실험제목	센서 기반 제조 공정을 위한 인공지능
실험목적	반도체 제조 공정의 센서 데이터를 이용한 생성 공정 관리

배경

경쟁이 치열하고, 제품의 변화가 빠르기 때문에 생산 공정의 유연한 조정과 제품 출시 시간 단축은 매우 중요함.

IoT 센서의 발전으로 제조 공정에서 다양한 데이터 수집이 가능해 졌으며, 그 데이터를 이용한 공정 관리는 복잡 해지고 있음.

제조 공정 관리를 위해서 빅데이터 와 인공지능 기술의 결합을 통해 유연하고 빠른 공정 관리가 가능 할 것으로 예상 됨.

반도체 제조 공정 데이터 (SECOM) 을 이용하여 생산 공정 관리를 위한 기계 학습 기술을 적용할 방법 연구.

실험 설계

1. 데이터 분석을 위한 알고리즘 선택
   - Boruta
   - Multivariate Adaptive Regression Spline (MARS)
   - PCA
2. 분류 알고리즘
   - Logistic Regression (LR)
   - Random Forest (RF)
   - Radient Boosted Trees (GBT)
3. 평가
   - FPR
   - Precision
   - recall
   - F-measure
   - accuracy

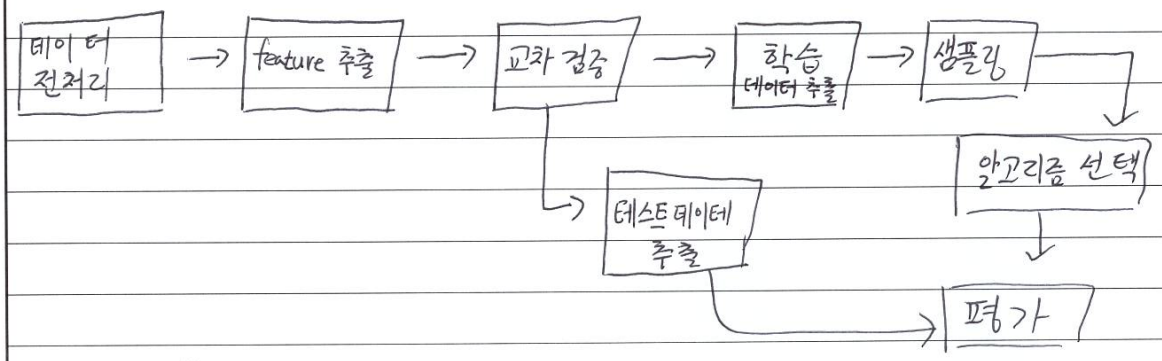
기록자 Written by	점검자 Witnessed or Understood by	점검자 Witnessed or Understood by
김 민 태		
일자 1/1	일자	일자

실험제목	센서 기반 제조 공정을 위한 인공지능
실험목적	반도체 제조 공정의 센서 데이터를 이용한 생성 공정 관리

데이터 분석.

- 100개 이상의 단계로 구성된 반도체 제조 공정에서 수집된 데이터
- 1567개의 샘플 데이터
- 샘플 데이터는 591개의 특성을 가짐.
- 측정이 되지 않은 값이 'NaN'으로 기록
- 제품에 문제가 있는 경우 1, 정상인 경우 -1로 나타냄

실험 설계



1. 데이터 전처리
   - 결측치 제거
   - 데이터 대, 비 정규화
2. feature 추출
   - Bruta
   - MARS
   - PCA
3. 교차검증
   - k-Fold cross validation
4. 샘플링
   - 불균형 데이터
   - 오버 샘플링
5. 알고리즘 선택
   - RF
   - LR
   - GBT

기록자 Written by	점검자 Witnessed or Understood by	점검자 Witnessed or Understood by
김 민 태		
일자 1/8	일자	일자

