BIG DATA ANALYTICS

WEEK-10 | Anomaly Detection

Yonsei University
Jungwon Seo

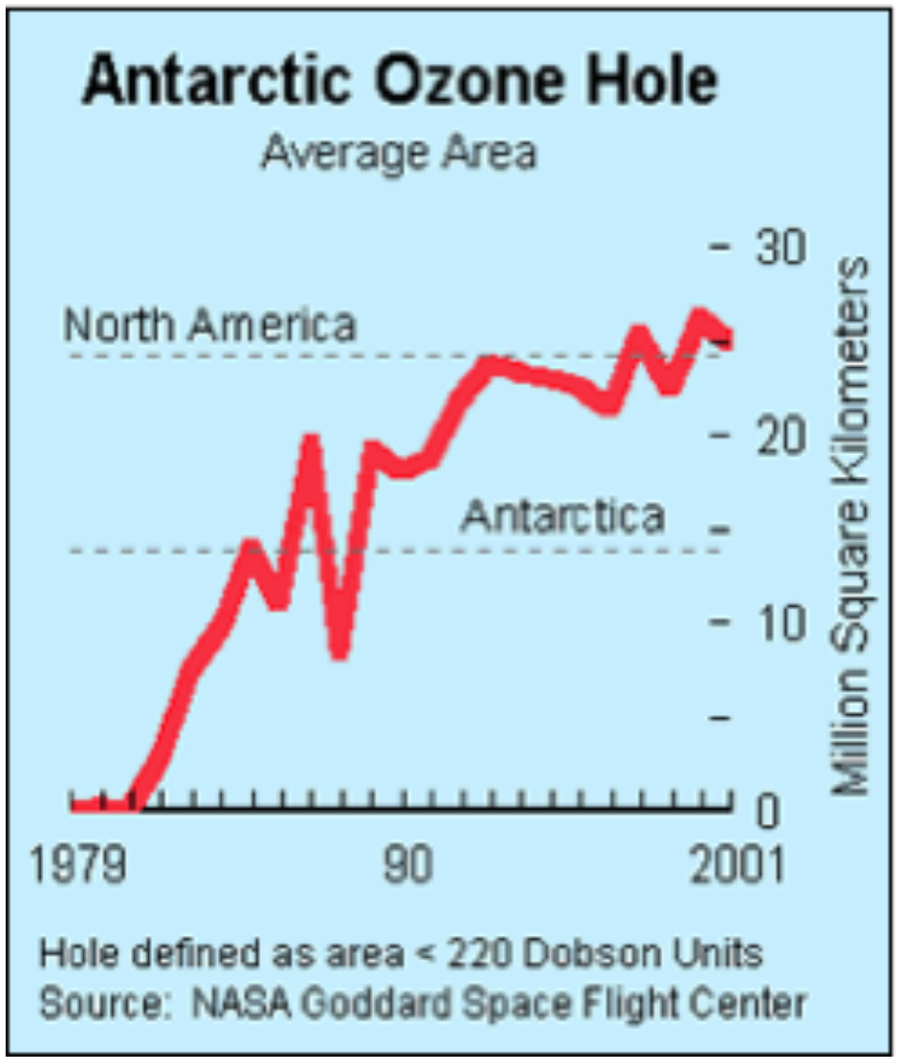
이상탐지의 중요성

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

- 응용 사례
  - 신용 카드 사기 탐지
  - 통신 사기 탐지
  - 네트워크 침입 탐지, 결함 탐지

이상탐지의 중요성

- 오존층 파괴 이력
- 1985 년 영국 남극 조사에서 수집 한 자료에 따르면 남극 대륙의 오존 수치가 정상 수치보다 10 % 낮음
- 오존 수준을 기록하기위한 장비가 장착 된 Nimbus 7 위성이 오존 농도를 비슷하게 기록하지 않은 이유는?
- 위성에 의해 기록 된 오존 농도는 너무 낮아서 컴퓨터 프로그램에 의해 이상치로 처리되어 버려짐



* Source: http://exploringdata.cqu.edu.au/ ozone.html
** Source: http://www.epa.gov/ozone/science/hole/size.html
*** Source: https://svs.gsfc.nasa.gov/11644

