BIG DATA ANALYTICS

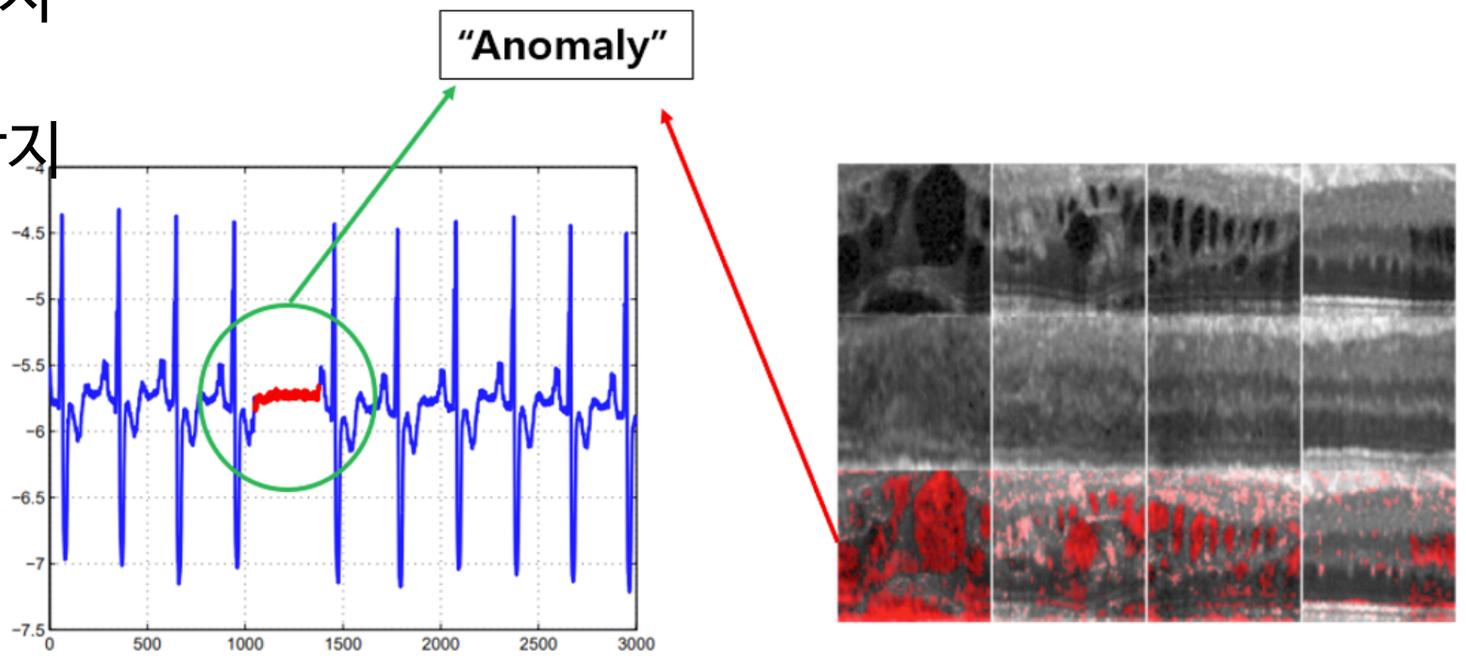
WEEK-11 | Anomaly Detection

Yonsei University
Jungwon Seo

이상탐지의 중요성

- Anomalies/Outliers란?
  - 나머지 데이터와 상당히 다른 데이터 포인트 세트

- 응용 사례
  - 신용 카드 사기 탐지
  - 통신 사기 탐지
  - 네트워크 침입 탐지, 결함 탐지
  - Video Surveillance
  - 제조업 공정과정에서 이상탐지



Reference
[1] Anomaly Detection of Time Series, 2010
[2] Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery, 2017

이상탐지의 중요성

* 구정모, "국내 자동차리콜 연 218만대…"사고 4300건 차량결함 추정", 연합뉴스, 2020. 02. 21

* 안정락, "[갤노트7 '전격 리콜'] 빠르게…통 크게…삼성, 불량률 0.0024%에도 "모두 바꿔주겠다", 한국경제, 2016. 09. 03

이상탐지의 중요성

연합뉴스 | 최신기사 정치 북한 경제 산업 사회 전국 세계 문화 라이프 연예 스포츠 오피니언

#코로나19 #사랑제일교회 #코로나 #확진 #확진자

뉴스홈 | 최신기사

미국 토네이도 재산피해만 8년 연속 100억 달러

송고시간 | 2016-01-05 02:04

장현구 기자

| 독일 재보험회사 "전 세계 재해 피해액은 6년 사이 최저"

