FPRI
Future Policy Research Institute

Anomaly Detection & AI

- 이상감지와 인공지능 V1.0 -

2019.07.27

Jason, Min

지진예측의 가치

예측 : 지진이 발생하는 시각, 장소, 강도 예측필요
비용 : 주민대피에 큰 비용이 들어감
유용한 예보 기준 : 50%의 정확도, 하루 정도의 정확도, 50Km 이내

그러나 지진에 대한 예측 성공사례는 없음

이상징후탐지(Anomaly Detection)

1) 과거 시계열 데이터 또는 2) 비슷한 시점의 다른 보편적인 데이터와 상이하거나 그러한 징후가 있는 패턴을 찾아내는 데이터 분석의 한 분야

* 이상행위 탐지(상이)는 전수조사 시 100% 확인는 가능함

이상징후탐지시스템

데이터를 분석하여이상패턴을 탐지/저장/처리하는 시스템

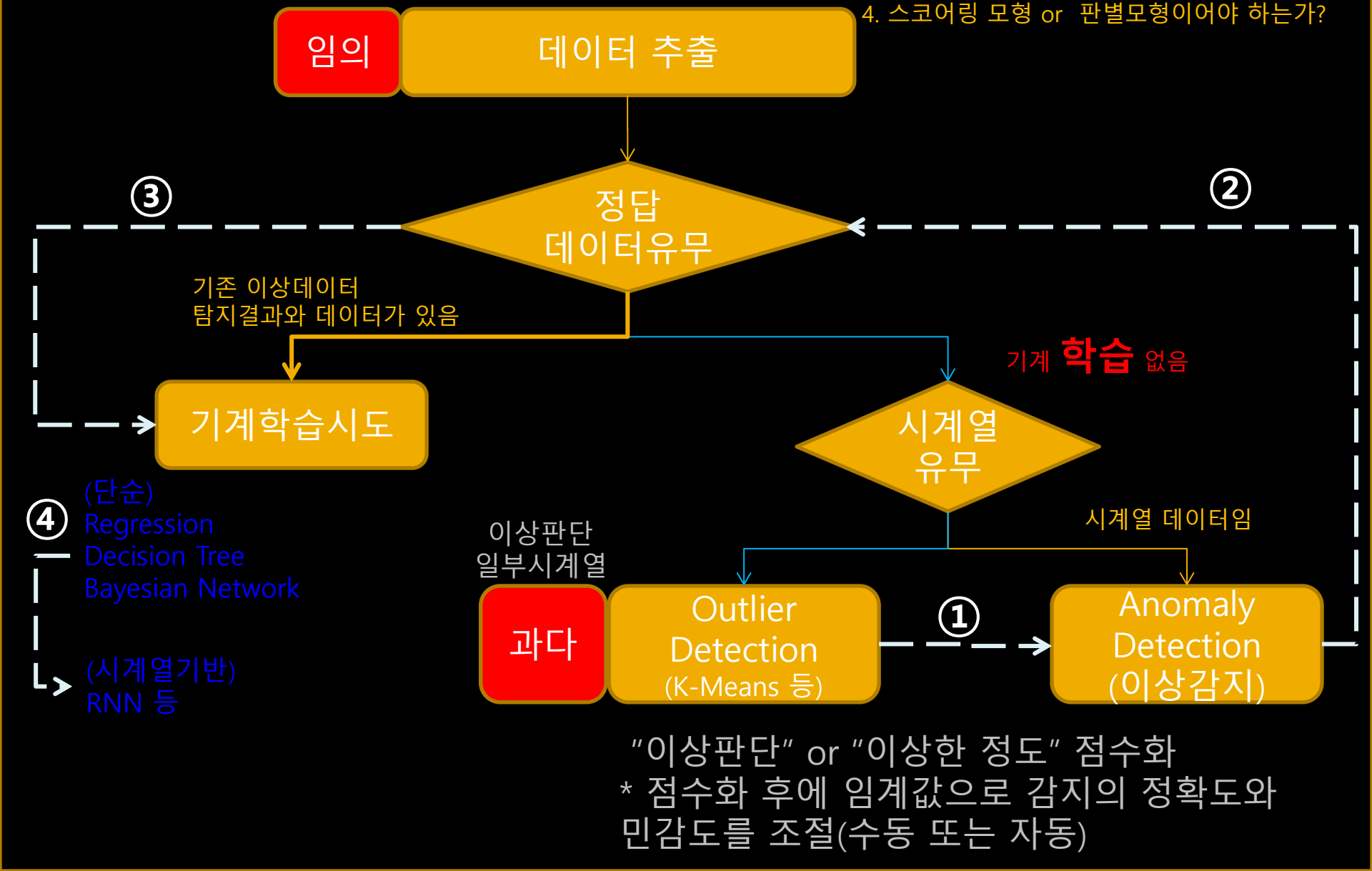
현재 : 어떤 대상의 현재까지의 상태가 이상한지를 감지하는 것 → 현 정리범위 & 가변임계치
미래 : 어떤 대상이 미래에 문제가 발생할 여지가 있는지 예측하는 것

