SDT와 민감도 기반 동적 이상점 제거

산업공학과 문현지

동기: 적절한 이상점 제거기준?

예측웹서비스 스타트업 운영 중 (qtell.co.kr) 예측정확도에 결정적인 [1] 이상점 처리방식 필요성

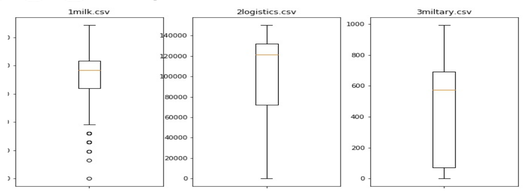
1. 산업별로 이상점 발생원인 다름 (그림1)

예) 군식당 이상점 요인
- 비상대기: 북한 핵 실험, 국가 사고
- 군사훈련: 을지 포커스렌즈
- 복날: 저렴한 닭 한 마리
- 기상상황 : 비

2. 데이터량별로 이상점 형태 다름

예) 한 달 vs 한 분기 vs 1년

→ 세 종류의 산업 수요데이터 바탕으로 적합한 이상점 제거 기준탐구



Same Outlier cutoff Method?

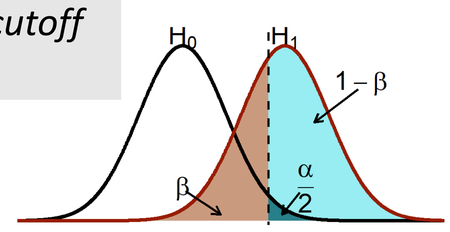
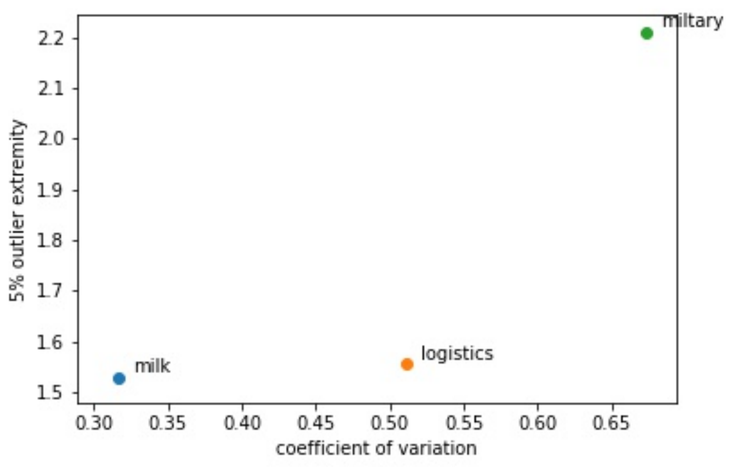


그림1. 산업별 다른 이상점 특성 (우유, 물류, 군식당)

제안: 데이터의 이상점 특성별, 데이터 개수별로 변하는 이상점 제거기준



높은 민감도
= 높은 변동성 (표준편차/평균)
= 높은 이상점 극단성 (상위 5% 이상점과 나머지의 평균비)

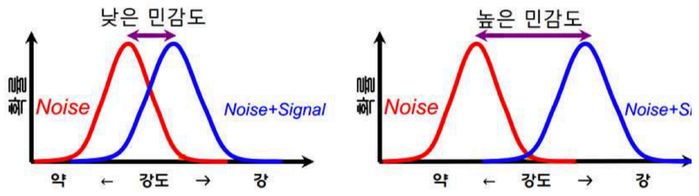


그림2. 민감도 관점 산업별 이상점 특성

Different beta depending on Data!

정적 제거
신호잡음 분리에 고정된 β

vs

동적 제거
민감도, 데이터 개수 따라 다른 β
(β = 이상점 극단성 + 데이터량 x축 sigmoid함수)

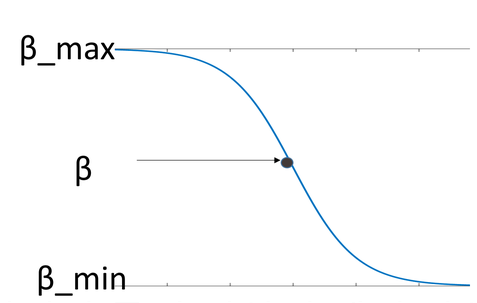


그림3. 정적과 동적 이상점 제거 방식 비교

결과 분석

	제거 x	정적 제거	동적 제거
우유	1.102	0.702	0.661
물류	0.865	0.688	0.688
군식당	1.422	1.174	1.174

표1. 산업과 이상점 제거 방식에 따른 예측오차

	제거 x	정적 제거	동적 제거
60일 (1회 예측)	0.818	0.713	0.713
120일 (3회 예측)	0.813	0.704	0.697
240일 (5회 예측)	0.791	0.714	0.695

표2. 데이터량과 이상점 제거 방식에 따른 예측오차

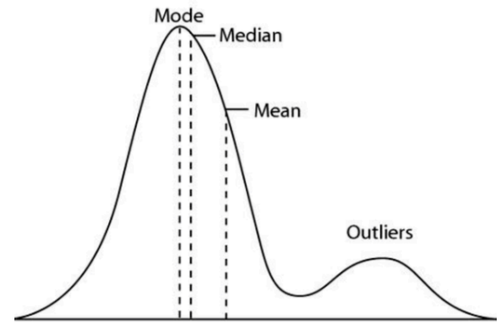
낮은 민감도 -> 동적 제거 방식 효과적!

많은 데이터량 -> 동적 제거 방식 효과적!

한계점 및 개선 방안

Outlier = Noise?

과거의 error 분석 통해 유사 error를 예방가능! e.g. 9.11테러같은 상상 초월 사건도 거시적인 관점에서는 이상치가 아닐 수 있다 [2] (네이트 실버)
이상점을 noise로 간주하고 삭제한 것이 한계
이상치제거에서 손실되는 정보량의 최소화를 위해
이상점 생성기작을 모형에 반영하는 mixture models [3] 통해 개선 가능



참고문헌
1) Rousseeuw, P. J., & Leroy, A. M. (2005). *Robust regression and outlier detection* (Vol. 589). John wiley & sons.
2) Silver, N. (2012). *The signal and the noise: the art and science of prediction*. Penguin UK.
3) Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.