독일 재보험회사 "전 세계 재해 피해액은 6년 사이 최저"

2004 Indian Ocean earthquake and tsunami (2004년 인도양 지진해일)

December 26, 2004

The 2004 Indian Ocean earthquake and tsunami occurred at 07:58:53 in local time on 26 December, with an epicentre off the west coast of northern Sumatra, Indonesia. It was an undersea megathrust earthquake that registered a magnitude of 9.1–9.3 Mw, reaching a Mercalli intensity up to IX in certain areas. Wikipedia

Depth: 30,000 m

Date: December 26, 2004

Number of deaths: 227,898

Location: Banda Aceh, Indonesia

Total damage: 15 billion USD

People also search for View 15+ more

Megatsu…

2011 Tōhoku earthqua...

1964 Alaska earthquake

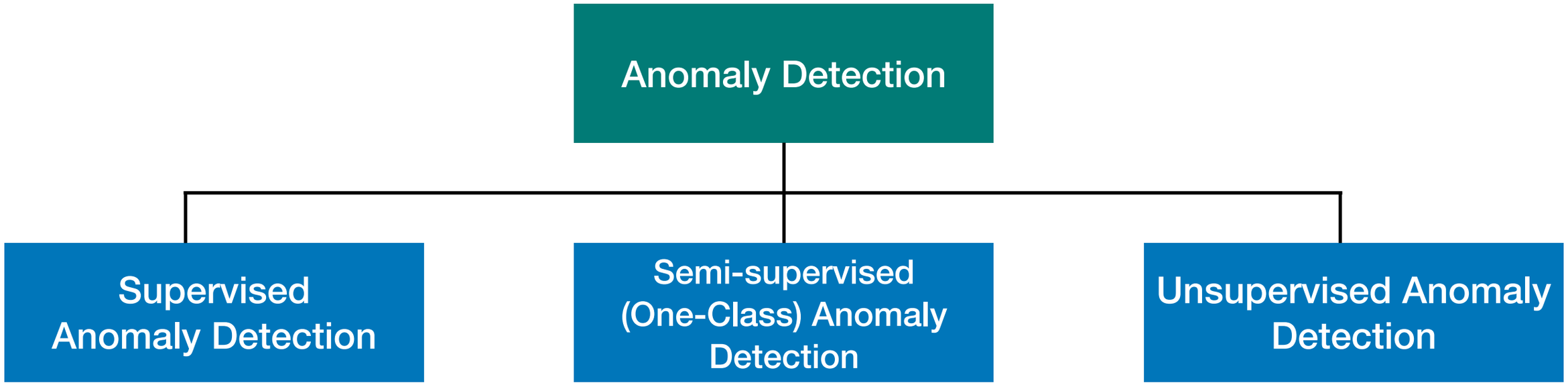
1960 Valdivia earthquake

2010 Chile earthquake

Feedback

*** 장현구, "미국 토네이도 재산피해만 8년 연속 100억 달러", 연합뉴스, 2016. 01. 05**

이상탐지

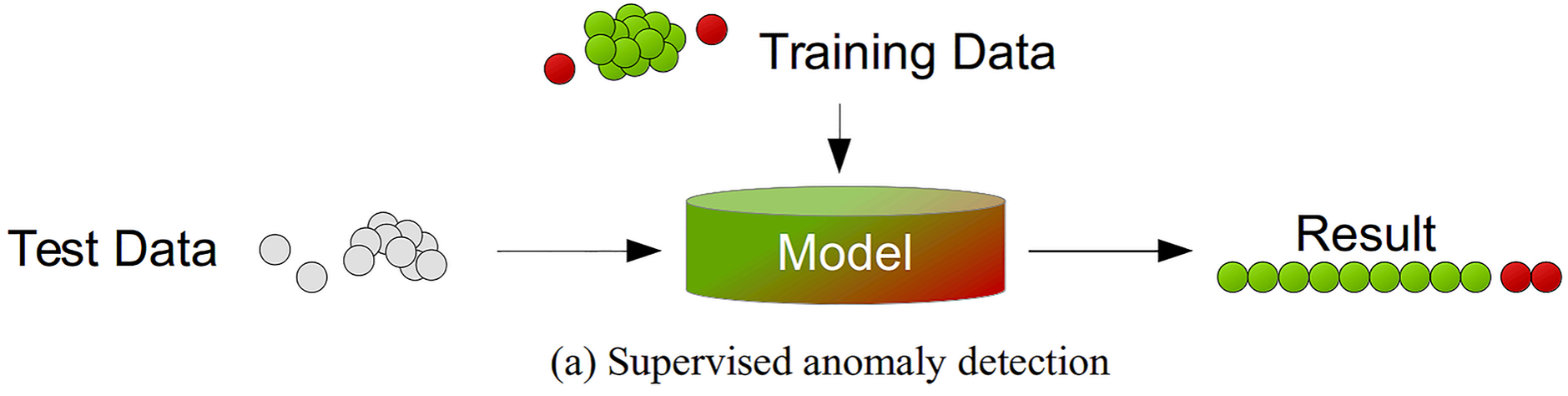


- 정상/비정상 Label이 주어진 경우
- 지도학습이므로 정확도가 높은편
- 일반적으로 비정상샘플이 정상샘플에 비해 적으므로 Class-Imbalance 문제에 직면함

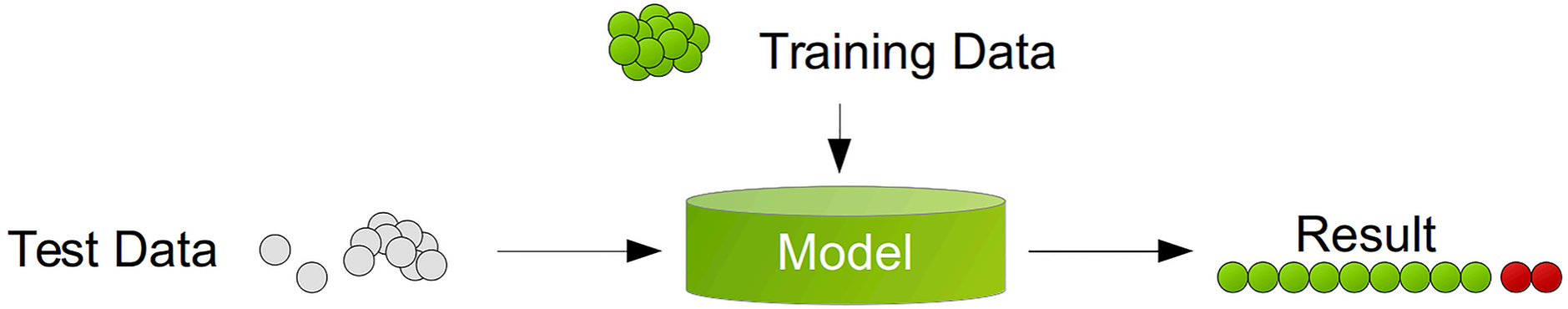
- 정상/비정상 Label이 주어진 경우
- **정상 샘플만을** 갖고 학습하여, 정상의 범주(boundary) 를 결정
- 정상 샘플만을 활용하기 때문에, Class-Imbalance 문제에 직면하지 않음
- 지도 이상 탐지에 비해 **성능이 떨어짐**

- **Label 없이 모두 정상이라고 가정**
- 클러스터링 및 거리 기반의 비지도학습 알고리즘 사용
- PCA나 AutoEncoder를 이용하여, 원본과 복원본을 비교하여 차이를 기준으로 판별하는 방법도 많이 쓰임
- 정확도가 높지 않고, hyperparameter에 의해 영향을 많이 받음