이상탐지

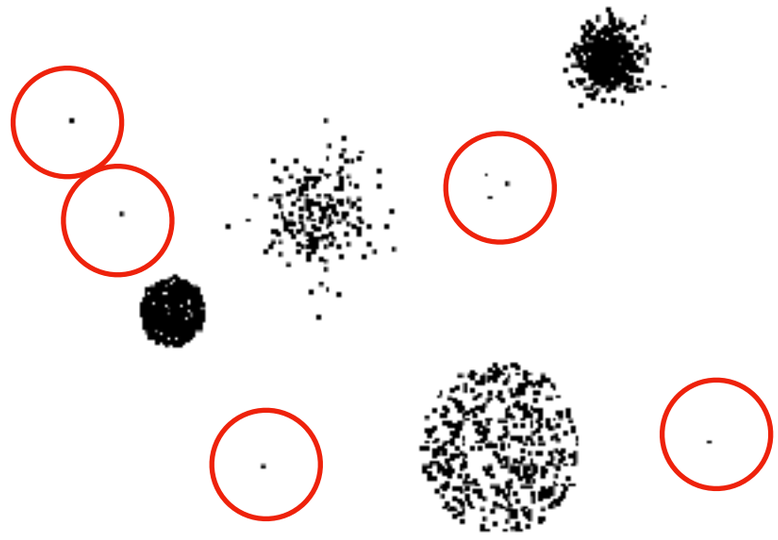
- Challenges
  - 데이터에 몇 개의 특이 치가 있나?
  - Unsupervised Learning
  - 클러스터링과 마찬가지로 유효성 검사는 매우 어려움
  - 사막에서 바늘 찾기

- 가정
  - 데이터에 "비정상"관측치 (이상치 / 이상)보다 상당히 많은 "정상"관측치가 있다.

이상 탐지 체계

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

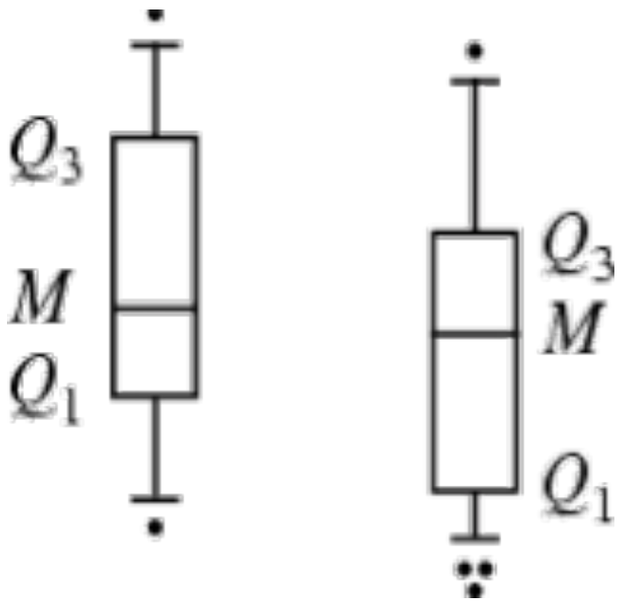
- 이상 감지 유형 체계
  - Graphical
  - Model-based
  - Distance-based
  - Clustering-based

