공학석사 학위논문

비지도학습 이상치 탐지의 데이터 이상정도 측정 지표에 관한 연구

A study on an Outlierness metric for Unsupervised Outlier Detection

2019년 2월

서울과학기술대학교 일반대학원

데이터사이언스학과

나 선 민

비지도학습 이상치 탐지의 데이터 이상정도 측정 지표에 관한 연구

A study on an Outlierness metric for Unsupervised Outlier Detection

지도교수 조남욱

이 논문을 공학석사 학위논문으로 제출함

2019년 2월

서울과학기술대학교 일반대학원

데이터사이언스학과

나 선 민

나선민의 공학석사 학위논문을 인준함

2019년 2월

심사위원장 금영정 (인)

심사위원 조남욱 (인)

심사위원 권혁윤 (인)

요 약

제 목 : 비지도학습 이상치 탐지의 데이터 이상정도 측정 지표에 관한 연구

이상치 탐지는 데이터 마이닝을 기반으로 한 데이터 분석 기법 중의 하나로서, “어떤 데이터 안에서 다른 관측값들과 다른 방법에 의해 생성되었다고 의심되는 관측값” (Hawkins, 1980)인 이상치를 탐지하는 기법이다. 이상치 탐지는 IT 보안, 의료진단, 제조공정에서의 모니터링 등 다양한 산업분야에 적용되고 있으며 최근 데이터의 생산량이 늘어남과 함께 그 활용분야가 점차 확대되고 있다.

최근 이상치 탐지에 관한 기존 연구들은 지도학습(supervised learning)에 기반하여 제안된 알고리즘의 성능을 비교하거나 특정 데이터셋에 맞는 알고리즘 추천에 관한 연구가 주를 이루고 있다. 주어진 데이터의 이상치 유무를 알 수 없는 경우 일반적으로 비지도 학습(unsupervised learning) 기반의 이상치 탐지 기법이 적용된다. 비지도학습 이상치 탐지에서 모델 성능 평가는 연구자의 주관에 상당부분 의존할 수밖에 없으며(Aggarwal, 2013), 어떤 데이터셋에 실제로 이상치가 포함되어있는지, 또 얼마나 포함되어있는지를 객관적으로 증명하는 외부지표(external measure)에 대한 필요성이 커지고 있지만 이에 관련된 연구는 아직까지 미흡한 실정이다.

본 연구는 이러한 한계점을 극복하고자 데이터의 이상정도(outlierness)를 측정할 수 있는 두 가지 지표의 활용방안을 제안하고자 한다. 데이터 불순도를 측정하는 지표인 엔트로피(entropy)와 데이터 분포의 불평등을 측정하는 지니계수(Gini index)를 이용하여 인공적으로 생성한 데이터셋에 대해 실험을 진행한 후 두 지표가 데이터의 이상 정도를 객관적으로 판단할 수 있는 지표임을 확인하였다. UCI machine learning repository에서 수집한 10개의 실제 데이터셋에도 실험을 진행하여 지표의 실제 효용성을 확인하였다. 엔트로피와 지니계수 모두 데이터의 이상 정도를 측정할 수 있는 지표임을 확인했으며, 제안된 지표를 이용하여 데이터의 이상정도를 정량화 할 수 있음을 보였다.

본 연구는 이상치 탐지 단에 앞서 주어진 데이터가 이상치를 포함하고 있는지를 확인할 수 있는 새로운 지표를 제안했다는 점에서 그 의의가 있으며, 추후 이를 활용하여 다양한 산업분야에서 응용이 가능할 것으로 기대된다.

목 차

표 목 차

그림목차

1. 서론

1.1 연구의 배경

이상치 탐지는 데이터 마이닝 분야의 한 분야로서 데이터셋에 존재하는 소량의 이상치를 탐지하는 데이터 분석 기법이다. 최근 모바일 디바이스 및 IoT 기술 등의 발전으로 많은 양의 데이터가 생성되고 있는 환경에서, 이상치 탐지에 관한 연구는 네트워크 보안 시스템에서 침입을 탐지하거나 제조공정시스템에서 불량품을 탐지 하는 등 대부분의 산업분야에서 활용 가능하다는 점에서 더욱 의미가 있다.

이상치란 “어떤 데이터 안에서 다른 관측값들과 다른 방법에 의해 생성되었다고 의심되는 관측값” (Hawkins, 1980) 또는 “이상치는 같은 표본 안에 존재하는 다른 관측값들과 비교하여 현저히 다른 관측치“(Barnet and Lewis, 1994)를 말한다. 대개 이상치는 데이터 안에 극소량 존재하기 때문에 훈련 데이터를 구축하는 것이 쉽지 않아 일반적인 기계학습의 분류 문제로 접근하는 것이 어렵다. 이에 따라 이상치를 탐지하는 알고리즘이 꾸준히 제시되어 학계 및 산업분야에서 활용되고 있다.

기존의 이상치 탐지 연구들은 다양한 이상치 탐지 알고리즘을 데이터셋에 적용하여 모델이 탐지해낸 예상 이상치 검출률을 기반으로 각 모델의 성능을 비교하거나 데이터 특성에 맞는 알고리즘을 제안하는데 제한되어 있다. 그러나 이러한 선행 연구들은 이상치 검출을 지도학습 방법을 사용하여 접근하기 때문에 정답에 관한 정보가 존재하지 않는 실제 데이터셋을 다루기에는 적합한 방법이라고 볼 수 없다. 또, 비지도학습의 성격을 띄는 이상치 탐지 문제 특성상 이상치 선별은 연구자의 주관에 따를 수 밖에 없으며(Aggarwal, 2013), 이 외에도 주어진 데이터셋에 이상치가 존재하는지조차 알 수 없는 경우가 대부분이다. 따라서 이러한 이상정도를 측정할 수 있는 객관적인 외부지표(external measure)의 필요성이 커지고 있지만, 데이터 분석기법인 클러스터링 기법에서 엔트로피를 기반으로 한 외부성능지표가 활용된 사례(Zimek et al., 2014)를 제외하고 아직 이상치 탐지 분야에서 이에 관련한 연구는 활발히 진행되고 있지 않다.

따라서 데이터의 이상정도(outlierness)를 측정할 수 있는 지표가 제안된다면 이상치를 선별해내는 탐지 단계에 앞서 데이터를 이해하고 그에 맞는 알고리즘을 적용하는 데 도움이 될 수 있을 것으로 기대된다.

1.2 연구의 목적

본 연구의 목적은 주어진 데이터 안에 실제로 이상치가 포함되어 있는지를 객관적인 수치를 통해 확인할 수 있는 외부 지표를 제안하고, 실험을 통해 지표의 효용성을 검증하는 데 있다. 제안된 지표는 데이터 자체의 이상치 정도를 객관적으로 제시함으로써 데이터 자체의 이상치 정도를 객관적으로 측정할 수 있다. 이를 통해 기존 연구에서 주로 사용되던 내부 지표(정확도, 재현율, 정밀도 등)의 한계점을 극복하고 데이터셋 자체의 이상치정도를 측정하는 외부지표를 제시하고자 한다.

1.3 연구 방법

이상치 정도를 나타내는 지표의 타당성을 평가하기 위해 먼저 실험 데이터셋을 만들어 이상 정도를 조금씩 늘려가며 지표의 변화 정도를 관찰하고 외부지표의 적용가능성을 모색하였다. 이를 토대로 데이터셋의 이상치를 나타낼 수 있는 지표 후보군을 도출하였다. 그 후 UCI(University of California, Irvine)에서 제공하는 기계학습용 데이터셋 중 최근 10년간 이상치 탐지 논문에서 자주 쓰인 데이터셋 10개를 선정하여 실험을 통해 도출된 지표의 효용성을 검증하였다.

1.4 연구의 개요

본 논문은 5장으로 구성된다. 제 2장에서는 관련 문헌 연구와 본 연구에서 쓰인 알고리즘 LOF(local outlier factor)와 K-NN(K-nearest neighbor) 알고리즘을 적용한 이상치 탐지 방법 알고리즘을 요약 설명한다. 제 3장에서는 본 연구에서 제시하는 지표의 배경이론인 엔트로피(entropy)와 지니계수(Gini index)에 관한 설명과 함께 이 지표를 실험 데이터셋에 적용하는 과정을 설명한다. 제 4장에서는 3장에서 제안한 지표를 데이터셋에 적용하여 실험한 결과와 그 변화 양상을 서술하며 마지막 제 5장에서는 연구의 결론 및 한계점을 정리한다.

2. 이론적 배경

2.1 이상치 탐지

2.1.1 이상치 탐지 방법론

이상치 탐지는 크게 두 가지 방법론으로 나뉜다. 먼저 지도학습 이상치 탐지(supervised outlier detection)의 경우, 정상 데이터와 이상치 데이터에 관한 클래스 정보가 존재하여 이를 바탕으로 훈련 데이터를 구성할 수 있을 때 쓰이는 방법이다. 훈련 데이터로 모델을 학습 시킨 후 새로운 데이터가 주어졌을 때 이 데이터가 이상치인지 정상치인지를 판별하는 방법론이다. 그러나 대부분의 실제 이상치 탐지 문제는 정답이 존재하지 않는 비지도학습의 성격을 띄기 때문에 이러한 지도학습적 접근법은 효용성이 떨어진다고 볼 수 있다.

그에 반해 비지도학습 이상치탐지(unsupervised Outlier Detection)은 모델을 학습시킬 데이터가 구축이 어렵다. 비지도학습 특성 상 훈련할 수 있는 이상치에 관한 과거 정보를 학습할 수 없기 때문에 모델과 그 모델 안에서 조절할 수 있는 모수(parameter) 선택에 따라 이상치 검출 결과가 상이할 수 있다는 점이 가장 큰 한계점으로 꼽힌다. 본 연구에서는 이러한 비지도학습 이상치 탐지의 한계점을 극복하고자 하며 그에 맞춰 앞으로는 비지도학습 방법론을 적용한 이상치 탐지에 제한하여 서술하고자 한다.

2.1.2 비지도 학습 이상치 탐지 알고리즘의 종류

최근까지 비지도학습 이상치 탐지를 위한 많은 알고리즘 및 방법론이 제시되었다. 이를 몇 가지 군으로 묶어 분류한 분류체계가 다수의 기존 연구에서 제시되었지만 본 연구에서는 (Pimentel et al., 2014)이 제시한 분류 체계를 따르고자 한다.

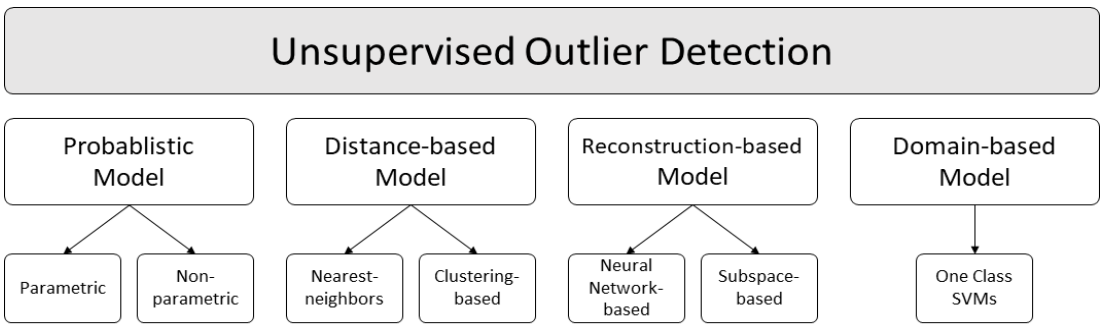

Fig 2.1 비지도학습 이상치 탐지의 분류체계 (Pimentel et al., 2014)

이 외에도 통계학적 접근을 기반으로 한 알고리즘(HBOS; Histogram-based Outlier Score) 등 다양한 접근법이 있지만 현재 널리 쓰이는 이상치 탐지 알고리즘은 거리기반(distance-based) 알고리즘, 특히 인접이웃(nearest-neighbors)을 기반으로 한 알고리즘이 대부분이며 많은 데이터셋을 대상으로 한 실험에서 인접이웃기반 알고리즘이 다른 알고리즘보다 평균적으로 우수한 성능을 보였다(Goldstein and Uchida, 2016). 따라서 본 연구에서 제안하는 지표 제안에 있어 인접 이웃 관련 알고리즘을 선택하였고, 그렇게 선택한 인접이웃 이상치 탐지 알고리즘에 관해 먼저 설명하도록 한다.