이상탐지



(a) Supervised anomaly detection



(b) Semi-supervised anomaly detection

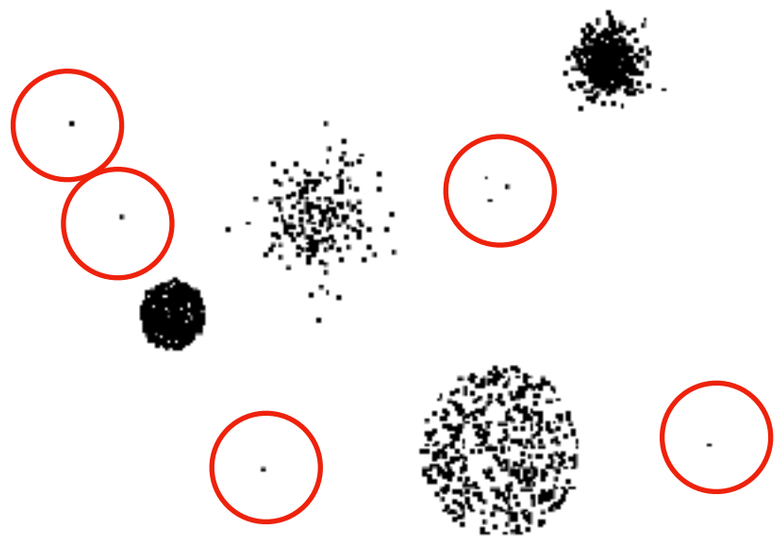
(c) Unsupervised anomaly detection

doi: https://doi.org/10.1371/journal.pone.0152173.g001

이상 탐지 접근법

- 일반적인 순서
  - 정상적인 행동의 기준 설립
    - 전체 모집단의 패턴 또는 요약 통계
  - 정상기준을 사용해 이상탐지
    - 이상 현상은 특성이 정상 기준과 크게 다른 관측치

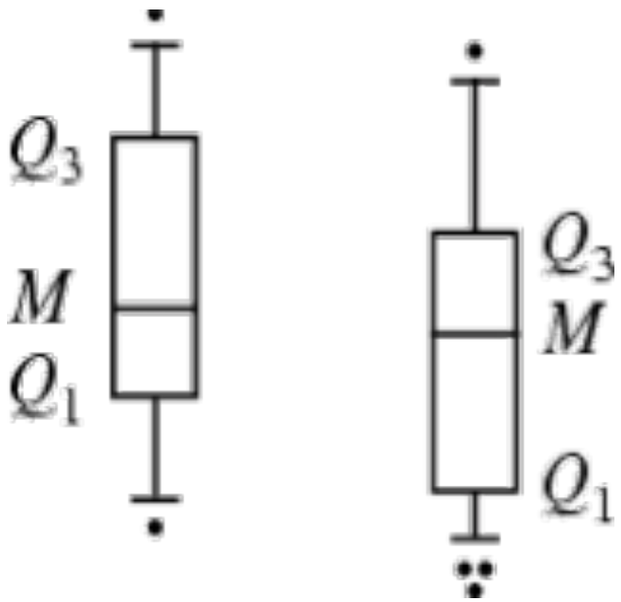
- 이상 탐지 접근법
  - Graphical
  - Model-based
    - Label이 존재하는 경우
  - Distance-based
  - Clustering-based

