

## D (토론 1 - 6 조, 토론 2 - 5 조)

토론 1 - 6 조 5.77 3월 31 일

### 문제 정의 분석 보고서

#### [데이터 설명]

- 반도체 제조 공정 데이터인 SECOM 데이터를 사용.
- 측정 시간, 공정 센서 데이터, 양품/불량품 여부에 관한 1567개의 object를 가짐.  
-> 데이터의 개수가 많지 않으므로 sample을 함부로 삭제하지 않아야 함.
- 591개의 feature을 가지는 다변량 데이터이고 측정 시간을 제외한 독립변수명 공개하지 않음.  
-> 분류에 악영향을 주거나 무의미한 feature을 제거 통해 차원을 감소시켜 overfitting 방지해야 함.
- Time 변수를 제외한 나머지 독립변수들은 수치형 변수이며, 종속변수는 양품(1) 및 불량품(-1)으로 구성.
- 이 데이터를 통해 가장 높은 분류 성능을 가지는 모델을 구축하고, 각 모델에서 제공하는 도구를 이용하여 중요한 신호를 파악.

3

데이터에 대한 설명이 상세합니다.

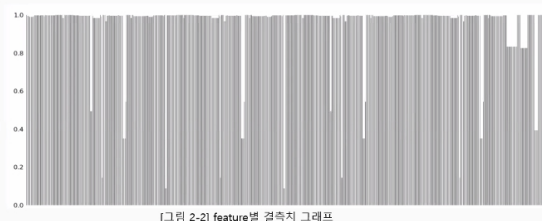
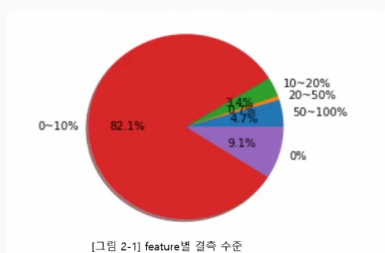
对数据的说明详细。

### 문제 정의 분석 보고서

#### 1. 결측치

해당 데이터의 feature에서 많은 결측치가 있음을 발견.  
결측이 없는 feature의 비율은 9%로 결측치가 있는 열을 모두 제거할 경우 데이터의 손실이 큼.  
즉, 결측치 수준에 따른 결측치 처리 방법을 다르게 결정.

대체 : 결측치를 유의미한 관측치로 대체  
삭제 : 결측치의 비율이 매우 높은 feature을 제거



5

데이터 결손값 구분이 상세합니다.

数据缺失值划分细。

## 문제 정의 분석 보고서

### 2. Feature 별 분포 종류

[그림 3-1]

[그림 3-2]

[그림 3-3]

**- 신호 = 의미 있는 정보 + 의미 없는 정보 + 노이즈**

- 이상치가 다수 있는 feature set  
 -> outlier이지만 의미 있는 정보라고 판단되고, 변수명도 알지 못하므로 함부로 대치, 삭제 하기 보다는 유지.
- feature가 근접 공분산(near-zero variance) 인 경우  
 -> 종속변수에 영향을 끼치지 않는 의미 없는 정보라고 판단되어 feature 삭제.
- 분류에 악영향을 주는 노이즈(ex) 측정 에러 등  
 -> 노이즈 대치.

6

구체적으로 삭제된 데이터를 예시합니다.

例举具体删除的数据。

토론 2 - 5 조 5.47 5월 4일

**결측치**

- 55%의 threshold를 두고 결측치가 많은 열들을 제거  
 그 후 mean heuristic과 nearest neighbor heuristic 기법 이용
- EM 알고리즘 이용 (Cluj-Napoca, Romania, 2017)

Zoom 회의

장혜연 / Zhang Huiyan(20#...)

강상엽 / Kang S...

김규원 / Kim...

김병훈 (主持人)

聊天

김민규 / Kim, Min Kyu(20#...08)对所有人说: 안녕하십니까

이강산 / Lee, Kang San(20#...05)对所有人说: 저희 조 4시에 실시간강의가 있어서 다음에 발표해도 괜찮을까요?

发信: 所有人

输入消息...

이용한 기술과 알고리즘을 썼습니다.

写出了利用的技术和算法。

Zoom 会议

장혜연 / Zhang... 김병훈 강상엽 / Kang S... 김규연 / Kim ...

Introduction Data Description Existing Research Plan

Target 불균형

1) SMOTE 기법 사용  
(양진경&이동희&이초희&김광재,데이터 기반의 핵심 반도체 제조 공정 선정에 관한 연구,2017)

2) 생성적 적대 신경망(GAN) 사용  
(김희수&이현수,생성적 적대 신경망을 이용한 반도체 제조공정 데이터의 결측치 추정 및 공정 이상 진단 프레임워크,2018)

Target value

Pass 90.4% Fail 9.6%

Dimension Reduction

1) PCA기법

2) Lasso Regression

3) Top 25% Correlation analysis

4) MeanDiff 필터링 기법 (양진경&이동희&이초희&김광재,데이터 기반의 핵심 반도체 제조 공정 선정에 관한 연구,2017)

8

参会者 (0)

查找参会者

장혜연 / Zhang Huiyan(20#... (我) 解除静音

김병훈 (主持人) 邀请 解除静音

聊天

김민규 / Kim, Min Kyu(20####08)对所有人说 : 안녕하세요

이강산 / Lee, Kang San(20####05)对所有人说 : 저희 조 4시에 실시간강의가 있어서 다음에 발표해도 괜찮을까요?

发送给: 所有人 文件 ...

输入消息...

절차가 상세하게 설명되어 있습니다.

步骤写的详细。

Zoom 会议

장혜연 / Zhang... 김병훈 강상엽 / Kang S... 김규연 / Kim ...

Introduction Descriptive Statistics Existing Research Plan

실험 성능 평가

10-Fold Cross Validation 실험결과와 안정성을 확보하기 위해 10-Fold Cross Validation으로 데이터를 처리하여 진행할 계획

Accuracy & F1-score 일반적으로 분류문제의 성능 척도는 Accuracy를 계산하여 사용 불균형 문제가 심하기 때문에 F1-Score 사용

$$F1-Score = 2 * \frac{1}{\frac{1}{Precision} + \frac{1}{Recall}}$$

불균형 데이터 분류 문제에서 평가적으로 사용하고 있는 기하평균 사용

Geometric Mean  $GM = \sqrt{Sensitivity * Specificity}$

GM은 민감도와 특이도의 곱의 제곱근으로 계산되어 불균형 데이터에도 한 범주에 편향되지 않은 성능척도로 사용될 수 있음

10

参会者 (0)

查找参会者

장혜연 / Zhang Huiyan(20#... (我) 解除静音

김병훈 (主持人) 邀请 解除静音

聊天

김민규 / Kim, Min Kyu(20####08)对所有人说 : 안녕하세요

이강산 / Lee, Kang San(20####05)对所有人说 : 저희 조 4시에 실시간강의가 있어서 다음에 발표해도 괜찮을까요?

发送给: 所有人 文件 ...

输入消息...

계획 중 실험성능을 평가합니다.

在计划中评价实验性能。