2.1.3 인접이웃기반(nearest neighbors-based) 알고리즘

인접이웃기반 알고리즘에는 k-Nearest Neighbors(K-NN) 이상치 탐지 기법과, Local Outlier Factor(LOF), Local Outlier Probabilities(LoOP) (Kriegel et al., 2009), Local Correlation Integral(LOCI) (Papadimitriou et al., 2003) 등 다양한 알고리즘이 사용되고 있지만 그 중에서도 본 연구에서는 이상치 탐지 알고리즘 선택에 있어 가장 먼저 추천되는 두 개의 알고리즘(Goldstein and Uchida, 2016), K-NN과 LOF를 선정하여 활용하였다.

2.1.4 k-Nearest Neighbor (K-NN)

K-NN 기반의 이상치 탐지 기법은 전역적(global) 이상치 탐지 기법 중 하나이다. 하나의 데이터셋을 구성하고 있는 정상 관측치들과 몹시 다른 관측지를 전역적 이상치 라고 하며, K-NN은 이러한 전역적 이상치들을 탐지해 내는 데 가장 널리 쓰이는 알고리즘 중 하나이다. K-NN 모델에서는 정상치들은 이웃 정상치들과 근접해 있고, 이상치들은 이러한 정상치들에 멀리 떨어져 있음을 가정한다. (Hautamaki et al., 2004) 이웃들에 비해 멀리 떨어져 위치한 데이터 포인트는 이상치로 간주되며, 이러한 거리 계산에는 주로 유클리디안 거리(euclidean distance)가 주로 사용되나 마할라노비스(Mahalanobis) 또는 민코프스키(Minkowski)와 같은 거리 계산 방법도 사용될 수 있다(Pimentel et al., 2014). k-NN 접근법에서는 k개의 근접 이웃까지 거리를 이용하여 각 데이터에 대한 이상치 점수를 계산한다. 아래 그림에서처럼, 파란색 데이터의 이웃까지의 거리가 다른 데이터들에 비해 크기 때문에 이 데이터 포인트는 이상치라고 볼 수 있다.

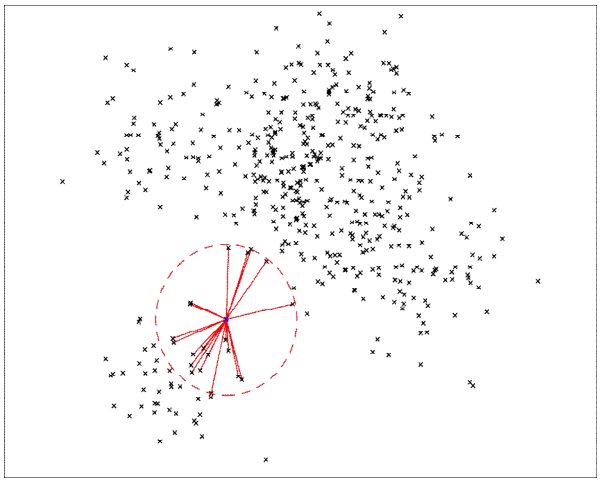

Fig 2.2 K-NN 기반 이상치 탐지 알고리즘 도식화

본 연구에서는 이웃간의 거리를 계산할 때 민코프스키 거리를 사용했다. 유클리드 공간에서 점$(x_1, x_2, .., x_n)$과 점 $(y_1, y_2, .., y_n)$이 주어졌을 때, p차 민코프스키 거리는 식 (1)과 같이 정의된다.

$$d(x,y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{1/p} \quad (1)$$

2.1.5 Local Outlier Factor (LOF)

Local Outlier Factor(LOF) 는 (Breunig et al., 2000) 에 의해 제안된 밀도 기반 지역적(local) 이상치 탐지 기법으로 데이터 공간 내에서 다른 개체들에 비해 밀도가 낮게 측정 되는 데이터들이 이상치라고 가정한다. 이 알고리즘 역시 데이터 포인트 마다 이상치 점수(LOF score)를 계산하며 이 점수가 높을수록 이상치일 확률이 높음을 암시한다. LOF는 데이터 포인트의 주변 밀도를 고려하여 이상치를 탐지하기 때문에, 전역 이상치(global outlier) 뿐만 아니라 지역 이상치(local outlier)를 선별할 수 있다는 점에 그 장점이 있다(김승 et al., 2010). 아래 그림에서 O1은 전역적 이상치로 K-NN과 같은 전역적 이상치 탐지 알고리즘에서는 탐지가 가능하지만 O2와 같은 지역적 이상치는 탐지가 어렵다. 그러나 밀도기반인 LOF 알고리즘에서는 가장 가까운 이웃들의 밀도에 비해 현저히 낮은 밀도를 보여주는 O2도 이상치라고 판별하기 때문에

이와 같은 지역적 이상치를 탐지해내는 문제에 있어 높은 성능을 보여주는 알고리즘이다.

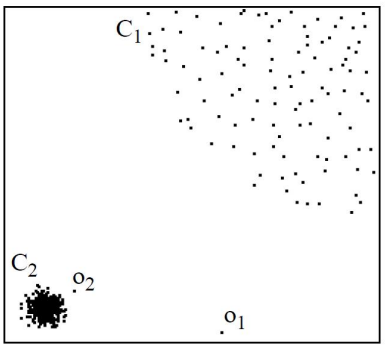

Fig 2.3 LOF에서 탐지가능한 이상치 (Breunig et al., 2000)

LOF 알고리즘은 아래와 같은 과정을 거쳐 데이터셋 내에 존재하는 모든 데이터 포인트에 대해 이상치 점수를 부여한다.

1) 모든 데이터 개체 p에 대해 k번째 인접이웃사이의 거리 k-distance(p)를 구한다.

2) 모든 데이터 개체 p와 데이터 셋의 다른 개체 q까지의 도달가능거리(reachability distance)를 아래와 같이 계산한다.

$$Reachability-dist_k(p,q) = \max\{k-distance(q), dist(p,q)\} \quad (2)$$

3) 모든 데이터 개체 p에 대해 지역 도달 가능 밀도 (local reachability density: lrd)를 아래와 같이 계산한다.

$$lrd_k(p) = \frac{|N_k(p)|}{\sum_{O \in N_k(p)} reachability-distance_k(p,o)} \quad (3)$$

여기서 $N_k(p)$는 데이터 포인트 p의 k번째 인접이웃 집합 안에 존재하는 개체 수이다. 즉, 데이터 p의 lrd는 p로부터 다른 개체들까지의 도달가능거리들

의 평균값을 거꾸로 뒤집은 것과 같다.

4) 모든 데이터 개체 p에 대해 최종적으로 LOF점수를 아래와 같이 계산한다.

$$LOF_k(p) = \frac{\sum_{q \in N_k(p)} \frac{lrd_k(q)}{lrd_k(p)}}{|N_k(p)|} \quad (4)$$

이렇게 나온 점수는 개체 p의 k번째 인접 이웃에 속한 q의 지역 도달가능밀도의 평균을 p의 k번째 인접이웃 집합 안에 존재하는 개체 수만큼 나눈 것이기 때문에 데이터 포인트 p가 얼마나 이상치인가를 나타내는 지표라고 볼 수 있다. LOF는 데이터 내에 존재하는 밀집된 클러스터에서 조금만 떨어져 있는 값이라도 이상치로 탐지해 낼 수 있기 때문에, 흔히 쓰이는 거리 기반 알고리즘으로는 탐지가 어려운 지역이상치를 탐지하는데 좋은 알고리즘이라고 알려져 있다(Goldstein and Uchida, 2016).

2.2 이상치 탐지 알고리즘의 성능 측정

이상치 탐지 알고리즘의 성능을 측정하는 것은 쉽지 않다고 알려져 있다. 이상치는 대개 데이터셋 안에 극소량 존재하며 특히 비지도학습 이상치 탐지의 경우 지도학습 분류와 같이 정상치 혹은 이상치를 판별할 수 있는 클래스 정보가 존재하지 않기 때문이다. 이상치 탐지 문제는 대개 데이터내 불균형이 심하기 때문에, 알고리즘의 성능을 측정할 때 기존의 정확도를 사용하는 것은 부적절하다(Tang and He, 2017). 이와 같은 이유로 지금까지 이상치 탐지 연구 논문에서는 이상치 탐지를 불균형이 심한 분류문제로 접근하여 알고리즘의 성능을 정밀도와 재현율로 측정하는 시도가 활발히 이루어 지고 있다. 데이터셋에 존재하는 아주 드문 클래스를 잠재이상치라고 선정한 후 이 이상치들을 얼마나 정확하게 탐지해 내는가에 관해 정밀도(precision)와 재현율(recall)을 평가함으로써 모델의 성능을 평가할 수 있다. 정밀도와 재현율은 아래와 같이 계산된다.

Table 2.1 정밀도와 재현율 계산을 위한 confusion matrix

		실제 정답	
		true	false
모델 실험 결과	true	True Positive (TP)	False Positive (FP)
	false	False Negative (FN)	True Negative (TN)

$$\text{정밀도 } (precision) = \frac{TP}{TP + FP}$$

$$\text{재현율 } (recall) = \frac{TP}{TP + FN}$$

이외에도 ROC AUC(Area Under the Receiver Operating Characteristics Curve)를 이용한 성능 측정 방법이 널리 쓰이고 있으며, 이상치 탐지에 있어서는 ROC AUC가 가장 효율적인 지표임이 증명된 적 있다. (Campos et al., 2016) 또, ROC AUC는 상대 빈도를 이용하여 클래스 불균형 문제를 본질적으로 다루기 때문에 이상치 탐지 모델의 성능평가에서 가장 널리 쓰이고 있는 지표이다(Zimek et al., 2014).

그러나 앞서 언급한 것처럼, 비지도학습 이상치 탐지에서 모델 성능 평가 과정은 연구자의 주관에 상당부분 의존할 수밖에 없으며(Aggarwal, 2013), 이러한 한계점을 극복하고자 어떤 데이터셋에 실제로 이상치가 포함되어있는지, 또 얼마나 포함되어있는지를 객관적으로 증명하는 지표에 관한 연구는 아직까지 미흡한 실정이다.

2.2.1 불순도 평가 지표

본 연구에서는 데이터셋의 불순도를 평가하는 두 가지 지표 엔트로피(entropy)와 지니계수(Gini index)를 활용함으로써 데이터의 객관적인 이상치 정도(outlierness)를 측정하고자 한다.

2.2.2 엔트로피(Shannon entropy)

엔트로피(Shannon entropy)는 정보이론(information theory)에서 제시된 개념으로, 불확실성을 측정할 때 쓰이는 지표이다. 분포 P를 따르는 확률변수 H(X)의 엔트로피는 모든 사건 정보량의 기댓값으로, 아래와 같이 표현할 수 있다.

$$H(P) = H(X) = E_{X\sim P}[I(X)] = -E_{X\sim P}[\log P(X)] \quad (5)$$

이 때, k가지 상태를 가진 이산 변수의 경우 엔트로피는 아래와 같이 표현 가능하다.

$$H(X) = -\sum_{k=1}^{K} p(X=k)\log p(X=k) \quad (6)$$

엔트로피는 어떤 정보량을 평가하거나 어떤 변수의 불확실성을 평가할 수 있는 강력한 지표이며 (Shannon, 1948) 또, 어떤 시스템의 무작위성(randomness)을 측정할 수 있는 지표이기도 하다. 어떤 한 데이터셋을 구성하는 데이터 포인트 중에서 이상치가 많이 섞여있을 경우 그 데이터셋의 전체 엔트로피가 증가하기 때문에 엔트로피는 이상치 탐지에 접목함에 있어 매우 직관적인 개념이며 (Toshniwal and Eshwar, 2014) 최근 다양한 이상치 탐지 분야에서 연구되고 있다. (Liu et al., 2008; Jiang et al., 2010; Koufakou et al., 2007; He et al., 2005)

2.2.3 지니계수(Gini index)

지니계수(Gini index)는 어떤 데이터셋에 이질적인 것이 얼마나 섞여있는지를 측정하는 불순도 측정 지표이며 데이터 마이닝 분야에서는 결정나무모델(decision tree)에서 분기를 할 때 분기의 효율성을 측정하는 지표로서 활용되고 있다. 어떤 집합에 있는 데이터 포인트들이 모두 같다면 지니계수는 최솟값(0)을 갖게 되며, 이 집합은 불순도가 0인 집합이다. $i \in \{1, 2, \ldots, m\}$인 i가 있을 때 f_i를 i로 표시된 집합 안의 항목 부분이라고 했을 때, 지니 계수는 아래와 같이 계산된다.

$$G.I.(f) = \sum_{i=1}^{m} f_i(1-f_i) = 1 - \sum_{i=1}^{m} f_i^2 \quad (7)$$

아직까지 지니계수를 활용한 이상치 탐지에 관한 연구는 진행된 바 없으며 따라서 본 연구에서는 지니계수와 엔트로피를 활용하여 기존의 비지도학습 이상치 탐지의 객관적인 이상치정도(outlierness)를 판단할 수 있는 지표를 제안하고자 한다. 이어지는 제 3장에서는 연구 방법론에 대해 서술한다.