현재 (+과거) **이상감지** : 시간 또는 순서가 있는 흐름에 따른 패턴이 보편적인 상황 또는 보편적인 패턴들과 다른 것들을 찾아내는 것
현재 (현재 ONLY) **아웃라이어** : 시간과 관련이 없이 대상을 표현하는 숫자들의 위치를 보고 보편적인 대상과 벗어난 것을 찾아내는 것

FPRI
Future Policy Research Institute

[방법론 선정 체크리스트]

1. 기계학습을 할 수 있는 정답데이터가 있는가?
2. 시계열 데이터인가? 아닌가?
3. 단변량인가? 다변량인가?
4. 스코어링 모형 or 판별모형이어야 하는가?



"이상판단" or "이상한 정도" 점수화
* 점수화 후에 임계값으로 감지의 정확도와 민감도를 조절(수동 또는 자동)

FPRI
Future Policy Research Institute

분석설계

분석목적정의 → 보유정보확인 → 분석가능여부 판단 → 기능설계 (데이터&알고리즘& 결과처리방법/리소스)

* 분석가능정보 확인

* 정보 특성 분석

* 데이터 알고리즘 처리방법 투입리소스 결정

개발 (솔루션 영역)

데이터 수집 → 데이터 가공 → 알고리즘-분석 → 탐지결과 출력 (시각화,연관정보제공, 출력건수 조정 등)

* 탐지목적에 적합한 데이터가 수집되는가?
* 수집데이터의 신뢰성을 검증하였는가?

* 탐지목적에 적합하도록 데이터가 가공되는가?

* 탐지목적에 적합한 **데이터**와 **알고리즘**이 적용 되었는가?

업무처리

탐지결과 검토 (총건수 D) → 세부검토 (총건수 C) → 최종결과저장 (총이상건수 A) → 업무효율성 검토

투입가능 리소스는 탐지결과 건수에 비례, 즉, 탐지효율성에 반비례함

* 탐지건수는 목표한 **리소스 내에서 처리**가 가능한가? (처리를 위해 팀내 모든 인원이 투입되어야 하는가?)

* 소명요청 등

* 상세검토는 동일시스템 내에서 판단이 가능한가? C/D (* 소명요청은 회사 내 추가 리소스가 투입되는것임)

* 실 이상건수/탐지건수 비율은 적정한가?

A/D≒0 ?

C/D, A/C, A/D
D 미포함 이상건수
각단계별 처리시간
A (Anormal)
C (Check)
D (Detection)

FPRI
Future Policy Research Institute

[기계학습을 통한 이상감지]

통계기반 검출(룰/시나리오) & 비지도학습 기반 검출 → 지도학습 기반 검출

이상감지를 위한 비지도학습모형(unsupervised learning)

1) 지도학습을 위한 이상 데이터 수집을 위함

일반적인 행위데이터 수준으로는 다른 정상인것과 뚜렷하게 구분되지 않으므로
데이터 수집대상 검토에 큰 노력을 기울여야 함

→ 비지도 학습모형은 군집분석 수준으로, 군집분석이 해당 조직이나 부서에 적용가능한지 사전 검토가 필수적임
(예시 : 전사단위로는 분석의 의미는 없음, 특정 부서 내 대부분의 부서원이 동일한 패턴으로 업무를 수행하는 경우에만 탐지가능)

* **기계학습으로도 탐지(20~50차원 이상)가 안되는 데이터의 경우, 이는 "단순 룰 기반 탐지시나리오(3~4차원)" 또한 아무 의미가 없음**
→ 이 경우는 컴플라이언스 준수, 임직원 교육/계도 수준으로 리소스 투입을 감소시켜야 함

이상감지를 위한 지도학습모형(supervised learning)

1) 기존에 발견된 사례 수집 후 데이터셋 구축
2) 판별하는 예측 모델 학습

훈련데이터(Training set) 확보 어려움 : 즉 일반적인 행위데이터 수준으로는 다른 정상인것과 뚜렷하게 구분되지 않는다.
* 구분할 수 있는 명확한 변수(variable), 자질(feature)을 찾기 어려움
→ 이상행위를 판별할 수 있는 정보가 입력값으로 추가되도록 하여야 함
(예시 : **상담원의 경우 CTI, ARS 이력,**
임직원의 경우 업무수행요청과 관련된 정보)

위의 문제로 흔히 말하는 분류(classification)에 사용하는 Decision Tree, SVM, Regression, DNN 같은 알고리즘들은 이상징후감지에 사용하기 위해 많은 준비과정이 필요함
→ **최대한 다양한 데이터를 데이터셋으로 준비하여 향후 지도학습에서 활용할 수 있도록 저장하여 두어야 함**
→ 위의 방법으로 실패하는 경우, 정보가 부족하여 기계학습 분석이 불필요한 상황임. (* 분석가능성 자체가 없는 노이즈 데이터로 볼 수 있음)

FPRI
Future Policy Research Institute

[업무패턴 예시 – 월1회 분석의 경우]