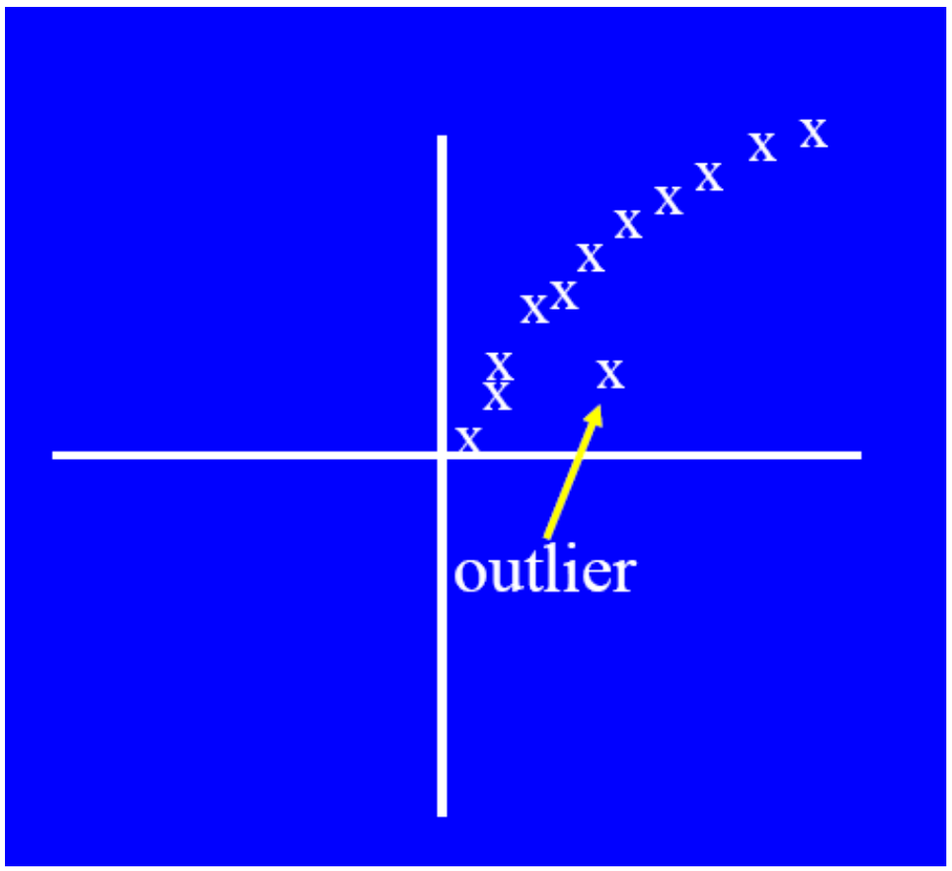


시각적 접근법

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

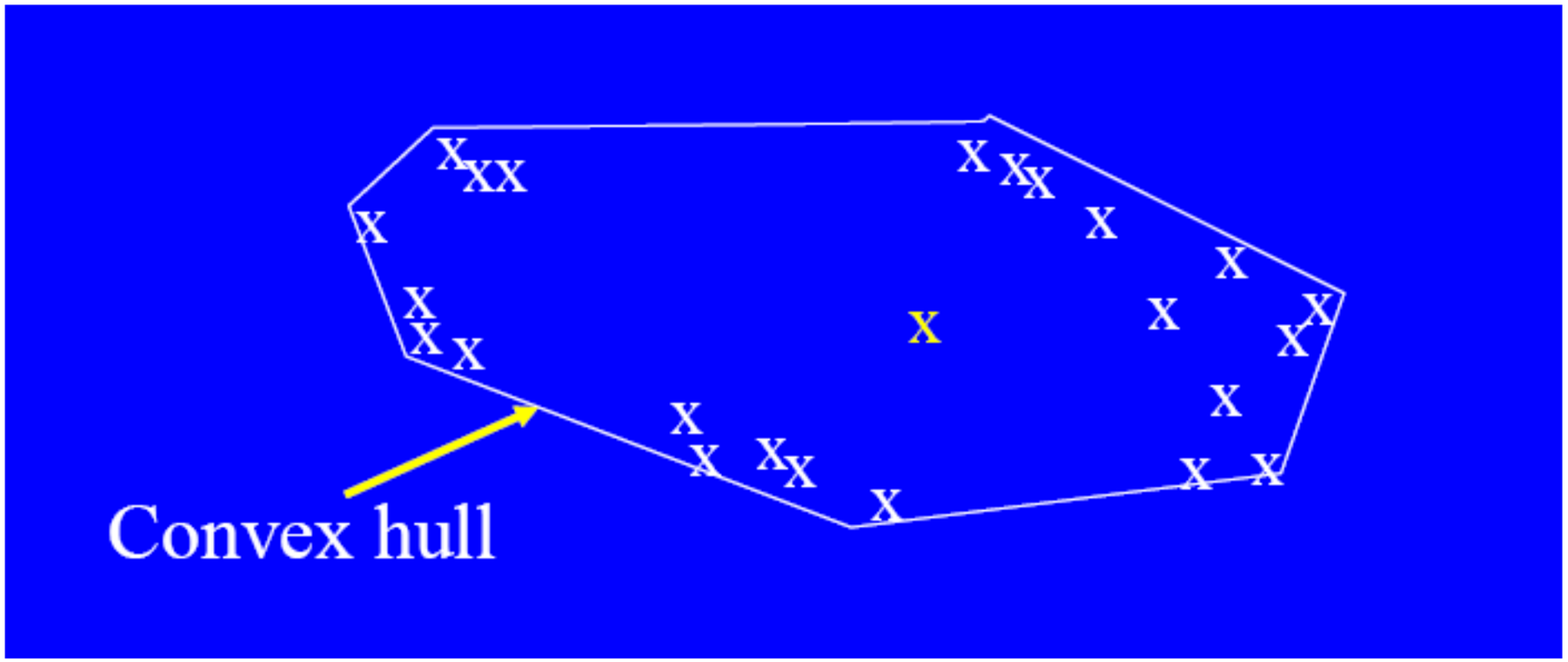
- 한계
  - Time consuming
  - 주관적





Convex Hull Method

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

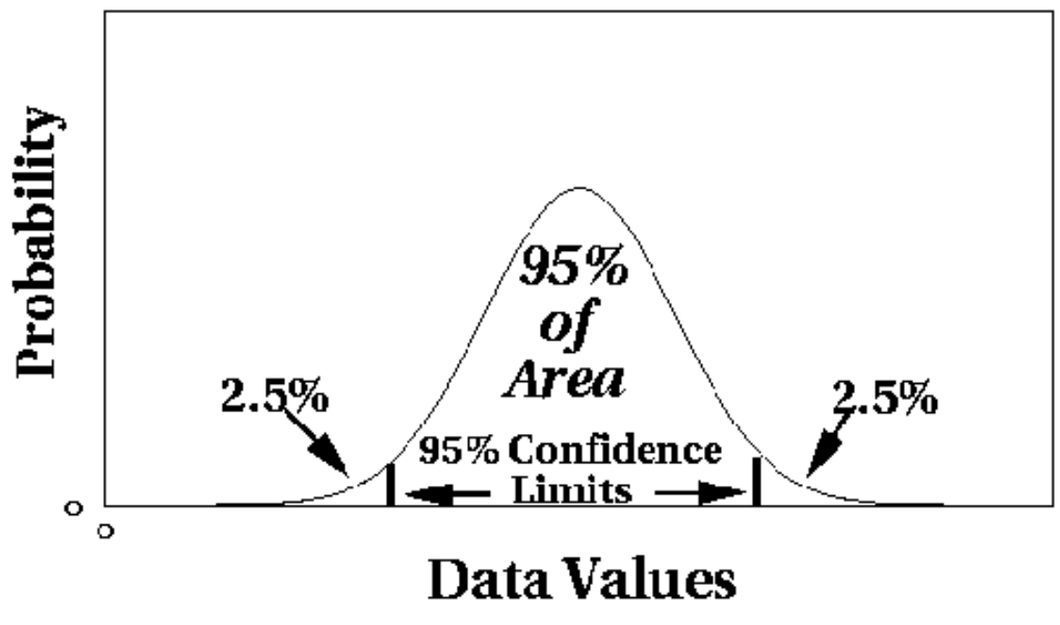


* Source: http://cgm.cs.mcgill.ca/~godfried/teaching/projects97/belair/alpha.html

통계적 접근법 Model-based