3. 연구 방법

3.1 연구의 구성

본 연구는 아래와 같이 구성된다. 먼저 첫 번째 단계에서는 앞서 제 2장에서 서술한 두 가지 불순도 지표(엔트로피, 지니계수)를 바탕으로 실험 데이터셋을 만들어 무작위성(randomness)을 일정한 단위로 추가한 후 그때의 엔트로피와 지니계수의 증감률을 관찰한다. 실험 데이터셋에 무작위성을 추가하는 알고리즘은 아래와 같다.

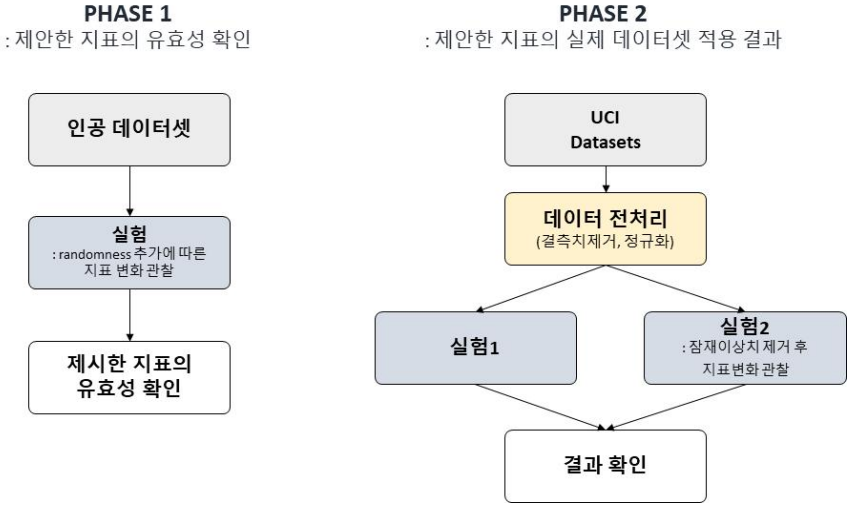

Fig 3.1 연구프레임워크

Algorithm: Add-Randomness

Input: Dataset D containing 400 data points equally distributed in 20x20 grid

1: **start** with $\eta = 0.2$, l = length of input data, $0 \leq x, y \leq 20$

2: **select** $\eta * l$ data points from D

3: **generate** new data points by the number of selected data points
: new$x_i = rand(0,1) * 20$, new$y_i = rand(0,1) * 20$

4: **replace** the data points with newly generated data in 3.

5: **do** Entropy and Gini index calculation

6: **increase** η by 0.2

7: **foreach** η **repeat** 2~5

8: **End**

Output: η percentage of randomness added 5 dataset $D_1^*, ..., D_5^*$

이렇게 실험 데이터셋에 대해 실험한 결과를 바탕으로 두 지표가 실제로 그 효용성을 입증함을 확인한 후 UCI 데이터셋 10개를 대상으로 같은 실험을 반복하여 실제 데이터셋에서도 그 효용이 입증됨을 확인한다.

3.2 UCI 데이터셋 실험 구성

UCI machine learning repository에서 수집한 데이터셋은 최근 10년간 이상치 탐지 분야에서 가장 많이 쓰인 데이터셋 10개를 선정했으며 해당 데이터셋에 관한 정보는 표1에서 찾을 수 있다. 몇 개의 데이터셋에서는 기존 문헌 연구(Campos et al., 2016)에서와 같이 다운샘플링을 통해 이상치의 수를 줄임으로써 실제 이상치 탐지 문제에 가깝도록 재설계하였다.

Table 3.1 UCI repository dataset

No.	Dataset	Normal class	Outlier class	# of instances	# of feature	Outlier (%)
1	PageBlock	1	2, 3, 4, 5	5,473	10	560 **(10.2%)**
2	Cardio	1(normal)	3 (pathologic)	1,831	21	176 **(9.6%)**
3	HTRU2	0	1	17,898	8	1,639 **(9.2%)**
4	Shuttle	1	2,3,4,5,6,7	12,345	9	867 **(7%)**
5	Wilt	N (others)	W (diseased trees)	4,839	5	261 **(5.4%)**
6	Glass	Others	6	214	7	9 **(4.2%)**
7	Waveform	Others	0 (downsampled)	3,443	21	100 **(2.9%)**
8	WDBC	Benign	Malignant (downsampled)	367	30	10 **(2.7%)**
9	Annthyroid	3, 2	1	3,772	21	93 **(2.5%)**
10	PenDigits	Others	4 (downsampled)	9,868	16	20 **(0.2%)**

실험에 앞서 전처리 단계에서는 전체 데이터 셋에 대해 결측치가 있는 데이터는 삭제하였으며 변수별로 정규화를 통해 이상치 탐지가 어떤 한 변수에 크게 의존하지 않도록 했다. 이후 제안한 두 지표(엔트로피, 지니계수)의 객관적인 이상치정도 측정 지표로서의 효용성을 측정하기 위해 아래와 같이 실험을 설계하였다.

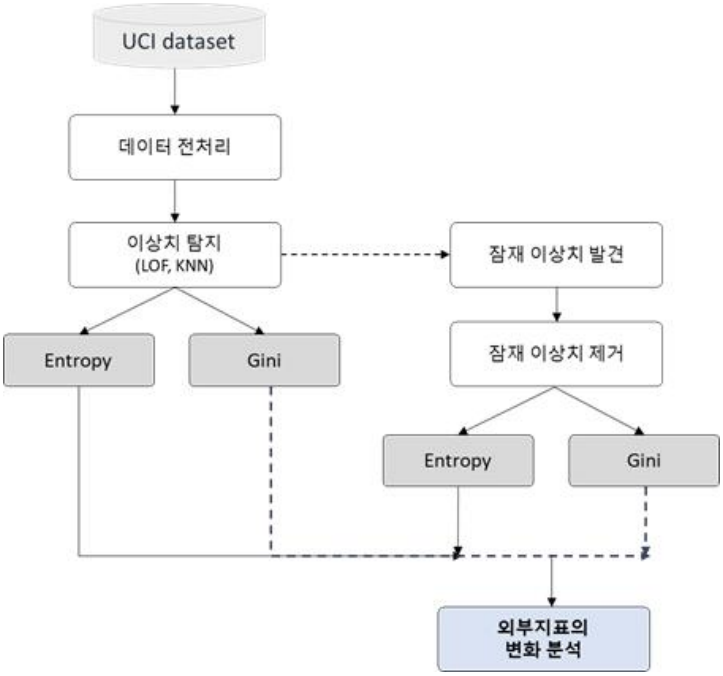

Fig 3.2 UCI 데이터셋을 대상으로 한 실험 개요

먼저 데이터 전처리 후 두 개의 이상치 탐지 모델(LOF, k-NN)을 통해 나온 이상치 점수를 바탕으로 전체 데이터셋의 엔트로피와 지니계수를 기록한다. 그 후 모델을 통해 탐지된 이상치를 일정 부분 제거한 후 다시 전체 데이터셋의 엔트로피와 지니계수를 기록하여 두 점수가 어떤 증감률을 보이는지 관찰하였다. 이러한 과정을 모든 UCI 데이터셋 10개에 적용하여 기록한 결과를 제4장에서 서술한다.

4. 연구 결과

4.1 임의 데이터셋 실험

4.1.1 임의 데이터셋 실험 환경 설정

먼저 제안하는 두 지표를 임의로 생성한 데이터셋을 대상으로 실험하여 그 효용성을 입증하고자 한다. 무작위성이 0인 기본 데이터셋에서 무작위성(randomness)을 0.2단위로 점차적으로 증가시켜 아래와 같이 총 6개의 데이터셋을 구성하였다.

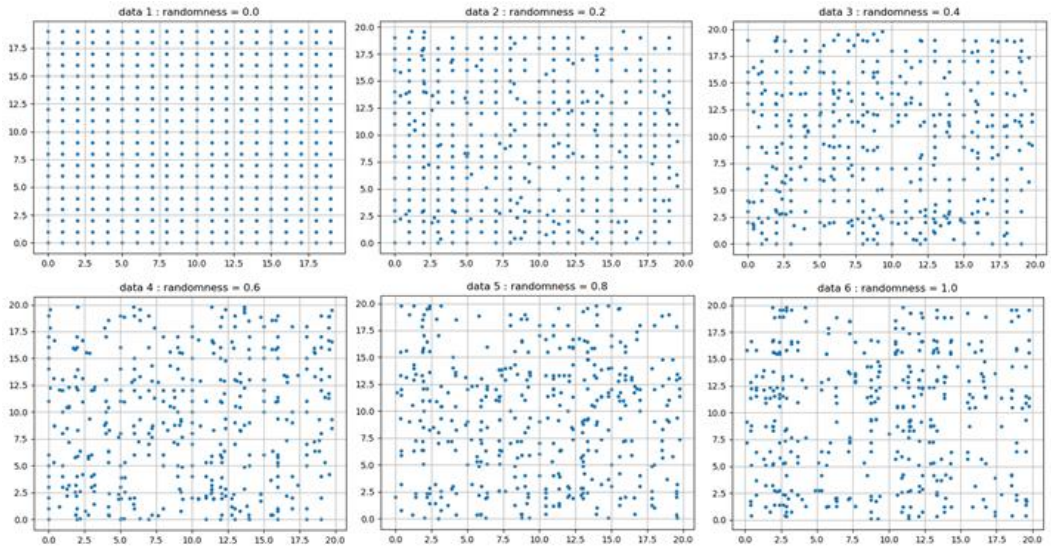

Fig 4.1 무작위성을 증가시킨 실험데이터셋

이렇게 구성한 실험데이터셋을 대상으로 연구의 실험을 진행하기 앞서 먼저 무작위성(Randomneses)이 데이터의 이상정도(Outlierness)를 대표할 수 있는지를 검증하기 위해 무작위성과 이상정도의 대응정도를 측정하고자 한다. 무작위성을 점진적으로 증가시키며 생성한 데이터셋을 바탕으로 각 데이터에 LOF 알고리즘을 적용하여 이상치 탐지를 수행한 결과 무작위성이 증가할수록 LOF 점수가 높음을 확인할 수 있었다. 따라서 무작위성은 데이터의 이상정도를 객관적으로 대표할 수 있는 대응조건임을 확인할 수 있었다.

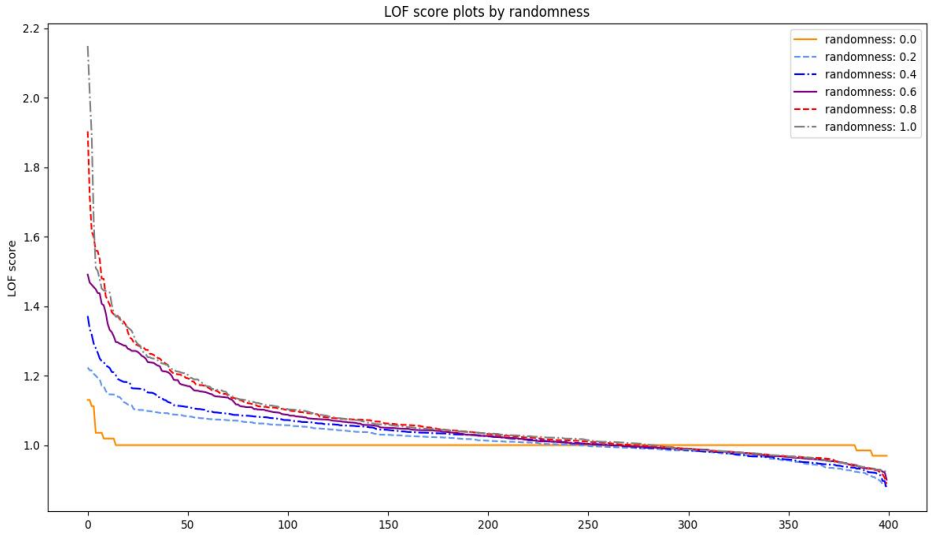

Fig 4.2 무작위성과 이상정도의 대응성을 검증하기 위한 실험결과

데이터의 분포에 영향을 많이 받는 엔트로피의 특성을 반영하기 위해 알고리즘 계산 결과 이상치 점수를 반올림할 자릿수 파라미터 $\epsilon = \{4,2\}$를 도입하고 이에 따라 실험 결과가 어떻게 달라지는지를 관찰하였다. 또, 본 연구에서 사용한 두 개의 이상치 탐지 알고리즘(LOF, k-NN)은 모두 인접이웃기반 이상치 탐지 기법이기 때문에 인접한 이수의 수 k에 따라 상이한 결과를 보여 줄 수 있으므로 $k = \{5, 10\}$ 즉, 이웃이 5개일 때와 10개일 때를 나누어 각각 실험하여 그 변화 추이를 기록하였다. 실험적으로 생성한 데이터셋의 크기가 크지 않기에 이상치 탐지 결과가 매번 상이할 수 있음을 고려하여 동일한 조건 하에 각 10번씩의 실험을 시행한 후 10번의 실험결과의 평균값을 기록과하고 그 분산 역시 아래에 기록하였다.