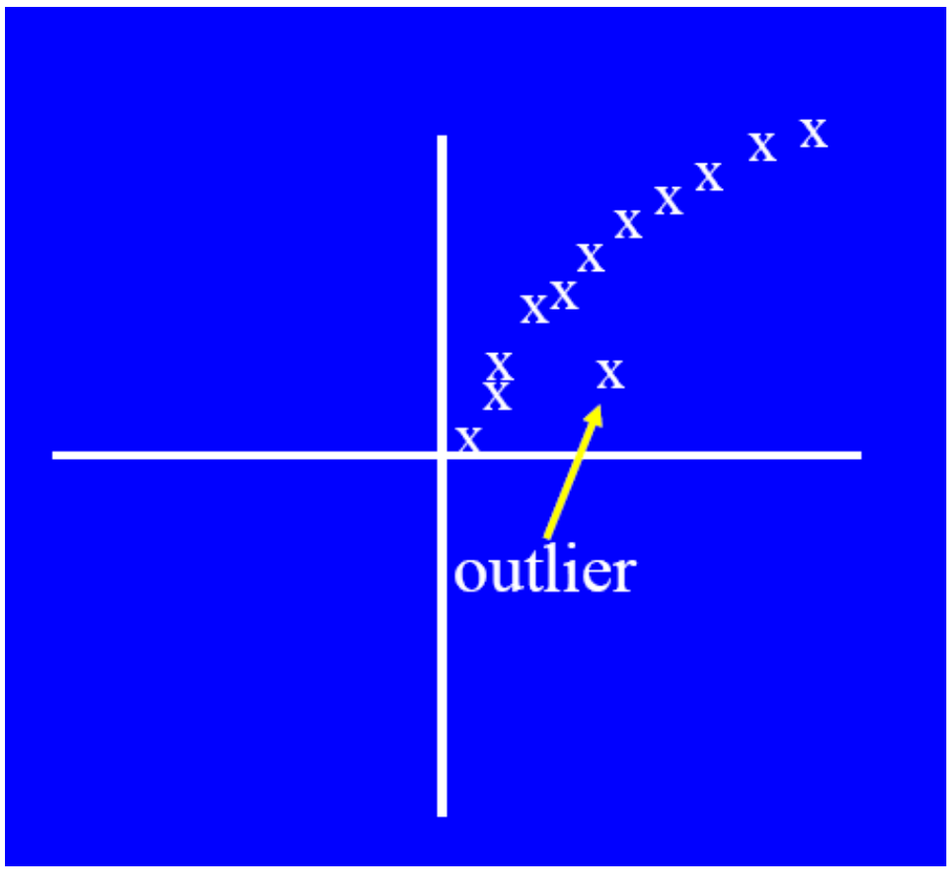


시각적 접근법

- Boxplot (1-D), Scatter plot (2-D), Spin plot (3-D)