- 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)



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

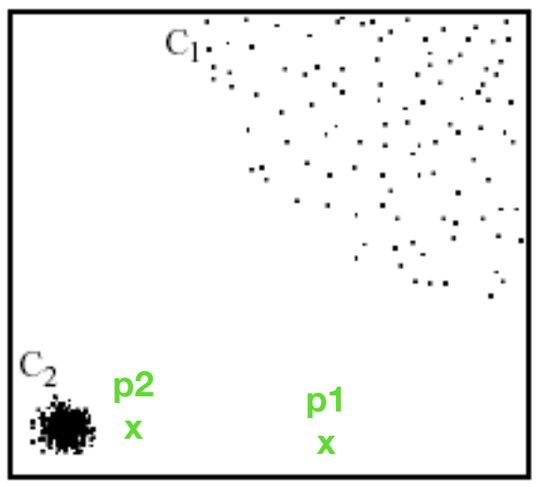
- 특이치를 정의하는 다양한 방법
  - 거리 D 내에서 인접 포인트가 p보다 적은 데이터 포인트
  - 가장 가까운 k 번째 이웃까지의 거리가 가장 큰 상위 n 개의 데이터 포인트
  - k 개의 가장 가까운 이웃까지의 평균 거리가 가장 큰 상위 n 개의 데이터 포인트

밀도기반 접근법

- 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

클러스터링 기반 접근법

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

E.O.D