4.1.2 실험 데이터셋 실험 결과 (1): $\epsilon = 4$, base detector = LOF

표4.1은 LOF를 base detector로 하여 k=5일 때 이상치 탐지 알고리즘을 적용하여 나온 전체 데이터셋의 이상치 점수를 대상으로 엔트로피와 지니계수를 계산한 결과이다. 표4.2는 같은 실험 조건 하에 k=10 일 때의 실험 결과이다.

Table 4.1 실험 데이터셋 실험결과(1): $\epsilon = 4$, LOF, k=5

	rand = 0	rand = 0.2	rand = 0.4	rand = 0.6	rand = 0.8	rand = 1
Entropy	0.3906 ±0	5.8267 ±0.0008	5.8443 ±0.0001	5.8306 ±0.0008	5.8422 ±0.001	5.8297 ±0.0003
Gini	0.0027 ±0	0.0339 ±0	0.0398 ±0	0.0499 ±0.0001	0.0545 ±0	0.0636 ±0.0001

Table 4.2 실험 데이터셋 실험결과(1): $\epsilon = 4$, LOF, k=10

	rand = 0	rand = 0.2	rand = 0.4	rand = 0.6	rand = 0.8	rand = 1
Entropy	3.1687 ±0	5.7697 ±0.0017	5.7872 ±0.0013	5.8053 ±0.0006	5.8134 ±0.0006	5.7985 ±0.0006
Gini	0.0075 ±0	0.0217 ±0	0.0269 ±0	0.0317 ±0	0.0376 ±0	0.041 ±0

그 결과 무작위이 증가할수록 엔트로피와 지니계수 모두 유의미하게 점진적으로 증가함을 확인할 수 있었다. 두 지표 값들이 어떻게 변화하는지를 쉽게 확인할 수 있도록 시각화한 결과는 아래 그림 4.3와 같다.

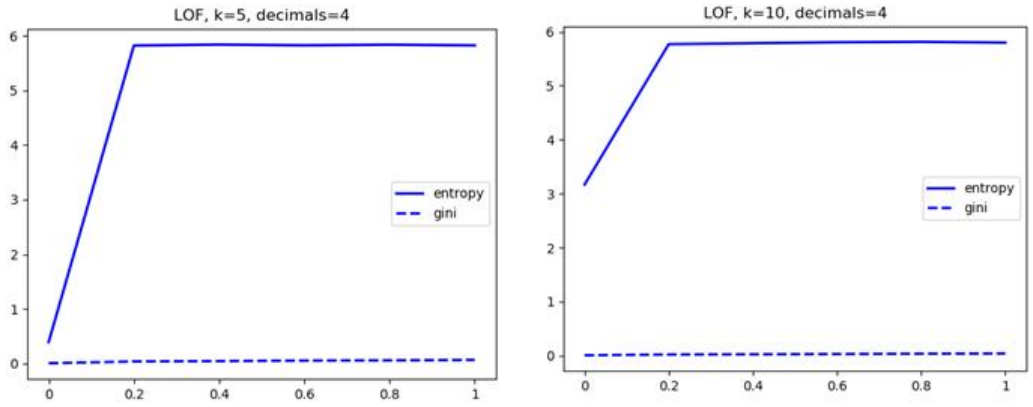

Fig 4.3 실험 데이터셋 실험결과(1)

4.1.3 실험 데이터셋 실험 결과 (2): $\epsilon = 4$, base detector = K-NN

표4.3은 K-NN을 base detector로 하여 k=5일 때 이상치 탐지 알고리즘을 적용하여 나온 전체 데이터셋의 이상치 점수를 대상으로 엔트로피와 지니계수를 계산한 결과이다. 표4.4는 같은 실험 조건 하에 k=10 일 때의 실험 결과이다.

4.3 실험 데이터셋 실험결과(2): $\epsilon = 4$, K-NN, k=5

	rand = 0	rand = 0.2	rand = 0.4	rand = 0.6	rand = 0.8	rand = 1
Entropy	0.056 ±0	2.9987 ±0.0163	4.3071 ±0.002	5.2581 ±0.0062	5.7664 ±0.0013	5.821 ±0.0003
Gini	0.0041 ±0	0.1022 ±0	0.1301 ±0	0.155 ±0.0001	0.169 ±0.0002	0.1915 ±0.0002

4.4 실험 데이터셋 실험결과(2): $\epsilon = 4$, K-NN, k=10

	rand = 0	rand = 0.2	rand = 0.4	rand = 0.6	rand = 0.8	rand = 1
Entropy	0.5254 ±0	3.276 ±0.0183	4.6637 ±0.0104	5.4748 ±0.0072	5.8276 ±0.0084	5.8842 ±0.0023
Gini	0.0214 ±0	0.085 ±0	0.1039 ±0.0001	0.1223 ±0.0001	0.1388 ±0.0001	0.1521 ±0.0003

그 결과 base detector로 K-NN을 사용했을 때도 무작위이 증가할수록 엔트로피와 지니계수 모두 유의미하게 점진적으로 증가함을 확인할 수 있었다. 두 지표 값들이 어떻게 변화하는지를 쉽게 확인할 수 있도록 시각화한 결과는 아래 그림 4.4과 같다.

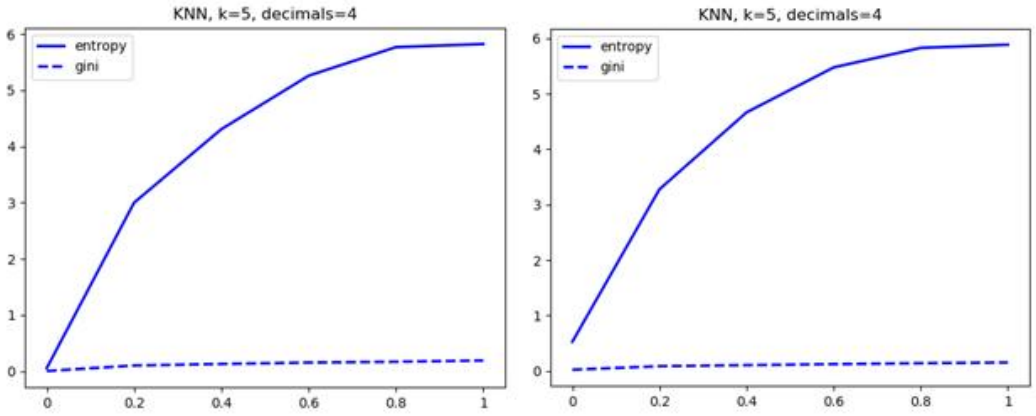

Fig 4.4 실험 데이터셋 실험결과(2)

4.1.4 실험 데이터셋 실험 결과 (3): $\epsilon = 2$, base detector = LOF

표4.5는 LOF를 base detector로 하여 k=5일 때 이상치 탐지 알고리즘을 적용하여 나온 전체 데이터셋의 이상치 점수를 대상으로 엔트로피와 지니계수를 계산한 결과이다. 표4.6은 같은 실험 조건 하에 k=10 일 때의 실험 결과이다.

Table 4.5 실험 데이터셋 실험결과(3): $\epsilon = 2$, LOF, k=5

	rand = 0	rand = 0.2	rand = 0.4	rand = 0.6	rand = 0.8	rand = 1
Entropy	0.3906 ±0	3.1229 ±0.0026	3.2321 ±0.0021	3.3637 ±0.0022	3.4173 ±0.0083	3.5529 ±0.005
Gini	0.0028 ±0	0.0326 ±0	0.0412 ±0	0.0475 ±0	0.0595 ±0	0.0602 ±0

Table 4.6 실험 데이터셋 실험결과(3): $\epsilon = 2$, LOF, k=10

	rand = 0	rand = 0.2	rand = 0.4	rand = 0.6	rand = 0.8	rand = 1
Entropy	1.3183 ±0	2.6675 ±0.0028	2.8727 ±0.0092	2.9772 ±0.0065	3.0747 ±0.0084	3.1539 ±0.0035
Gini	0.0076 ±0	0.0228 ±0	0.0284 ±0	0.0314 ±0	0.0383 ±0	0.0394 ±0

그 결과 무작위이 증가할수록 엔트로피와 지니계수 모두 유의미하게 점진적으로 증가함을 확인할 수 있었다. 두 지표 값들이 어떻게 변화하는지를 쉽게 확인할 수 있도록 시각화한 결과는 아래 그림 4.5와 같다.

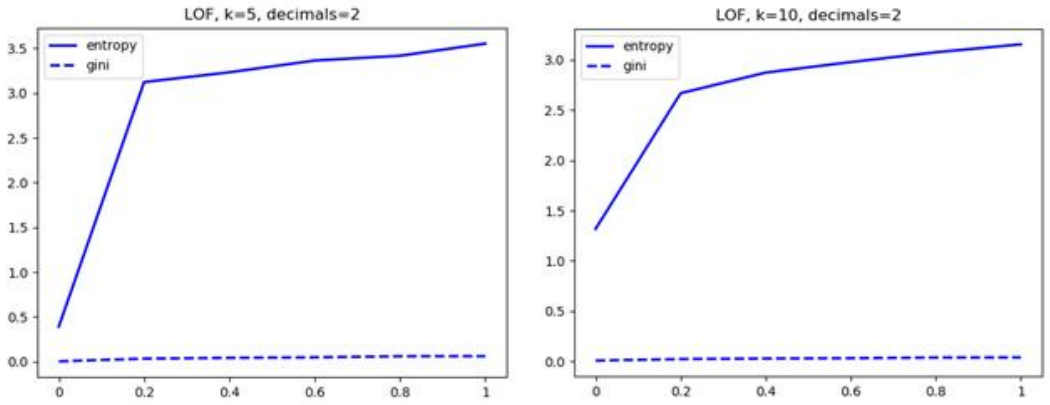

Fig 4.5 실험 데이터셋 실험결과(3)

4.1.5 실험 데이터셋 실험 결과 (4): $\epsilon = 2$, base detector = K-NN

표4.7은 K-NN을 base detector로 하여 k=5일 때 이상치 탐지 알고리즘을 적용하여 나온 전체 데이터셋의 이상치 점수를 대상으로 엔트로피와 지니계수를 계산한 결과이다. 표4.8은 같은 실험 조건 하에 k=10 일 때의 실험 결과이다.

Table 4.7 실험 데이터셋 실험결과(4): $\epsilon = 2$, K-NN, k=5

	rand = 0	rand = 0.2	rand = 0.4	rand = 0.6	rand = 0.8	rand = 1
Entropy	0.056 ±0	2.6779 ±0.0082	3.6983 ±0.0059	4.3395 ±0.004	4.6958 ±0.0014	4.7219 ±0.0035
Gini	0.0041 ±0	0.0978 ±0	0.1293 ±0	0.1466 ±0.0001	0.175 ±0.0002	0.1882 ±0.0001

Table 4.8 실험 데이터셋 실험결과(4): $\epsilon = 2$, K-NN, k=10

	rand = 0	rand = 0.2	rand = 0.4	rand = 0.6	rand = 0.8	rand = 1
Entropy	0.5254 ±0	2.9334 ±0.0059	3.9831 ±0.0082	4.5736 ±0.0019	4.8308 ±0.0036	4.87 ±0.0032
Gini	0.0217 ±0	0.0847 ±0	0.1101 ±0.0001	0.1249 ±0.0001	0.1442 ±0.0001	0.1555 ±0.0002

그 결과 base detector로 K-NN을 사용했을 때도 무작위이 증가할수록 엔트로피와 지니계수 모두 유의미하게 점진적으로 증가함을 확인할 수 있었다. 두 지표 값들이 어떻게 변화하는지를 쉽게 확인할 수 있도록 시각화한 결과는 아래 그림 4.6과 같다.