Case	10월	11월	12월	1월	2월	3월	4월	5월	6월	7월	8월	9월	10월	11월	12월
A1	1	1	50	5	5	1	1	1	1	1	1	1	1	1	50
A2	1	1	100	5	5	1	1	1	1	1	1	1	1	1	100
A3	1	1	100	5	5	100	1	1	100	1	1	100	1	1	100
A4	1	1	100	5	5	1	1	1	100	1	1	1	1	1	100
B1	1	1	1	50	50	50	50	50	50	50	50	50	50	50	50
B2	50	50	50	5	5	0	0	0	0	0	0	0	0	0	0
B3	50	45	100	50	45	55	50	45	200	50	45	55	50	50	1400

□ **탐지률 - 직전 2개월 평균 30배의 경우**

Case A1~2 (매년 1회 12월 업무수행 경우) : **12월 오탐, 1,2월 미탐**

Case A3~4 (매년 분기, 반기 업무수행 경우) : **매 분기 or 반기 오탐**, **1,2월 미탐**

Case : B1~2 부서이동의 경우: 부서이동후 일정기간 오탐 또는 **미탐**

Case : B3 매월 동일업무 수행의 경우 : 임계치 미만 확인시 **미탐 발생**

업무 데이터

: 개인, 부서별 업무특성 사전확인/기록 후 반영 필요
인사이동, 장기휴가 등 업무특성에 영향을 줄 수 있는 근태정보 반영 필요
해당 기록의 카운트 원인이 되는 정보를 인터뷰 후 최대한 반영
* 업무요청기록, 콜상담 기록 등

기술적 데이터

: 보안솔루션, 업무시스템 등에서 추출 가능한 부가정보를 Factor로 활용
* SQL 정보추출, USB, 출력기록 등

심리 분석 데이터

: 위험수준 변화 판단을 위한 퇴사예정 여부, 연봉변화, 스트레스지수 등 반영필요

이상징후 시스템 관리

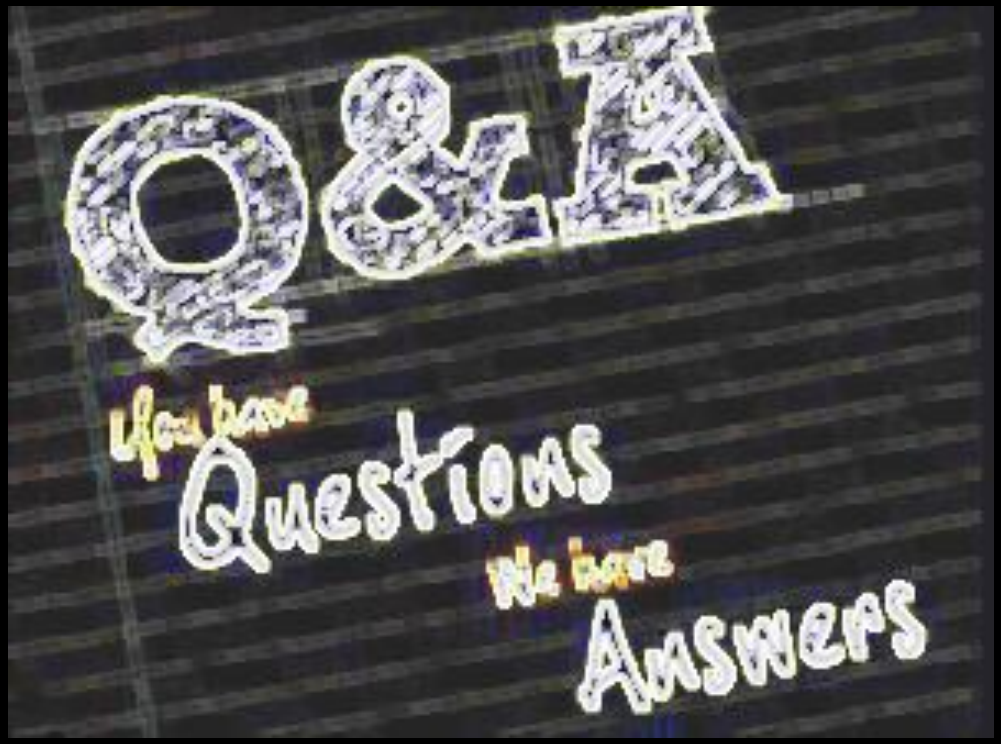
: C/D, A/D Ratio 관리를 통한 시스템 수준 진단 필요

FPRI
Future Policy Research Institute

[참고자료]

우발과 패턴(Ubiquity), 마크 뷰캐넌
인투더데이터, (intothedata.com)

FPRI
Future Policy Research Institute



"만약 당신이 미래를 꿈꾸지 않거나 지금 기술개선을 위해 노력하지 않는다면 그건 곧 낙오되고 있는 것이나 마찬가지 입니다."

그윈 쇼트웰(Gwynne Shtwell, SpaceX CEO, COO)

감사합니다

(facebook.com/sangshik, mikado22001@yahoo.co.kr)