- 한계
  - Time consuming
  - 주관적

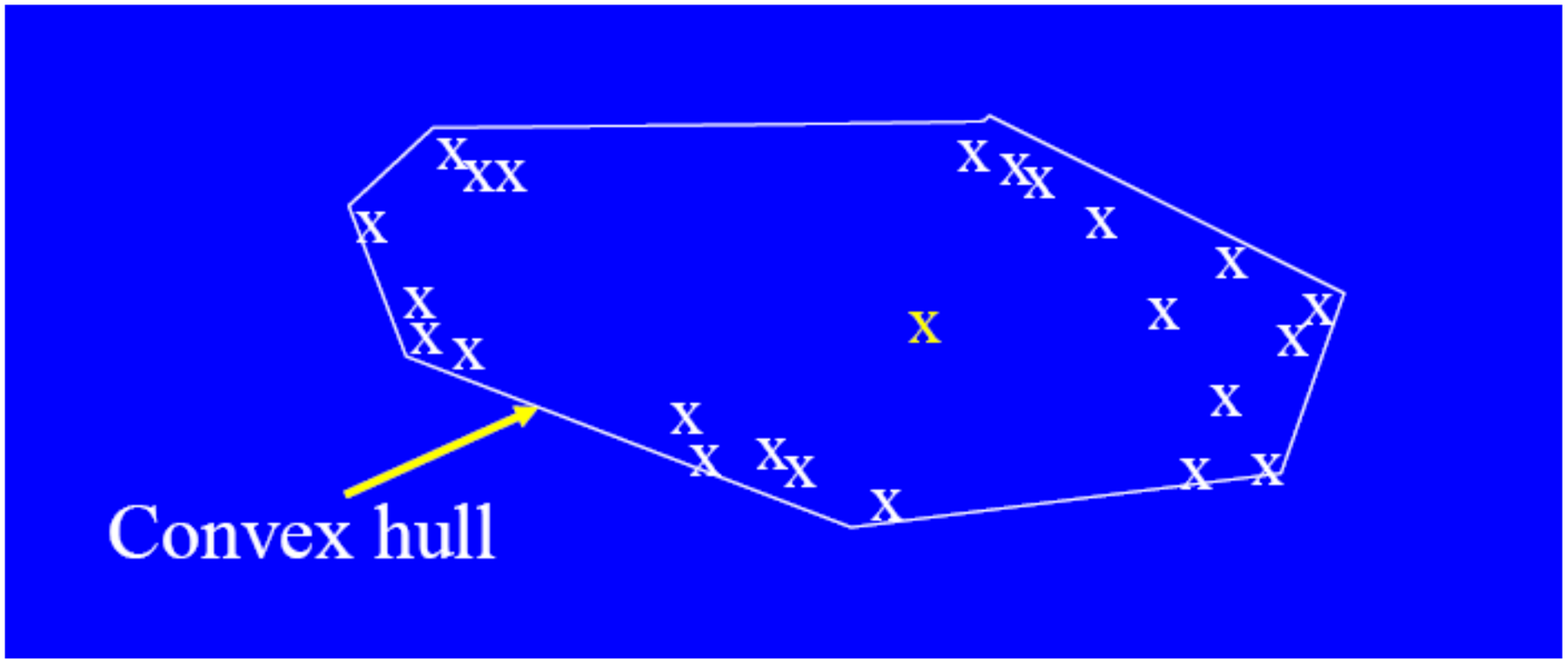




* Source: Tan, Steinbach, Kumar, Introduction to datamining, 2004.04.18

Convex Hull Method

- 극단 점은 이상치 인 것으로 가정
- 볼록 껍질 방법을 사용하여 극한값 감지

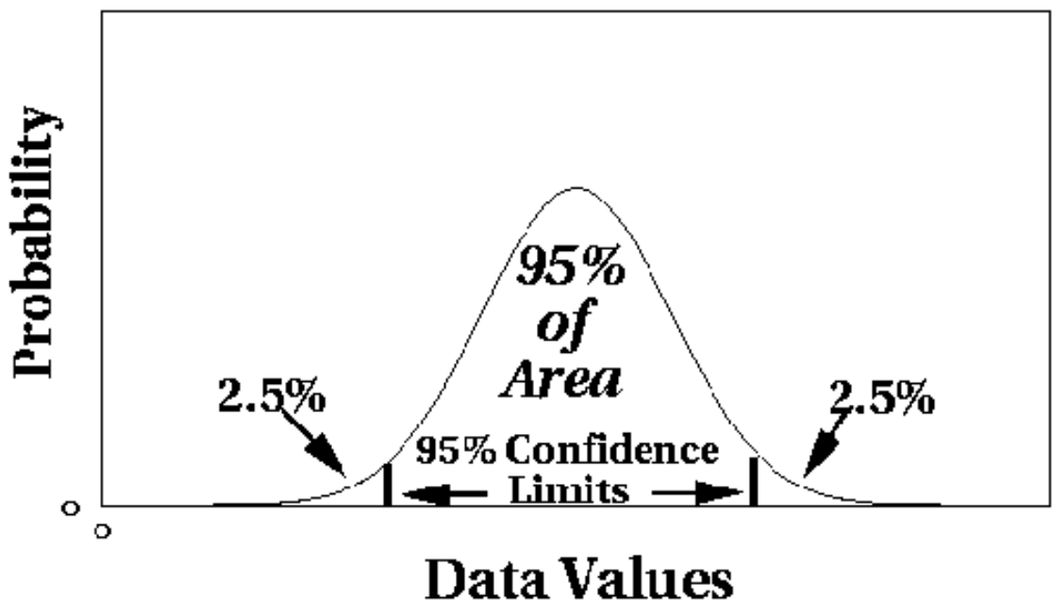


* Source: Tan, Steinbach, Kumar, Introduction to datamining, 2004.04.18

통계적 접근법

- 데이터 분포 (예 : 정규 분포)를 설명하는 모수 모델을 가정

- 다음에 의존하는 통계 테스트를 적용
  - 데이터 배포
  - 분포 모수 (예 : 평균, 분산)
  - 예상 이상치 수 (신뢰 제한)



* Source: Tan, Steinbach, Kumar, Introduction to datamining, 2004.04.18

통계적 접근의 한계

- 대부분의 테스트는 단일 속성에 대한 것
- 대부분의 경우, 데이터 분배 / 모델을 알 수 없음
- 고차원 데이터의 경우 실제 분포를 추정하기 어려움