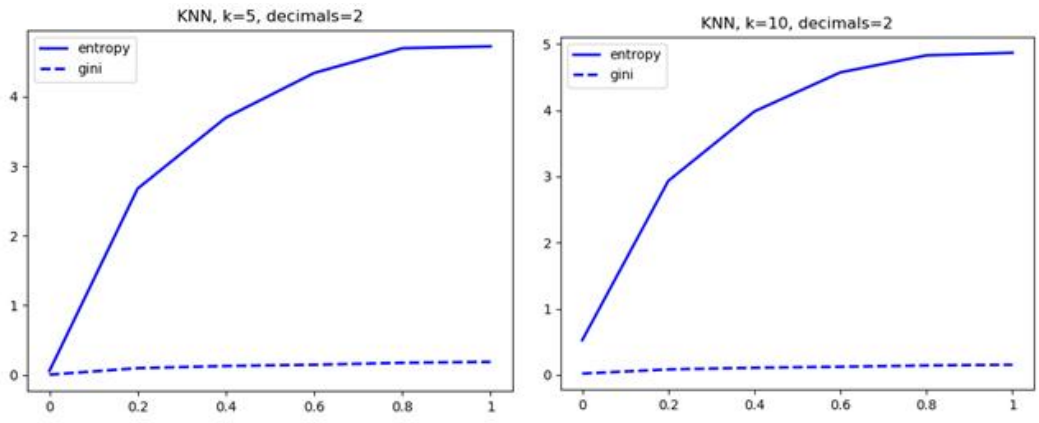

Fig 4.6 실험 데이터셋 실험결과(4)

4.1.6 실험 데이터셋 실험 결과 요약

다양한 환경에서 실험한 결과 base detector의 선택이나 이상치 점수를 반올림하는 파라미터 ϵ의 조건에 상관없이 무작위성이 증가할수록 엔트로피와 지니계수 모두 점진적으로 증가하는 양상을 보임을 확인하였다. 그러나 엔트로피는 무작위성이 증가함에 따라 일정 수준 이상에서는 더딘 증가폭을 보이는 반면, 지니계수는 무작위성이 증가함에 따라 일관적으로 같이 증가하는 양상을 보이고 있기에 지니계수가 객관적인 외부 지표로서의 범용성이 더 높다고 해석할 수 있다. 해당 실험을 통해 임의 데이터셋에서 두 지표의 효용성이 입증되었다고 볼 수 있으므로 UCI mcahine learning repository에서 습득한 10개의 데이터셋에도 동일한 조건 하에 실험하여 지표의 실제 효용성에 대해서 관찰하고자 한다.

4.2 UCI 데이터셋 실험 결과

4.2.1 PageBlock 데이터셋 실험결과

표4.9와 4.10은 PageBlock 데이터셋에 대한 실험 결과이다. 잠재이상치를 제거한 후 두 지표의 증감률을 관측하기 위해서 먼저 데이터셋에 포함된 잠재이상치를 판단할 기준이 필요하다. 이를 위해 원데이터셋을 대상으로 LOF와 K-NN 두 알고리즘을 적용하여 이상치 점수를 내림차순으로 정렬한 결과를 그림4.7과 같이 시각화했으며 그래프의 변곡점을 임계치(cut-off)라고 했을 때, 해당 데이터셋의 이상치 점수가 임계치 이상이면 잠재이상치로 가정했다. 주어진 데이터의 잠재이상치를 제거 했을 때 엔트로피, 지니계수의 변동을 측정한 결과는 다음과 같다.

Table 4.9 PageBlock 엔트로피 실험결과

Entropy	ϵ	original	deleted
LOF (k=10)	4	6.2416	5.8755
	2	1.8527	1.4116
LOF (k=20)	4	6.3394	5.9677
	2	1.9906	1.5426
kNN (k=10)	4	7.6622	7.4427
	2	3.6293	3.3148
kNN (k=20)	4	7.6353	7.4146
	2	3.6314	3.2886

Table 4.10 PageBlock Gini 실험결과

Gini	ϵ	original	deleted
LOF (k=10)	4	0.3445	0.2266
	2	0.3386	0.2182
LOF (k=20)	4	0.3432	0.2355
	2	0.3444	0.2372
kNN (k=10)	4	0.068	0.0461
	2	0.068	0.0461
kNN (k=20)	4	0.0698	0.0454
	2	0.0698	0.0454

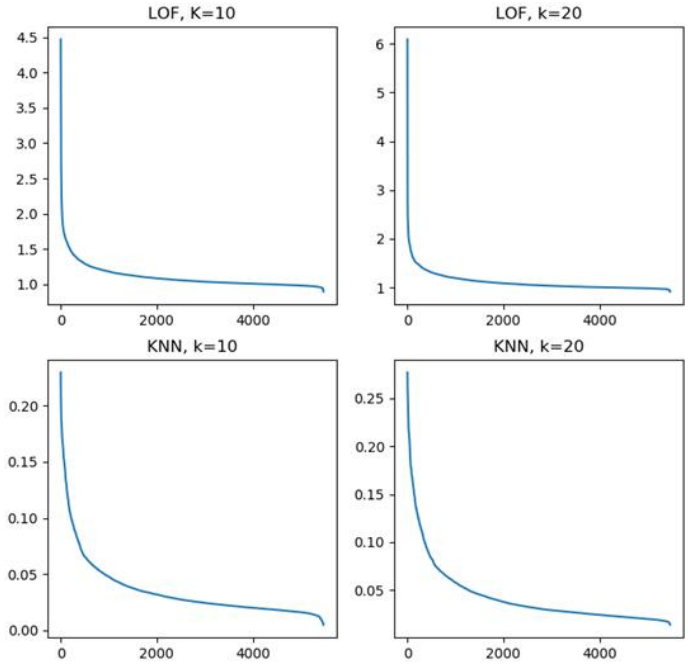

Fig 4.7 PageBlock LOF, KNN 탐지 결과

4.2.2 Cardio 데이터셋 실험결과

표4.11과 4.12는 Cardio 데이터셋에 대한 실험 결과이다. 잠재이상치를 제거한 후 두 지표의 증감률을 관측하기 위해서 먼저 데이터셋에 포함된 잠재이상치를 판단할 기준이 필요하다. 이를 위해 원데이터셋을 대상으로 LOF와 K-NN 두 알고리즘을 적용하여 이상치 점수를 내림차순으로 정렬한 결과를 그림4.8과 같이 시각화했으며 그래프의 변곡점을 임계치(cut-off)라고 했을 때, 해당 데이터셋의 이상치 점수가 임계치 이상이면 잠재이상치로 가정했다. 주어진 데이터의 잠재이상치를 제거 했을 때 엔트로피, 지니계수의 변동을 측정한 결과는 다음과 같다.

Table 4.11 Cardio 엔트로피 실험결과

Entropy	ϵ	original	**deleted**
LOF (k=10)	4	6.2982	6.0296
	2	2.0978	1.7052
LOF (k=20)	4	6.381	6.1282
	2	2.2125	1.8006
kNN (k=10)	4	7.014	6.8227
	2	3.4593	3.0873
kNN (k=20)	4	7.038	6.8503
	2	3.432	3.092

Table 4.12 Cardio Gini 실험결과

Gini	ϵ	original	deleted
LOF (k=10)	4	0.258	0.1861
	2	0.2589	0.1866
LOF (k=20)	4	0.254	0.1778
	2	0.2554	0.1792
kNN (k=10)	4	0.0605	0.0346
	2	0.0606	0.0347
kNN (k=20)	4	0.057	0.035
	2	0.057	0.035

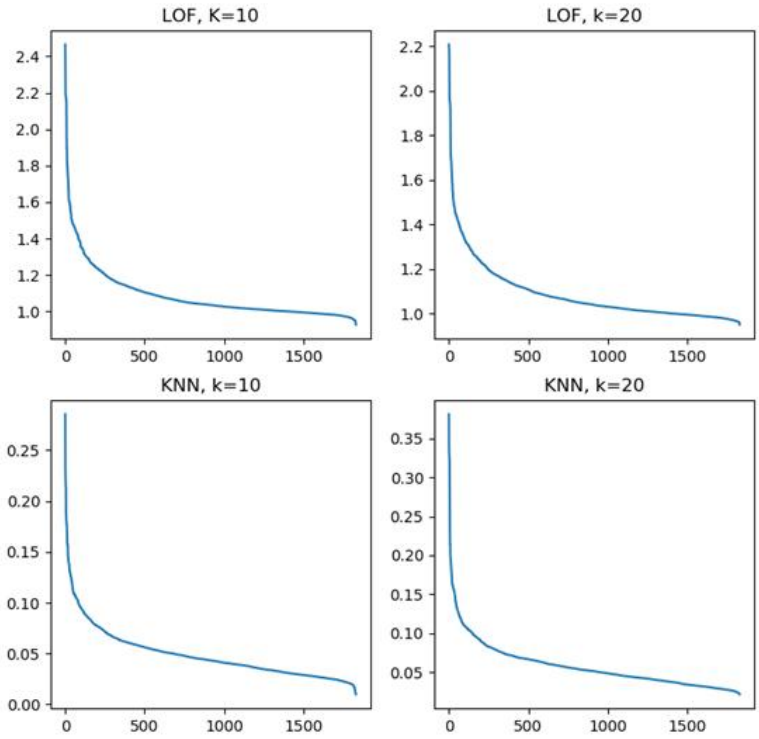

Fig 4.8 Cardio LOF, KNN 탐지 결과

4.2.3 HTRU2 데이터셋 실험결과

표4.13과 4.14는 HTRU2 데이터셋에 대한 실험 결과이다. 잠재이상치를 제거한 후 두 지표의 증감률을 관측하기 위해서 먼저 데이터셋에 포함된 잠재이상치를 판단할 기준이 필요하다. 이를 위해 원데이터셋을 대상으로 LOF와 K-NN 두 알고리즘을 적용하여 이상치 점수를 내림차순으로 정렬한 결과를 그림4.9과 같이 시각화했으며 그래프의 변곡점을 임계치(cut-off)라고 했을 때, 해당 데이터셋의 이상치 점수가 임계치 이상이면 잠재이상치로 가정했다. 주어진 데이터의 잠재이상치를 제거 했을 때 엔트로피, 지니계수의 변동을 측정한 결과는 다음과 같다.

Table 4.13 HTRU2 엔트로피 실험결과

Entropy	ϵ	original	deleted
LOF (k=10)	4	5.8498	5.4515
	2	1.4448	1.041
LOF (k=20)	4	6.0184	5.6249
	2	1.6915	1.2924
kNN (k=10)	4	7.5557	7.3575
	2	3.1291	2.8777
kNN (k=20)	4	7.5367	7.3386
	2	3.1079	2.844

Table 4.14 HTRU2 Gini 실험결과

Gini	ϵ	original	deleted
LOF (k=10)	4	0.4073	0.2604
	2	0.4084	0.2595
LOF (k=20)	4	0.4024	0.2554
	2	0.42	0.2793
kNN (k=10)	4	0.0417	0.0272
	2	0.0417	0.0272
kNN (k=20)	4	0.0435	0.027
	2	0.0435	0.027

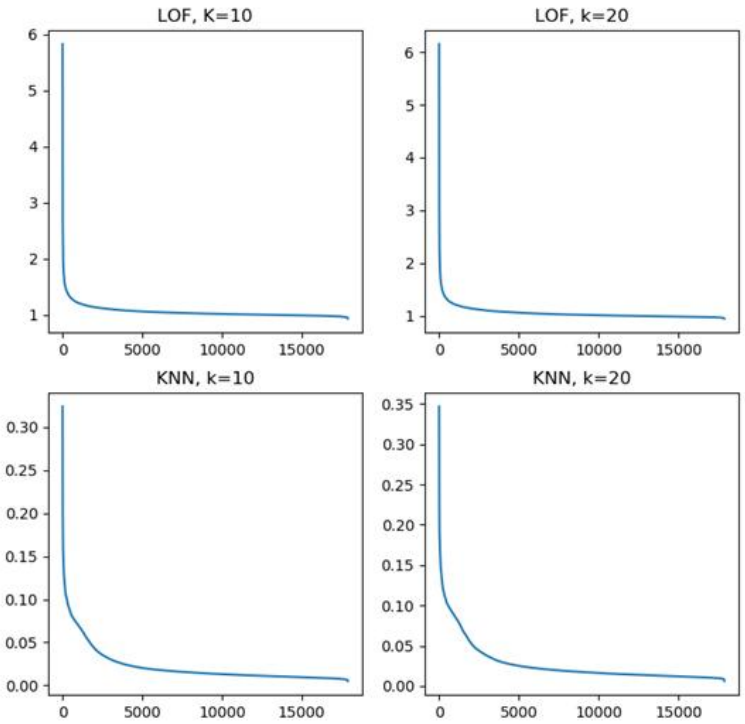

Fig 4.9 HTRU2 LOF, KNN 탐지 결과

4.2.4 Shuttle 데이터셋 실험결과

표4.15와 4.16는 Shuttle 데이터셋에 대한 실험 결과이다. 잠재이상치를 제거한 후 두 지표의 증감률을 관측하기 위해서 먼저 데이터셋에 포함된 잠재이상치를 판단할 기준이 필요하다. 이를 위해 원데이터셋을 대상으로 LOF와 K-NN 두 알고리즘을 적용하여 이상치 점수를 내림차순으로 정렬한 결과를 그림4.10과 같이 시각화했으며 그래프의 변곡점을 임계치(cut-off)라고 했을 때, 해당 데이터셋의 이상치 점수가 임계치 이상이면 잠재이상치로 가정했다. 주어진 데이터의 잠재이상치를 제거 했을 때 엔트로피, 지니계수의 변동을 측정한 결과는 다음과 같다.