실험제목	센서 기반 제조 공정을 위한 인공지능
실험목적	반도체 제조 공정의 센서 데이터를 이용한 생산공정 관리

데이터 전처리

-55% 이상 데이터가 결측된 경우 해당 특성 제거

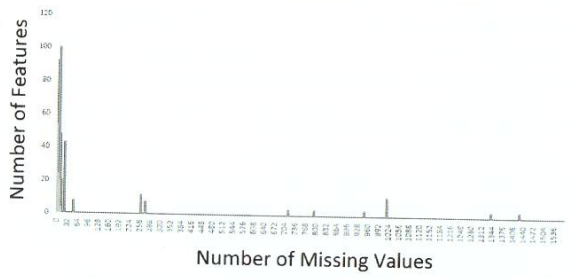


Fig. 2. Features Distribution by the Number of Missing Values

- 각 샘플 데이터에서 결측값 수를 시각화
- 591개의 특성중 24개가 제거 됨.
- 567개의 특성을 고려하여 사용

-'NaN' 값은 전, 후 값이 정상 치일 경우 전, 후 값의 평균 이용

- 전, 후 값 중 'NaN'이 존재할 경우 정상치를 그대로 이용

- 상수 값을 제거 하여 총 116개의 특성이 추가로 제거됨

- 451개의 특성이 남으며, 다음과 같은 식을 통해 0~1로 정규화

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

~~feature~~ 추출

-분류에 영향을 미치지 않는 특성 제거

TABLE II
COMPARISON OF THE FEATURES SELECTION TECHNIQUES

Features Selection Algorithm	Number of Selected Features	Characteristics
Boruta	22	by default it uses the Random Forest (RF) algorithm
MARS	10	it is a form of regression analysis
PCA	111	it uses orthogonal linear transformations

- Boruta는 76번의 반복 수행을 통해 22개의 특성 추출
- MARS는 RSS와 GCV를 기반으로 10개의 특성을 추출
- PCA는 111개의 특성을 추출

기록자 Written by	점검자 Witnessed or Understood by	점검자 Witnessed or Understood by
김 민 태		
일자 1/15	일자	일자

실험제목	센서 기반의 제조 공정을 위한 인공지능
실험목적	반도체 제조 공정의 센서 데이터를 이용한 생산 공정 관리

샘플링

- 소수 클래스와 다수 클래스의 불균형 문제
- 1567개의 데이터 중 104개는 실패, 1463개는 성공 데이터
- 1:14의 비율로 불균형 데이터임.

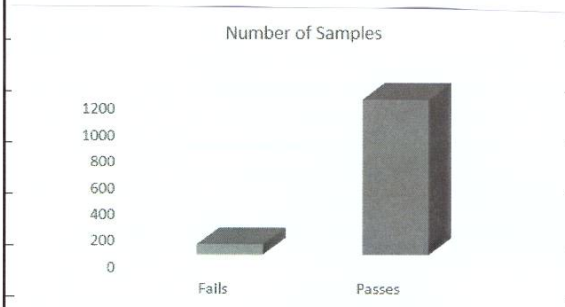


Fig. 6. Initial Distribution of the Fail Samples and the Pass Samples

- 불 균형을 해결하기 위해 다수 클래스에서 무작위로 데이터를 뽑는 언더 샘플링 이용
- 소수 클래스를 복사하는 오버 샘플링 이용

- 언더 샘플링에는 Spread Sub Sample 필터를 이용
- 오버 샘플링에는 Synthetic Minority Oversampling Technique 이용

데이터 분할

- 학습 데이터와 테스트 데이터를 ~~1:4~~ 4:1 (80% : 20%)로 분할
- 모델 검증을 위해 교차 검증을 이용
- k-fold cross validation 이용 하며, 5개로 분할하는 5-fold cross validation 이용
- 모델 평가에는 F-measure, recall, Precision, Accuracy, False Positive Rate (FPR)을 이용
- 테스트 데이터에서 과적합 문제를 검증하기 위해 샘플링은 적용하지 않음.

기록자 Written by	점검자 Witnessed or Understood by	점검자 Witnessed or Understood by
김민태		
일자 1/22	일자	일자

실험제목	센서 기반 제조 공정을 위한 인공지능
실험목적	반도체 제조 공정의 센서 데이터를 이용한 생산 공정 관리

분류 알고리즘

- LR은 Broyden - Fletcher - Goldfard - Shanno (LBFGS) 알고리즘 이용

- Random Forest 는 다음과 같이 구성

TABLE III
RANDOM FOREST PARAMETERS

Parameter	Value	Description
number of classes	2	the number of classification classes
number of trees	20	a parameter that describes how many trees are in the forest
features subset strategy	"auto"	how many features to use as candidates for splitting
impurity	"giny"	a parameter that describes the homogeneity of the labels
maximum depth	20	the maximum depth of the trees from the forest
maximum number of bins	100	the maximum number of bins that are used to discretize the features

