2002 Proceedings 6 (pp. 535-548). Springer Berlin Heidelberg.

Zhao, Y., & Hryniewicki, M. K. (2018, July). Xgbod: improving supervised outlier detection with unsupervised representation learning. In 2018 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.

Zhao, Y., Nasrullah, Z., & Li, Z. (2019). Pyod: A python toolbox for scalable outlier detection. arXiv preprint arXiv:1901.01588.

Zong, B., Song, Q., Min, M. R., Cheng, W., Lumezanu, C., Cho, D., & Chen, H. (2018, February). Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In International conference on learning representations.

Zhou, Z. H. (2018). A brief introduction to weakly supervised learning. National science review, 5(1), 44-53.

Abstract

Performance Comparison of Anomaly Detection Algorithms: in terms of Anomaly Type and Data Properties

Jaeung Kim* · Seung Ryul Jeong* · Namgyu Kim**

With the increasing emphasis on anomaly detection across various fields, diverse anomaly detection algorithms have been developed for various data types and anomaly patterns. However, the performance of anomaly detection algorithms is generally evaluated on publicly available datasets, and the specific performance of each algorithm on anomalies of particular types remains unexplored. Consequently, selecting an appropriate anomaly detection algorithm for specific analytical contexts poses challenges. Therefore, in this paper, we aim to investigate the types of anomalies and various attributes of data. Subsequently, we intend to propose approaches that can assist in the selection of appropriate anomaly detection algorithms based on this understanding. Specifically, this study compares the performance of anomaly detection algorithms for four types of anomalies: local, global, contextual, and clustered anomalies. Through further analysis, the impact of label availability, data quantity, and dimensionality on algorithm performance is examined. Experimental results demonstrate that the most effective algorithm varies depending on the type of anomaly, and certain algorithms exhibit stable performance even in the absence of anomaly-specific information. Furthermore, in some types of anomalies, the performance of unsupervised anomaly detection algorithms was observed to be lower than that of supervised and semi-supervised learning algorithms. Lastly, we found that the performance of most algorithms is more strongly influenced by the type of anomalies when the data quantity is relatively scarce or abundant. Additionally, in cases of higher dimensionality, it was noted that excellent performance was exhibited in detecting local and global anomalies, while lower performance was observed for clustered anomaly types.

Key Words : Anomaly Detection, Anomaly Types, Performance Comparison, Algorithm Selection

Received : August 26, 2023 Revised : September 7, 2023 Accepted : September 8, 2023
Corresponding Author : Namgyu Kim

* Graduate School of Business IT, Kookmin University
** Corresponding author: Namgyu Kim
Graduate School of Business IT, Kookmin University
77 Jeongneung-ro, Seongbuk-gu, Seoul 136-702, Korea
Tel: +82-2-910-5425, Fax: +82-2-910-4017, E-mail: ngkim@kookmin.ac.kr

저 자 소 개

김재웅

현재 국민대학교 비즈니스IT전문대학원 석사과정에 재학 중이며, 원광대학교 경영학과에서 경영학을 전공하여 학사 학위를 취득하였다. 주요 관심분야는 시계열 데이터 처리, 이상탐지, 빅데이터 분석 및 활용 등이다.

정승렬

미국 사우스 케롤라이나 대학에서 경영정보학 박사를 취득하고 현재 국민대학교 비즈니스IT전문대학원 교수로 재직 중이다. Journal of MIS, Communications of the ACM, Information and Management, Journal of Systems and Software, Online Information Review, APJIS, 경영과학, ISR, 정보처리학회지 등의 국내외 저널에 프로세스 혁신, ERP, 정보자원관리, 시스템 구현, 프로젝트 관리, 빅데이터 등의 주제와 관련하여 많은 논문을 발표하였다

김남규

현재 국민대학교 비즈니스IT전문대학원 및 경영정보학부 교수로 재직 중이다. 서울대학교 컴퓨터공학과에서 학사 학위를 취득하고, KAIST 테크노경영대학원에서 Database와 MIS를 전공하여 경영공학 석사 및 박사학위를 취득하였다. 한국지능정보시스템학회 부회장, 한국정보기술응용학회 부회장, 한국경영학회 상임이사, 한국경영정보학회 이사, 한국인터넷정보학회 이사 등을 역임하였으며, 주요 관심분야는 텍스트 마이닝, 딥러닝, 데이터 모델링 등이다.