Table 4.15 shuttle 엔트로피 실험결과

Entropy	ϵ	original	deleted
LOF (k=10)	4	5.8887	5.647
	2	1.5535	1.2846
LOF (k=20)	4	6.0774	5.8335
	2	1.6281	1.3308
kNN (k=10)	4	7.81	7.621
	2	3.5393	3.2618
kNN (k=20)	4	7.7556	7.5648
	2	3.4892	3.1926

Table 4.16 shuttle Gini 실험결과

Gini	ϵ	original	deleted
LOF (k=10)	4	0.3473	0.2103
	2	0.352	0.2175
LOF (k=20)	4	0.3571	0.207
	2	0.351	0.1983
kNN (k=10)	4	0.0771	0.0495
	2	0.0772	0.0494
kNN (k=20)	4	0.0878	0.0548
	2	0.0878	0.0548

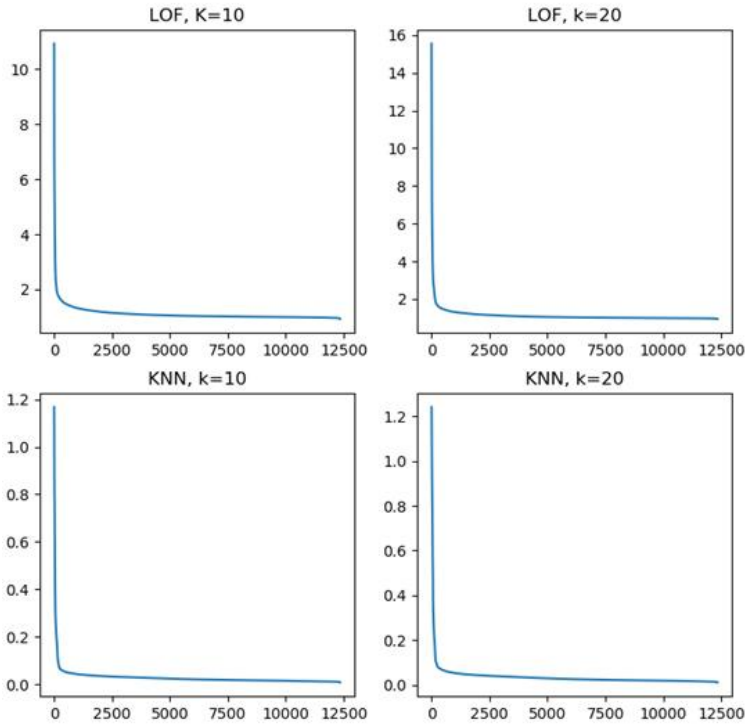

Fig 4.10 Shuttle LOF, KNN 탐지 결과

4.2.5 wilt 데이터셋 실험결과

표4.17와 4.18은 wilt 데이터셋에 대한 실험 결과이다. 잠재이상치를 제거한 후 두 지표의 증감률을 관측하기 위해서 먼저 데이터셋에 포함된 잠재이상치를 판단할 기준이 필요하다. 이를 위해 원데이터셋을 대상으로 LOF와 K-NN 두 알고리즘을 적용하여 이상치 점수를 내림차순으로 정렬한 결과를 그림4.11과 같이 시각화했으며 그래프의 변곡점을 임계치(cut-off)라고 했을 때, 해당 데이터셋의 이상치 점수가 임계치 이상이면 잠재이상치로 가정했다. 주어진 데이터의 잠재이상치를 제거 했을 때 엔트로피, 지니계수의 변동을 측정한 결과는 다음과 같다.

Table 4.17 wilt 엔트로피 실험결과

Entropy	ϵ	original	deleted
LOF (k=10)	4	5.6814	5.4189
	2	1.2849	0.9968
LOF (k=20)	4	5.8366	5.5798
	2	1.5347	1.2363
kNN (k=10)	4	7.4769	7.352
	2	3.3938	3.1451
kNN (k=20)	4	7.4592	7.3189
	2	3.3589	3.1168

Table 4.18 wilt Gini 실험결과

Gini	ϵ	original	deleted
LOF (k=10)	4	0.3595	0.2713
	2	0.3463	0.253
LOF (k=20)	4	0.3614	0.2632
	2	0.3735	0.2779
kNN (k=10)	4	0.0537	0.036
	2	0.0537	0.036
kNN (k=20)	4	0.0535	0.0361
	2	0.0535	0.0361

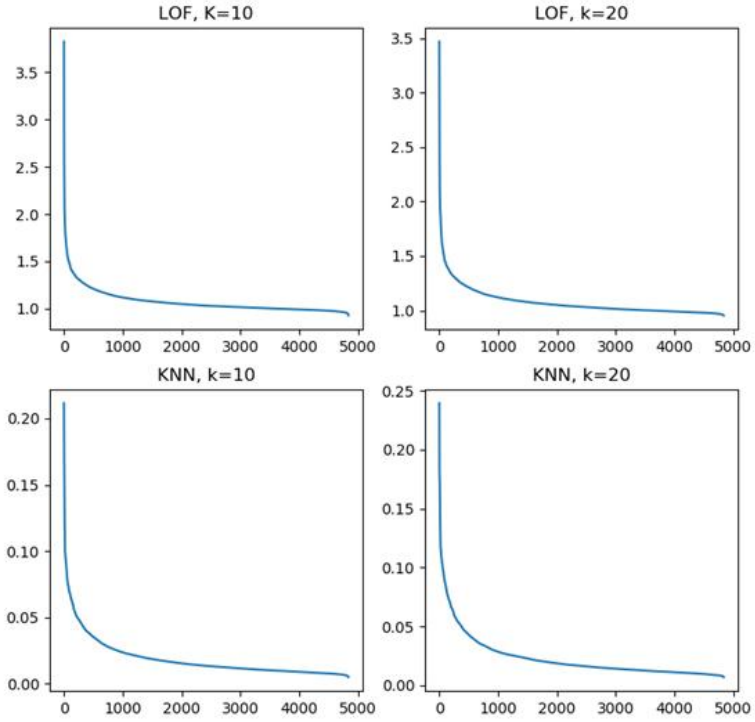

Fig 4.11 wilt LOF, KNN 탐지 결과

4.2.6 Glass 데이터셋 실험결과

표4.19과 4.20은 Glass 데이터셋에 대한 실험 결과이다. 잠재이상치를 제거한 후 두 지표의 증감률을 관측하기 위해서 먼저 데이터셋에 포함된 잠재이상치를 판단할 기준이 필요하다. 이를 위해 원데이터셋을 대상으로 LOF와 K-NN 두 알고리즘을 적용하여 이상치 점수를 내림차순으로 정렬한 결과를 그림4.12와 같이 시각화했으며 그래프의 변곡점을 임계치(cut-off)라고 했을 때, 해당 데이터셋의 이상치 점수가 임계치 이상이면 잠재이상치로 가정했다. 주어진 데이터의 잠재이상치를 제거 했을 때 엔트로피, 지니계수의 변동을 측정한 결과는 다음과 같다.

Table 4.19 Glass 엔트로피 실험결과

Entropy	ϵ	original	deleted
LOF (k=10)	4	5.2818	5.1522
	2	3.3651	3.04
LOF (k=20)	4	5.2623	5.1303
	2	3.5553	3.2263
kNN (k=10)	4	5.2988	5.1786
	2	3.9647	3.7233
kNN (k=20)	4	5.3206	5.2178
	2	3.9549	3.7156

Table 4.20 Glass Gini 실험결과

Gini	ϵ	original	deleted
LOF (k=10)	4	0.438	0.3297
	2	0.4367	0.328
LOF (k=20)	4	0.4164	0.323
	2	0.4164	0.3231
kNN (k=10)	4	0.1628	0.0895
	2	0.1629	0.0897
kNN (k=20)	4	0.1553	0.094
	2	0.1554	0.094

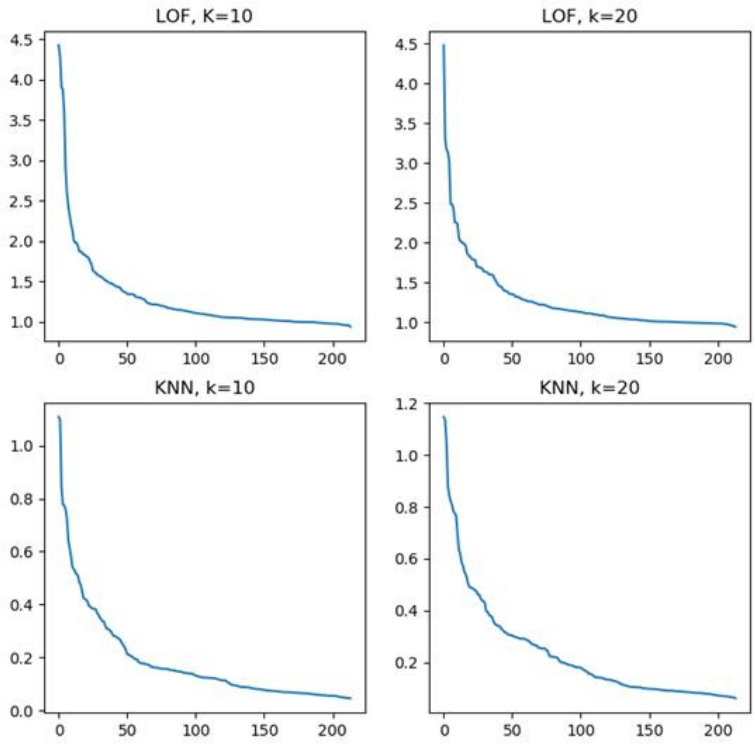

Fig 4.12 Glass LOF, KNN 탐지 결과

4.2.7 Waveform 데이터셋 실험결과

표4.20와 4.21은 Waveform 데이터셋에 대한 실험 결과이다. 잠재이상치를 제거한 후 두 지표의 증감률을 관측하기 위해서 먼저 데이터셋에 포함된 잠재이상치를 판단할 기준이 필요하다. 이를 위해 원데이터셋을 대상으로 LOF와 K-NN 두 알고리즘을 적용하여 이상치 점수를 내림차순으로 정렬한 결과를 그림4.13과 같이 시각화했으며 그래프의 변곡점을 임계치(cut-off)라고 했을 때, 해당 데이터셋의 이상치 점수가 임계치 이상이면 잠재이상치로 가정했다. 주어진 데이터의 잠재이상치를 제거 했을 때 엔트로피, 지니계수의 변동을 측정한 결과는 다음과 같다.

