A survey of outlier detection

Lee ji soo

Department of Statistics
Seoul National University

2018.08.09

type1 방법

type 1 방법

- 데이터에 대한 선험적 지식 없이 outlier를 판별한다
- unsupervised clustering과 유사하며 데이터를 정적 분포로 처리
- 모든 데이터가 처리되기 전에 정적이라는 전제가 필요
- 현재 상태의 확률이 직전 상태의 확률과 같아지게 되는 평형 상태에 도달한 확률 분포
- diagnosis -이상점의 가능성이 높은 점들을 주목하고 이러한 점들을 다음 가공에 배재하는 법

type 2 방법

- supervised classification과 유사하며 pre-labelled data를 필요로 한다
- classification model을 익히고 새로운 exemplar를 분류
- 새로운 exemplar가 normality region에 있으면 normal로 분류하고 그렇지 않으면 outlier로 분류
- normal과 abnormal data의 good spread가 필요 -classification이 아는 분포로 제한되어 있을 때 새로운 exemplar가 제대로 분류

type 3 방법

- normality만 모델링하거나 abnormality를 극히 조금의 case에서 모델링한다
- novelty detection으로 명명되고 semi-supervised recognition과 같다
- pre-classified된 data를 필요로 하지만 normal이라고 마크된 데이터만 학습
- normality의 경계를 정의하고 new exemplar를 판별한다
- type 2 방법과 다르게 unexpected region에서의 outlier도 다룰 수 있다
- abnormal data를 구하기 어려운 비행기 엔진 monitoring 같은 문제에 적합

Statistical model

- Grubb's method는 1차원 method, 평균과 분산으로부터 z value를 계산
- 1퍼센트, 5퍼센트 유의 수준 level하에 값들을 비교
- Laurikkala(2010) method는 box plot들을 이용해 outlier를 판별한다
- 차원이 증가할수록 processing time이 증가하고 큰 부피로 가면서 sparse해진다
- 이를 차원의 저주라 하고 data를 더 낮은 차원의 subspace로 project 하거나 pca 방법을 사용한다

Proximity-based techniques

- 데이터 분포 모델에 대한 선험적 가정이 필요없고 type1, type2 방법에 적합하다
- computational 복잡성이 차원수와 data 수들에 비례하기 때문에 knn 같은 방법은 고차원 data 의 경우 적합하지 않다
- Mahalanobis distance

$$\sqrt{(x-\mu)^T C^{-1}(x-\mu)} \tag{1}$$

- μ: 중앙에서 각 포인트까지의 거리 , C: Covariance matrix
- inter-attribute dependencies를 포함하므로 attribute combination들을 비교할 수 있다(다변량 자료에 적합)

optimised k-NN method

- Knorr and Ng(1998)이 제안, m개(m<k)의 이웃들이 d라는 특정 threshold 내에 있으면 normal로 분류
- Ramasway(2000)이 제안, 전체를 cell들로 partition한 후 k points보다 많으면 normal로 분류
- Wattschereck(1994), 다수 투표 방법을 제안
- features들의 수를 줄임(knn 방법 외에도 유용)

k-means method

- 새 데이터가 단지 k prototype와 비교하고 k prototype vector만 저장되면 되기 때문에 계산적인 복잡성이 감소한다
- k 개의 데이터 오브젝트를 임의로 추출하고, 이 데이터 오브젝트들을 각 클러스터의 중심으로 설정한다
- 각 데이터 오브젝트들에 대해 k 개의 클러스터 중심 오브젝트와의 거리를 각각 구하고, 각 데이터 오브젝트가 어느 중심점와 가장 유사도가 높은지 알아낸다
- 그리고 그렇게 찾아낸 중심점으로 각 데이터 오브젝트들을 할당한다

k-means method

- $$argmin \sum_{j=1}^{K} \sum_{n \in S_j} |x^n - \mu_j|^2 \quad (2)$$

- 각 중심점과 cluster에서 가장 먼 점까지의 거리를 반지름으로 설정한다
- 이 반지름이 normality의 경계를 정의하고 knn 방법들의 global 거리와 비교하면 지역적인 특색이 있다
- 포인트가 모든 클러스터들 바깥에 있을 때 outlier로 분류한다

k-medoids method

- Bolton and Hand(2001)이 fraud detection 발견을 위해 사용
- 클러스터의 중심을 실제값으로 쓰기 때문에 k-means의 경우처럼 가끔 poor 클러스터들로 구성되지 않는다
- 이 때문에 outlier detection에 효과적이다
- 하지만 k-medoids 의 경우 $O(n^2)$의 iteration당 실행 시간으로 k-means 의 경우 $O(n)$보다 길어 데이터 셋이 많을 경우 적합하지 않다

graph connectivity method

- Shekhar et al.(2001)이 교통 모니터링을 위해 도입
- 점들이 거리에 근거한 측정법보다는 위상적으로 연결되어 있는지에 관심
- 한 포인트의 값과 위상적인 이웃들의 평균 값의 차이가 한계점을 넘으면 outlier로 판단한다
- 교통 흐름 네트워크처럼 각 점이 연결된 네트워크에서 하나의 노드를 구성할 때 효과적