거리 기반 접근법

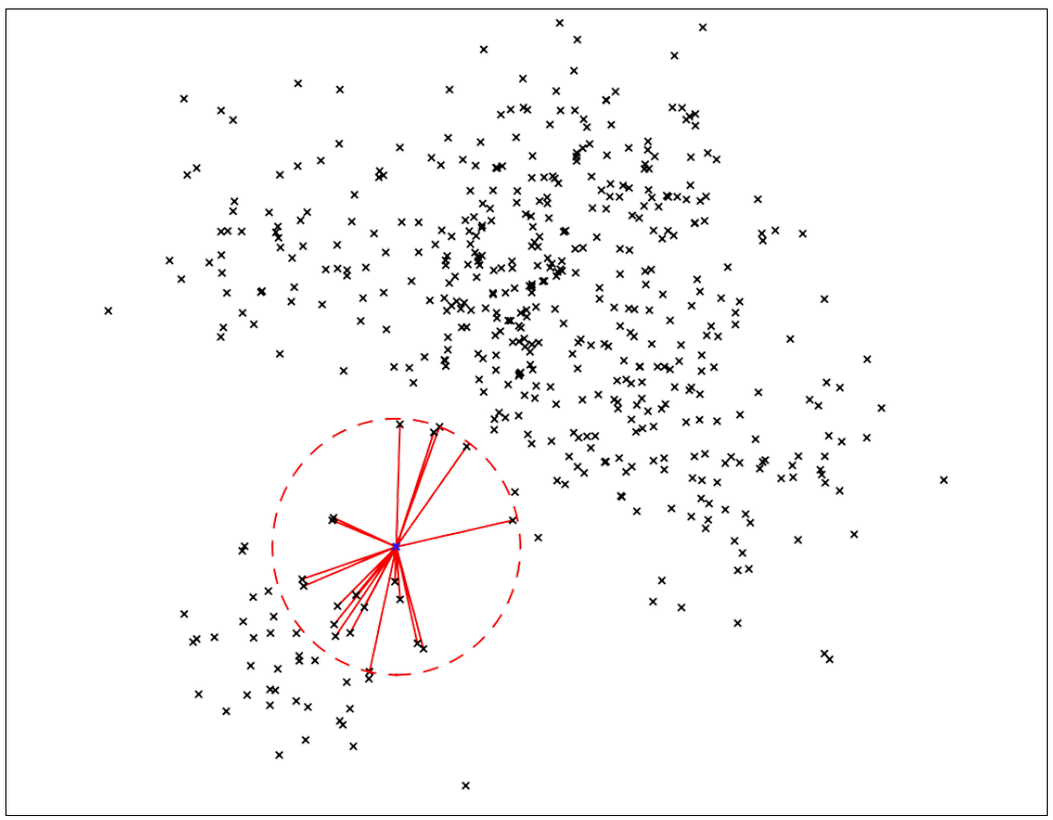
- 데이터는 feature들의 벡터로 표현됨

- 세 가지 주요 접근법
  - Nearest-neighbor based
  - Density-based
  - Clustering-based

Nearest-Neighbor 기반 접근법

- 모든 데이터 포인트 쌍 사이의 거리 계산

- 특이치를 정의하는 다양한 방법
  - 거리 D 내에서 인접 포인트가 p보다 적은 데이터 포인트
  - 가장 가까운 k 번째 이웃까지의 거리가 가장 큰 상위 n 개의 데이터 포인트
  - k 개의 가장 가까운 이웃까지의 평균 거리가 가장 큰 상위 n 개의 데이터 포인트



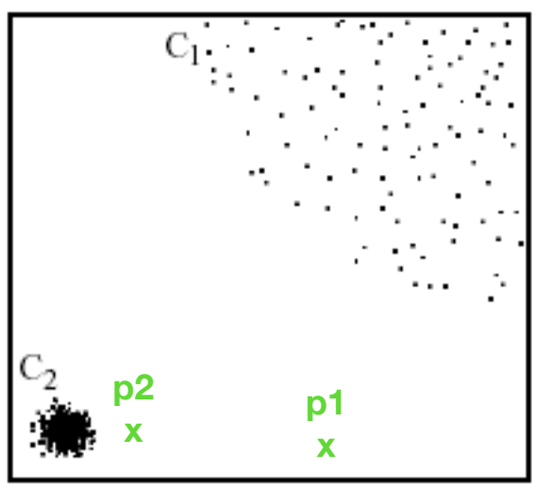
https://jayhey.github.io/novelty%20detection/2018/01/29/Novelty_detection_KNN/

밀도기반 접근법-LOF

- 각 점에 대해 해당 지역의 밀도를 계산
  - 예 : DBSCAN

- 샘플 p의 **local outlier factor**(국소 특이치 인자)를 샘플 p의 밀도와 가장 가까운 이웃의 밀도의 평균으로 계산

- 특이 치는 LOF 값이 가장 큰 점



NN 접근법에서 p2는 특이 치로 간주되지 않지만 LOF 접근법은 p1과 p2를 특이치로 찾음

다른 접근법 : 밀도 함수를 직접 사용
예 : DENCLUE의 밀도 함수

* Source: Tan, Steinbach, Kumar, Introduction to datamining, 2004.04.18

클러스터링 기반 접근법

- 아이디어 : 특이치라는 컨셉이 있는 군집 알고리즘을 사용
- 문제점 : 알고리즘에 어떤 매개 변수를 선택해야할까?
  - 예 : DBSCAN?
- 데이터의 x% 미만이 특이치어야함
  - x는 일반적으로 0.1과 10 사이에서 선택됨.
  - x는 다른 방법으로 결정가능
    - 예 : 통계 테스트

E.O.D