Table 4.20 Waveform 엔트로피 실험결과

Entropy	ϵ	original	deleted
LOF (k=10)	4	7.1734	7.0454
	2	2.9625	2.7637
LOF (k=20)	4	7.1878	7.0568
	2	2.9803	2.7844
kNN (k=10)	4	7.232	7.1082
	2	3.102	2.8816
kNN (k=20)	4	7.1912	7.0672
	2	3.0205	2.8027

Table 4.21 Waveform Gini 실험결과

Gini	ϵ	original	deleted
LOF (k=10)	4	0.0552	0.0468
	2	0.0552	0.0467
LOF (k=20)	4	0.0535	0.0453
	2	0.0535	0.0453
kNN (k=10)	4	0.0329	0.0258
	2	0.0328	0.0258
kNN (k=20)	4	0.0309	0.0242
	2	0.0309	0.0242

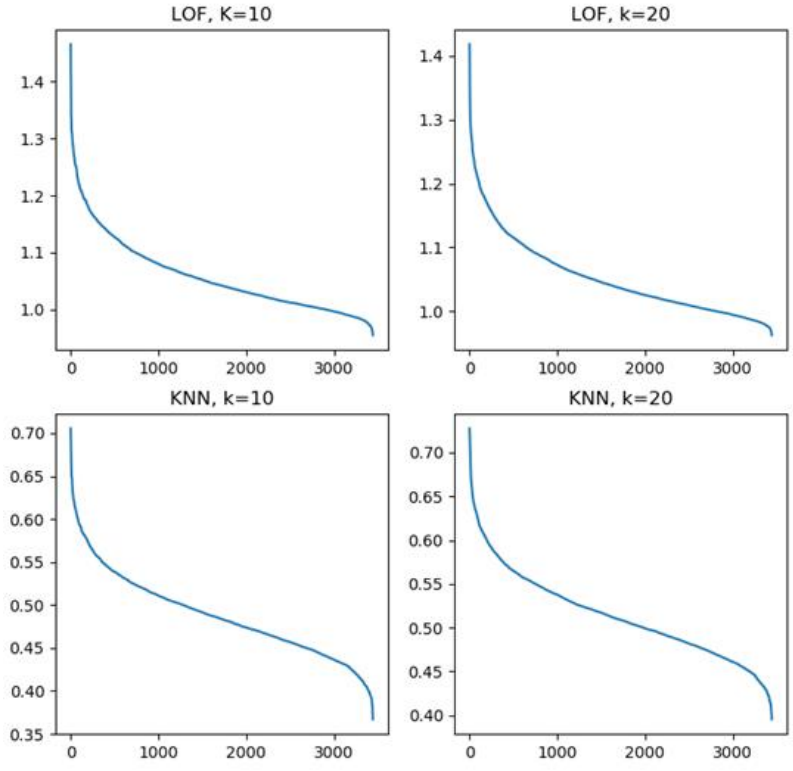

Fig 4.13 Waveform LOF, KNN 탐지 결과

4.2.8 WDBC 데이터셋 실험결과

표4.23과 4.24는 WDBC 데이터셋에 대한 실험 결과이다. 잠재이상치를 제거한 후 두 지표의 증감률을 관측하기 위해서 먼저 데이터셋에 포함된 잠재이상치를 판단할 기준이 필요하다. 이를 위해 원데이터셋을 대상으로 LOF와 K-NN 두 알고리즘을 적용하여 이상치 점수를 내림차순으로 정렬한 결과를 그림4.14과 같이 시각화했으며 그래프의 변곡점을 임계치(cut-off)라고 했을 때, 해당 데이터셋의 이상치 점수가 임계치 이상이면 잠재이상치로 가정했다. 주어진 데이터의 잠재이상치를 제거 했을 때 엔트로피, 지니계수의 변동을 측정한 결과는 다음과 같다.

Table 4.23 WDBC 엔트로피 실험결과

Entropy	ϵ	original	deleted
LOF (k=10)	4	5.8487	5.6856
	2	3.9119	3.5063
LOF (k=20)	4	5.8156	5.6471
	2	3.9473	3.557
kNN (k=10)	4	5.8081	5.634
	2	3.6374	3.3138
kNN (k=20)	4	5.8171	5.6883
	2	3.6754	3.3156

Table 4.24 WDBC Gini 실험결과

Gini	ϵ	original	deleted
LOF (k=10)	4	0.1959	0.1129
	2	0.1959	0.1128
LOF (k=20)	4	0.19	0.1088
	2	0.1899	0.1088
kNN (k=10)	4	0.0684	0.0471
	2	0.0686	0.0472
kNN (k=20)	4	0.0795	0.042
	2	0.0794	0.0419

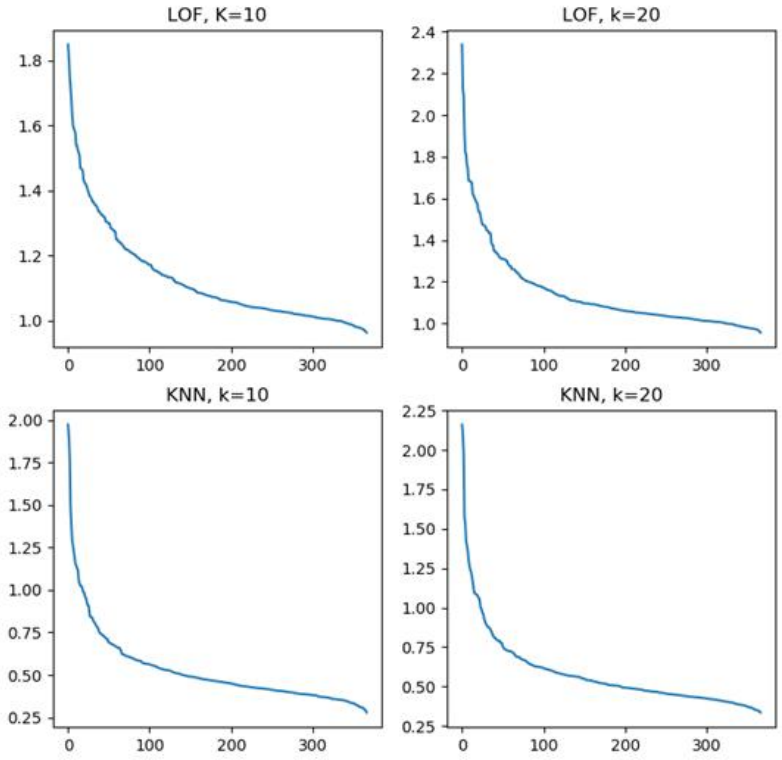

Fig 4.14 WDBC LOF, KNN 탐지 결과

4.2.9 Annthyroid 데이터셋 실험결과

표4.25과 4.26는 Annthyroid 데이터셋에 대한 실험 결과이다. 잠재이상치를 제거한 후 두 지표의 증감률을 관측하기 위해서 먼저 데이터셋에 포함된 잠재이상치를 판단할 기준이 필요하다. 이를 위해 원데이터셋을 대상으로 LOF와 K-NN 두 알고리즘을 적용하여 이상치 점수를 내림차순으로 정렬한 결과를 그림4.15와 같이 시각화했으며 그래프의 변곡점을 임계치(cut-off)라고 했을 때, 해당 데이터셋의 이상치 점수가 임계치 이상이면 잠재이상치로 가정했다. 주어진 데이터의 잠재이상치를 제거 했을 때 엔트로피, 지니계수의 변동을 측정한 결과는 다음과 같다.

Table 4.25 Annthyroid 엔트로피 실험결과 Table 4.26 Annthyroid Gini 실험결과

Entropy	ϵ	original	deleted
LOF (k=10)	4	6.6907	6.5413
	2	2.5627	2.311
LOF (k=20)	4	6.8944	6.7579
	2	2.8389	2.6376
kNN (k=10)	4	7.5979	7.5282
	2	4.0214	3.8005
kNN (k=20)	4	7.5891	7.5037
	2	4.0725	3.8333

Gini	ϵ	original	deleted
LOF (k=10)	4	0.645	0.4639
	2	0.6457	0.4652
LOF (k=20)	4	0.645	0.55
	2	0.6452	0.5502
kNN (k=10)	4	0.2613	0.1575
	2	0.2613	0.1575
kNN (k=20)	4	0.3059	0.1198
	2	0.3059	0.1199

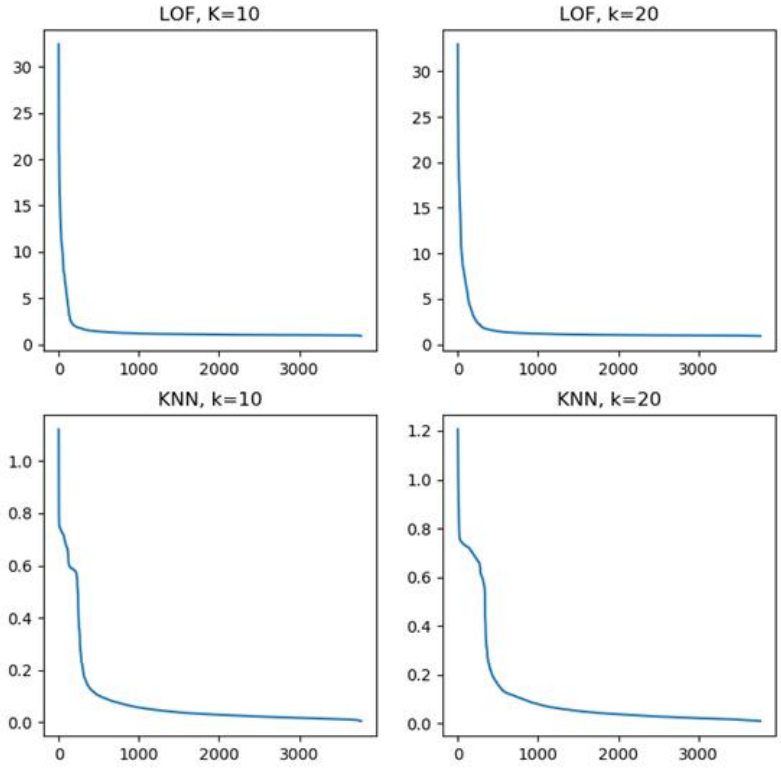

Fig 4.15 Annthyroid LOF, KNN 탐지 결과

4.2.10 PenDigits 데이터셋 실험결과

표4.27과 4.28은 PenDigits 데이터셋에 대한 실험 결과이다. 잠재이상치를 제거한 후 두 지표의 증감률을 관측하기 위해서 먼저 데이터셋에 포함된 잠재이상치를 판단할 기준이 필요하다. 이를 위해 원데이터셋을 대상으로 LOF와 K-NN 두 알고리즘을 적용하여 이상치 점수를 내림차순으로 정렬한 결과를 그림4.16와 같이 시각화했으며 그래프의 변곡점을 임계치(cut-off)라고 했을 때, 해당 데이터셋의 이상치 점수가 임계치 이상이면 잠재이상치로 가정했다. 주어진 데이터의 잠재이상치를 제거 했을 때 엔트로피, 지니계수의 변동을 측정한 결과는 다음과 같다.

Table 4.27 WBC 엔트로피 실험결과

Entropy	ϵ	original	deleted
LOF (k=10)	4	7.4072	7.3283
	2	7.4059	7.3279
LOF (k=20)	4	7.6345	7.5608
	2	7.6297	7.5549
kNN (k=10)	4	7.7184	7.651
	2	3.4286	3.3013
kNN (k=20)	4	7.677	7.6069
	2	3.3884	3.2646

Table 4.28 WBC Gini 실험결과

Gini	ϵ	original	deleted
LOF (k=10)	4	0.1961	0.1768
	2	0.1961	0.1768
LOF (k=20)	4	0.1985	0.1817
	2	0.1985	0.1817
kNN (k=10)	4	0.0539	0.0429
	2	0.0539	0.0429
kNN (k=20)	4	0.0537	0.0434
	2	0.0538	0.0434

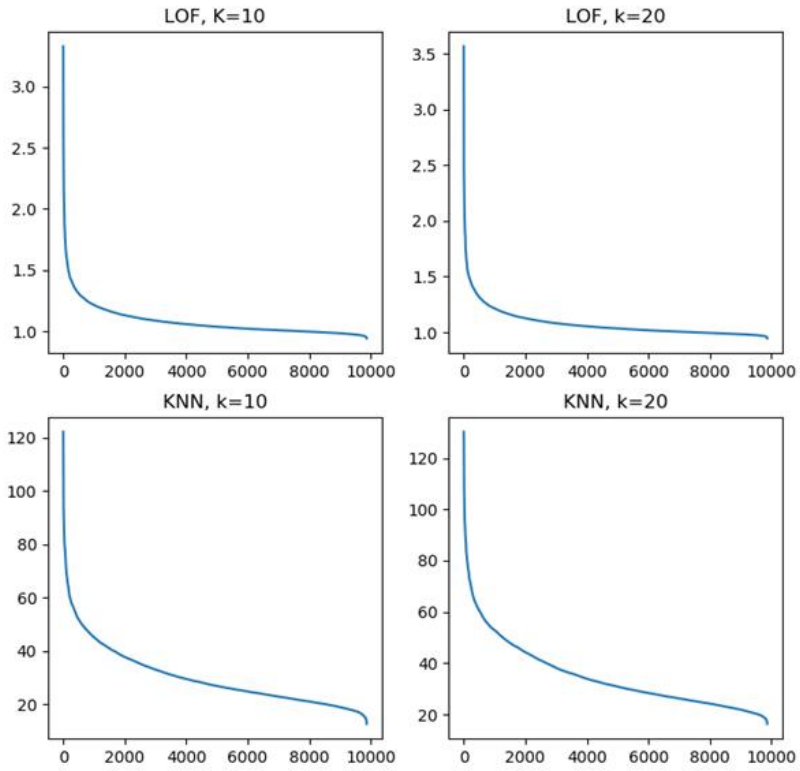

Fig 4.16 PenDigits LOF, KNN 탐지 결과

4.2.11 UCI 10개 데이터셋 실험 결과 요약

최근 이상치 탐지 분야에서 가장 빈번히 쓰인 10개의 벤치마크 데이터셋을 대상으로 실험한 결과 이상치를 제거하고 엔트로피와 지니계수를 측정했을 때 모두 유의미하게 지표가 감소함을 확인했다. 이를 바탕으로 두 지표 모두 데이터셋 내에 존재하는 이상정도를 측정하는 객관적인 지표로서 활용될 수 있음을 확인하였다.