Parametric methods

- MVE(Rousseeuw and Lroy,1996)은 데이터 분포 모델의 대다수 주위에 가장 작은 허용되는 ellipsoid volume을 fit한다
- convex peeling(Rousseeuw and Leroy,1996) 은 data 분포의 가장 자리부터 깎아나간다
- 각각의 점들에 깊이를 할당하고 미리 정한 횟수만큼 깊이가 같은 점들의 경계를 깎는다
- normal 점들을 많이 깎는 단점이 있다
- 차원이 늘어날수록 surface를 식별하기 어려우므로 저차원에서 주로 쓰인다

pruning and re-fitting

- $$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3}$$
- Torr and Murray(1993)은 위의 합을 최소화하는 방향으로 포인트들을 삭제한다

$$Median_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4}$$

- 위를 최소화하는 방법도 이상점의 수가 작을 때 쓰는데 계산적으로 간단하다

$$\sum_{i=1}^{h}(y_i - \hat{y}_i^2)_{i:n} \tag{5}$$

- least trimmed squares approach(Rousseeuw and Leroy,1996)
- 빠르게 converge하고 이상점의 숫자가 많을 때 사용가능하다

Non-parametric methods

- knn,k-means, convex hulls and regression 모두 데이터가 특정한 모델을 따른다는 가정
- non-parametric 방법은 더 유연하고 자율적임
- Dasguspta and Forrest(1996)은 machinery operation으로 이상점 탐지
- 시계열의 데이터들을 machinery operation으로 새로운 string(binary vector window)로 만들고 그것들과 매치하지 않은 detector들을 만든다
- 새로운 string이 detector와 매치하지 않으면 이상점으로 판단한다

semi-parametric methods

- semi-parametric 방법들은 local kernel 방법들을 사용
- 커널에 기반한 방법들은 input 공간의 밀도를 측정 후 낮은 밀도의 부분을 이상점으로 분류
- Gaussian mixture model(Roberto and tarassenko,1995)

$$p(t|x) = \sum_{j=1}^{M} \alpha_j(x)\phi_j(t|x) \tag{6}$$

- M은 커널 ϕ의 갯수, α_j는 혼합 계수, x는 투입 벡터, t는 대상 벡터

semi-parametric methods

- extreme value theory (Robert, 1998) 은 Gaussian mixture model 사용
- 이상점(extreme value)가 분포의 꼬리 부분에 나타남
- Gaussian mixture model(Roberto and tarassenko,1995)

$$p(extreme_x) = exp(-exp(-\frac{x_m - \mu_m}{\sigma_m})) \tag{7}$$

- 이상값들을 얻기 힘들거나 비쌀 경우 사용

semi-parametric methods

- Tax et al.(1999)와 DeCoste and Levine(2000)이 support vector machines 사용
- Support vector 함수는 이상점일 경우 음수로 표현됨

$$SV = sign(\sum_{j=1}^{n} \alpha_j L_j K(x_j, z) + b) \tag{8}$$

- K는 커널 함수 L_j는 클래스 라벨,b는 bias, z는 test input, x_j는 trained input
- 기계상태 진단과 의학적 분류에 주로 사용(정상과 이상점의 비율이 불균형일 때)

supervised neural methods

- Nairacet al.(1999)와 Bishop(1994)가 다층 퍼셉트론을 single hidden layer만 사용
- $$Error = \sum_{j=1}^{m} \int [y_j(x;w) - \langle t_j|x\rangle]^2 p(x)dx + \sum_{j=1}^{m} \int [\langle t_j^2|x\rangle - \langle t_j|x\rangle^2] p(x)dx \tag{9}$$
- $y_j(x;w)$는 function mapping, t_j는 target class,b는 bias, z는 test input, y_j는 actual class, p(x)는 unconditional 확률 밀도
- $y_j(x;w) = p\langle t_j|x\rangle$ 일 때 error가 최소화
- p(x)가 크면 error 함수가 네트워크에 패널티를 준다

Hopefield network

- Crook and Hayes(1995)가 type3 novelty detection에 사용

$$Energy = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} x_i x_j w_{ij} \tag{10}$$

- N 은 뉴런의 갯수 w_{ij}는 1 or -1
- Energy 값이 한계점을 넘으면 이상점으로 판별한다

decision tree

- decision tree는 다른 통계 방법들과 달리 데이터에 대한 선험적 지식이 필요없다
- curse of dimensionality에 영향받지 않음
- Arning et al.(1996)은 머신 러닝 방법을 기반으로 pruning을 시도
- 데이터를 sequence로 조사하고 차이점 함수를 사용해 새로운 점의 조사되었던 점과의 유사성 비교

rule-based system

- 일련의 조건들로 테스트한다는 점에서 decision tree와 유사
- 기존 룰에서 새로운 룰을 추가하거나 룰이 개정됨
- 학습된 룰은 프로파일링 모니터를 만들고 이상점을 찾아낸다
- 데이터를 sequence로 조사하고 차이점 함수를 사용해 새로운 점이 조사되었던 점과의 유사성 비교

Hybrid systems

- Smyth(1994)는 hidden markov model을 MLP에 적용
- MLP는 back-propagation을 사용해 다음 단계를 예상한다
- MLP는 상태들이 자주 바뀌는 경향이 있음
- HMM이 상태 계산값을 correlate하고 EM 알고리즘을 적용해 예측을 강화한다

JAM (java agents of meta-learning)

- Columbia 대학(Stolfo et al., 1997)이 5개의 머신러닝 테크닉을 한 아키텍쳐에 집약
- 다양한 모델을 사용한 결과를 합성해 한 개의 시스템 결과값을 생성한다
- ID 3 decision tree, CART, C4.5, Ripper, naive Bayes classifier에 기반
- 다양한 분류기들을 사용해 결과값을 내고 가장 적합한 테크닉을 사용한다

결론

- 모든 상황에서 최적인 한 개의 단일한 방법론은 존재하지 않는다
- 문제점이 clustering 접근, classification 접근, novelty 접근 중 무엇에 적합한지 먼저 생각
- feature extraction이나 prototyping 같이 먼저 가공하는 것을 고려해야 한다

The End