- GBT는 다음과 같이 구성

TABLE IV
GRADIENT BOOSTED TREES PARAMETERS

Parameter	Value	Description
number of classes	2	the number of classification classes
number of iterations	3	each iteration creates a tree
maximum depth	5	the maximum depth of the trees

샘플링 방법

TABLE VII
SAMPLING METHOD

Sampling Method	Description
Unsampled	the number of samples is unchanged
Undersampled Majority Class	the number of samples of the majority class is reduced to 78
Oversampled Minority Class	the number of samples of the minority class is increased to 1176

기록자 Written by	점검자 Witnessed or Understood by	점검자 Witnessed or Understood by
김민태		
일자 1/29	일자	일자

실험제목	센서 기반 제조 공정을 위한 인공지능
실험목적	반도체 제조 공정의 센서 데이터를 이용한 생산 공정 관리

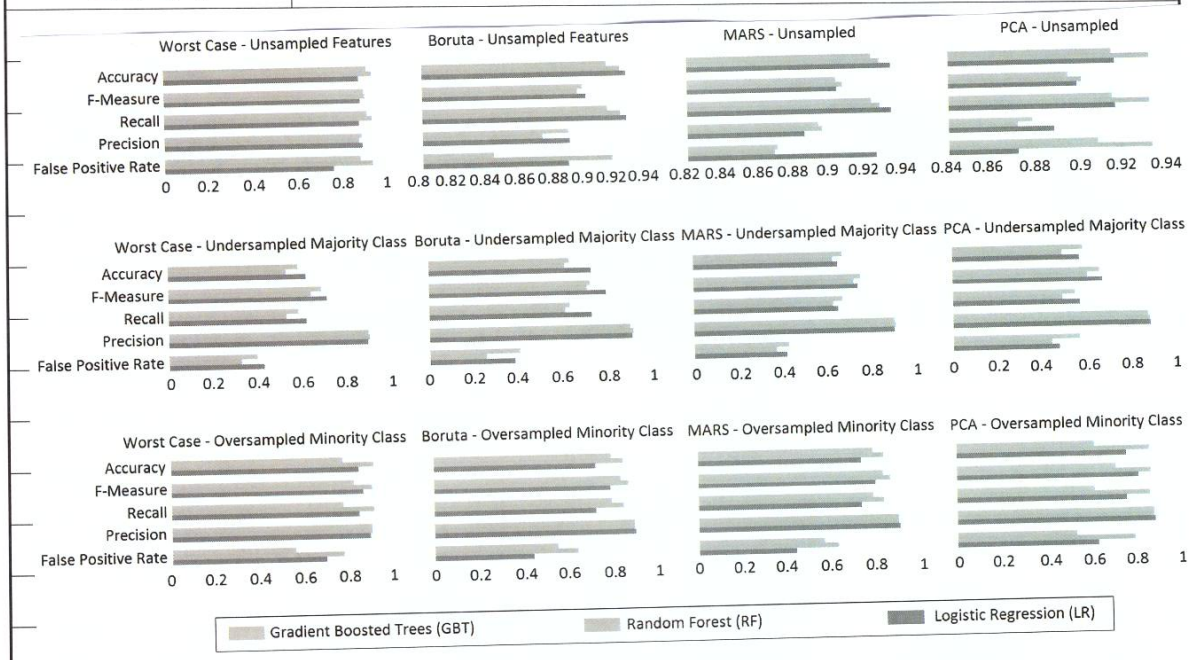


Fig. 8. Experimental Results Summary - Classification Algorithm Perspective

실험결과

- 데이터를 샘플링 하지 않은 경우 Boruta 알고리즘을 통해 특성을 추출 한뒤 LR을 적용 하였을 때 정밀도가 0.89440으로 가장 높게 나타남
- 다수 클래스 언더 샘플링 시 Boruta 알고리즘과 RF 결합시 정밀도가 0.90506으로 가장 높게 나타났음
- 소수 클래스 오버 샘플링 시 MARS 알고리즘과 LR 이용시 0.91541로 가장 높게 나타났음
- FPR을 기준으로 평가시에도 정밀도와 동일한 결과가 나타남

결론

- 대부분의 경우 Boruta 알고리즘과 RF를 이용한 경우 높은 성능이 나타났음

기록자 Written by	점검자 Witnessed or Understood by	점검자 Witnessed or Understood by
김민태		
일자 1/24	일자	일자