표 4.29는 데이터셋 마다 설정한 임계치에 따른 잠재 이상치를 제거한 후 재측정했을 때 지니계수의 감소율(-△%)과 삭제된 잠재이상치가 얼마만큼의 이상정도(outlierness)를 가지고 있었는지, 즉 이상치 하나가 평균적으로 가지고 있었던 이상정도를 기록한 표이다.

Table 4.29 ε=4, KNN(k=10) 일 때 데이터셋 지니계수의 감소율과 이상정도

no.	Dataset	cut-off (a)	Gini			
			original	deleted	-△% (b)	outlierness multiplier (b/a)
1	PageBlock	12.8%	0.068	0.0461	32.2%	**2.52**
2	Cardio	13.7%	0.0605	0.0346	42.8%	**3.12**
3	HTRU2	10.0%	0.0417	0.0272	34.8%	**3.48**
4	Shuttle	8.1%	0.0771	0.0495	35.8%	**4.42**
5	Wilt	8.0%	0.0537	0.036	33.0%	**4.13**
6	Glass	11.7%	0.1628	0.0895	45.0%	**3.85**
7	Waveform	7.3%	0.0329	0.0258	21.6%	**2.96**
8	WDBC	13.6%	0.0684	0.0471	31.1%	**2.29**
9	Annthyroid	5.3%	0.2613	0.1575	39.7%	**7.49**
10	PenDigits	2.5%	0.0539	0.0429	20.4%	**8.16**

이렇게 나온 결과를 바탕으로 열 개의 데이터셋을 그림4.17과 같이 이차원 공간에 나타내보았고 이는 다음과 같은 해석이 가능하다. 제 2사분면에 위치하는 두 개의 데이터셋 PenDigits, Annthyroid에서는 삭제한 잠재이상치의 수가 상대적으로 적지만 해당 이상치들이 나타내는 이상정도는 높은 경향을 보이고 있다. 즉, 소수의 이상치가 높은 이상정도를 가지고 있었다고 해석할 수 있다. 이와 반대로, 제 4사분면에 위치하는 Shuttle, HTRU2, Glass, Cardio, PageBlock, WDBC 여섯 개의 데이터셋의 경우에는 다수의 이상치가 비교적 낮은 이상정도를 가지고 있었음을 파악할 수 있다. 각 사분면에 위치한 데이터셋과 그 특징을 표4.30에 정리하였다.

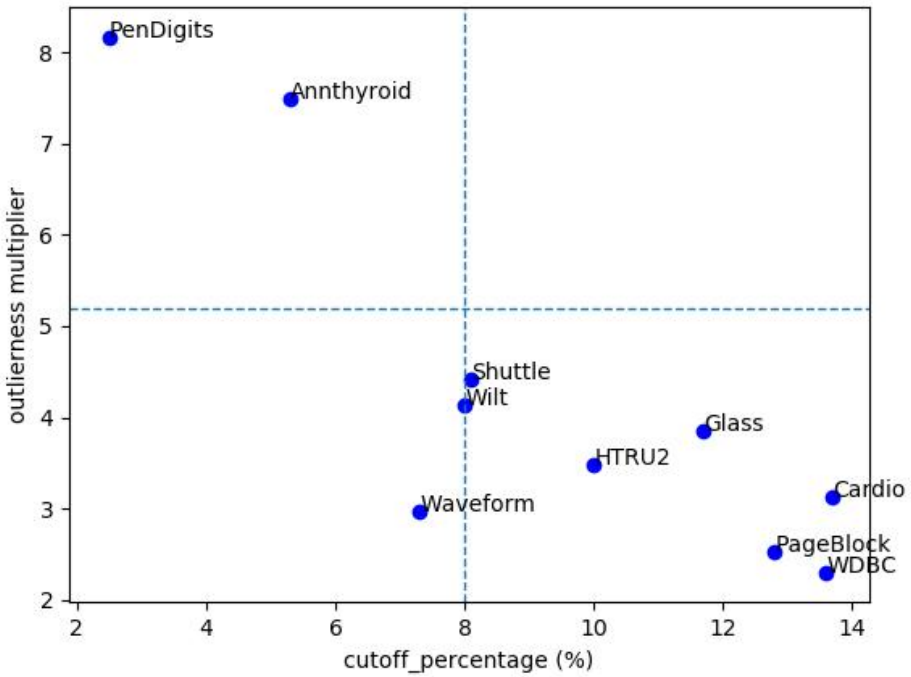

Fig 4.17 이상정도에 따라 분류한 10개의 데이터셋

Table 4.30 데이터셋의 이상치의 수와 이상치가 가진 이상정도

	데이터셋	설명
제2사분면	**Annthyroid, PenDigits**	**적은 수의 이상치가 높은 이상정도를 가짐**
제3사분면	Waveform, wilt	적은 수의 이상치가 적은 이상정도를 가짐
제4사분면	**Shuttle, HTRU2, Glass, Cardio, PageBlock, WDBC**	**많은 수의 이상치가 낮은 이상정도를 가짐**

5. 결론

5.1 연구의 결론

본 논문에서는 이상치 탐지에 앞서 데이터가 실제로 이상치를 포함하고 있는지를 알아볼 수 있는 객관적인 이상정도(outlierness)를 측정하는 지표를 제안하였다. 이 지표로는 기존 데이터 마이닝 분야에서 불순도 지표로 자주 쓰이던 엔트로피와 지니계수를 활용하였고 먼저 지표의 타당성을 확인하기 위해 실험 데이터셋을 만들어 무작위성이 증가할 때마다 두 불순도 지표가 유의미하게 증가함을 확인하였다. 이론적으로 두 지표가 데이터의 이상치 정도를 측정할 수 있음을 확인한 후 실제 데이터에 적용하여 지표의 효용성을 확인하고자 UCI machine learning repository에서 제공하는 열 개의 데이터셋을 대상으로 추가 실험을 진행하였다. 먼저 두 개의 이상치 탐지 알고리즘(LOF, K-NN)을 원데이터에 적용하여 나온 이상치 점수를 바탕으로 전체 데이터셋의 엔트로피와 지니계수를 기록했다. 그 후 각 모델을 통해 탐지된 이상치점수를 기준으로 내림차순 정렬한 후 상위 일정 부분의 잠재 이상치를 제거한 후 다시 전체 데이터셋의 엔트로피와 지니계수를 기록하였다. 그 결과 모든 데이터셋에서 잠재 이상치를 제거했을 때 엔트로피와 지니계수가 모두 낮아짐을 확인할 수 있었다. 이러한 결과를 바탕으로 엔트로피와 지니계수는 데이터의 이상정도를 확인할 수 있는 지표임을 확인했으며 추후 이상치 탐지 분야에서 데이터의 이상 정도를 먼저 알아볼 수 있는 방법으로 활용 될 수 있을 것으로 보인다.

5.2 연구의 한계점

본 연구에서는 비지도학습 이상치 탐지에서 데이터가 포함하고 있는 이상정도를 객관적으로 측정할 수 있는 지표를 제안했음에도 불구하고 다음과 같은 한계점을 가지고 있다. 먼저, 이상치 탐지 분야에 있어 가장 널리 사용되는 UCI repository의 10개의 데이터셋에 대해 실험을 진행하고 일관적인 결과를 확인했지만, 더 많은 데이터셋을 대상으로 하는 추가 실험을 통해 지표의 일관성을 더욱 공고히 할 필요가 있다. 또, 본 논문에서는 임계치를 이상치 점수를 내림차순 정렬하여 그래프를 그렸을 때 변곡점으로 설정했지만 이에 대한 수학적 논거가 미흡하기 때문에 추후 연구에서 이러한 점을 보완한다면 외부지표로서의 타당성을 더욱 높일 수 있을 것으로 기대된다.

참고문헌

김승, 조남욱, & 강석호. (2010). 대용량 자료 분석을 위한 밀도기반 이상치 탐지. 한국경영과학회지, 35(2), 71-88.

Aggarwal, C. C. (2013). Outlier ensembles: position paper. ACM SIGKDD Explorations Newsletter, 14(2), 49-58.

Barnet, V., & Lewis, T. Outliers in statistical data. 1994.

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000, May). LOF: identifying density-based local outliers. In ACM sigmod record (Vol. 29, No. 2, pp. 93-104). ACM.

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(4), 891-927.

Goldstein, M., & Uchida, S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. PloS one, 11(4), e0152173.

Hawkins, D. M. (1980). Identification of outliers (Vol. 11). London: Chapman and Hall.

He, Z., Deng, S., & Xu, X. (2005, August). An optimization model for outlier detection in categorical data. In International Conference on Intelligent Computing (pp. 400-409). Springer, Berlin, Heidelberg.

Jiang, F., Sui, Y., & Cao, C. (2010). An information entropy-based approach to outlier detection in rough sets. Expert Systems with Applications, 37(9), 6338-6344.

Koufakou, A., Ortiz, E. G., Georgiopoulos, M., Anagnostopoulos, G. C., & Reynolds, K. M. (2007, October). A scalable and efficient outlier detection strategy for categorical data. In Tools with Artificial Intelligence, 2007. ICTAI 2007. 19th IEEE International Conference on (Vol. 2, pp. 210-217). IEEE.

Kriegel, H. P., Kröger, P., Schubert, E., & Zimek, A. (2009, November). LoOP: local outlier probabilities. In Proceedings of the 18th ACM conference on Information and knowledge management (pp. 1649-1652). ACM.

Liu, X., Lu, C. T., & Chen, F. (2008, July). An entropy-based method for assessing the number of spatial outliers. In Information Reuse and Integration, 2008. IRI 2008. IEEE International Conference on (pp. 244-249). IEEE.

Papadimitriou, S., Kitagawa, H., Gibbons, P. B., & Faloutsos, C. (2003, March). Loci: Fast outlier detection using the local correlation integral. In Data Engineering, 2003. Proceedings. 19th International Conference on (pp. 315-326). IEEE.

Pimentel, M. A., Clifton, D. A., Clifton, L., & Tarassenko, L. (2014). A review of novelty detection. Signal Processing, 99, 215-249.

Shannon, C. E. (1948). A note on the concept of entropy. Bell System Tech. J, 27(3), 379-423.

Tang, B., & He, H. (2017). A local density-based approach for outlier detection. Neurocomputing, 241, 171-180.

Toshniwal, D., & Eshwar, B. K. (2014). Entropy Based Adaptive Outlier Detection Technique for Data Streams. In Proceedings of the International Conference on Data Mining (DMIN) (p. 1). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

Zimek, A., Campello, R. J., & Sander, J. (2014). Ensembles for unsupervised outlier detection: challenges and research questions a position paper. Acm Sigkdd Explorations Newsletter, 15(1), 11-22.

Abstract

A study on an Outlierness metric for Unsupervised Outlier Detection

La, Sunmin
(Supervisor Cho, Nam Wook)
Dept. of Data Science
Graduate School
Seoul National University of Science and Technology

Outlier detection is a data analysis method based on data mining technique to find an outlier in given dataset. Hawkins (Hawkins, 1980) defines an outlier *as an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism*. Outlier detection has been widely used in industry as the amount of data rises rapidly. Previous research focused on applying eminent outlier detection algorithms to multiple datasets and detecting outliers subjectively. Consequently the evaluation of the models also depends on researcher as well(Aggarawal, 2013). In this regard, the need of external measure has been highlighted over time but it has not been addressed yet. This study proposes two objective Outlierness metrics: Entropy and Gini index. To examine the effectiveness of proposed metrics, experiments have been conducted on both artificial and real-world datasets. As a result, both metrics are proved as legit external metrics that can be used in unsupervised outlier detection. The result also showed both metrics can quantify given dataset's outlierness.