지난 발표 자료 정리

1567개의 데이터 -> 변수 590개 -> PASS(1463/93.36%), FAIL(104/6.64%)

상대적으로 FAIL에 대한 데이터가 부족함 (약 15배 차이)

→ Over/Under sampling 을 통해 비율을 맞추는 방식을 고려

이상치 기준 ---> 평균 기준

이상치 처리 → 값을 대체 하는데…. 기준이..?

이상치가 15%를 넘어가는 열이 존재, 두 열을 제외하면 이상치는 평균에서 2%내외 분산

결측치 처리 → 결측치 50% 이상은 제거 / 나머지는 imputation

Imputation은… 기존 데이터 값을 바탕으로 추정치와 표준오차를 평균 공식을 통해 모수 세트를 만들어서 처리

모델링에 유용하지 않은 분산이 0이거나 0에 근접한 열을 제거하도록?

→ 왜?

상관관계, 다중공선성이 높은 특성끼리 페어 (변수변환)

======================================================================

발표 과제 (4단계)

1. 무의미한 변수 제거 (질문완료/답변대기)
2. 결측치 처리
3. 데이터간 병합
4. 변수처리
5. 범주형 변수를 가변수로 변환 (dummy variable)
6. 변수 간 scale 표준화 혹은 정규화

======================================================================

결측치 처리와 변수 스케일링에 초점을 맞추어 정리해 봤습니다.

< 생성적 적대 신경망을 이용한 반도체 제조공정 데이터의 결측치 추정 및 공정 이상 진단 프레임워크 >

1. 서론 생략

2. 배경 및 문헌연구

2.1 결측치 처리 연구

반도체 공정은 높은 복잡도를 가짐. 특히 수율 관리부문에서 대부분의 문제가 발생. 따라서 높은 수율 획득을 위해 선제적인 조치가 필요함. 대부분의 공정은 센서로부터 공정변수를 일정 시간 간격으로 측정, 모니터링 한다.

결측값 처리는 대부분 대체 값을 채우거나, 제거하는 것으로 해결한다. 불완전한 속성을 제거하는 방식은 결측된 수가 상대적으로 적고, 영향을 거의 미치지 않을 때 사용하며, 대체적으로는 결측치 값의 추정 및 대체방법을 사용한다.

일반적인 결측치 보정방법으로는 평균/중앙값 대체, k-Nearest Neighbor (K-NN), 인공신경망(Neural Network), 회귀 모델 (Regression)등을 통한 대체 방법 등도 있다.

[ Yuan - 12 ] 단일 추정값으로의 대체가 아닌 불확실성을 가진 대체 방법인 Multiple Imputation을 제안함.

[ Kerdprasop - 13 ] 데이터마이닝 기법 중 하나인 Problem Rule Induction Method (PRIM)에서 누락된 값의 체계적인 처리와 프로세스 개선을 위한 Missing-value PRIM을 제안.

본 논문에서는 GAN 구조에서의 학습과 역전파 알고리즘을 통해 결측치를 처리하는 방법을 제안하고, 그 유효성을 증명한다.

+) GAN – 생성적 적대 신경망은 주로 비지도 학습에 사용됨

표 1. 이상감지 및 분류 프레임워크와 방법 비교

* 두번째 행만 Imputation을 했습니다.

그림 1. 결측치 보정 및 오버샘플링을 통한 이상감지 및 분류 프레임워크

* 이건 전체 흐름에 도움이 되는 것 같아서..허허

2.2 오버샘플링을 통한 이상감지 및 분류

실제 현장에서 얻어진 데이터는 일부 클래스에 편향되어 있는 경우가 많다. 분류 모델을 구축하기 위해서는 데이터 불균형 문제를 해결하는 것이 중요한 선결조건이 된다.

Son.Ko와 Kim(1) 은 오류검출 및 분류 (Fault Detection and Classification : FDC)를 위해 공정결과와 연관성이 높은 데이터 스트림의 구조적인 특징들을 추출하여 의사결정 트리의 입력으로 사용하여 분류하는 모델을 제안하였다.

표 1에서 보다시피 기존 연구 대부분은 결측치를 고려하지 않았다.

Lee와 Hwang은 GAN을 사용하여 데이터를 오버샘플링한 후 분류 모델에 적용하였지만 결측치에 대한 고려 없이 가상 데이터를 생성하였다.

대부분의 기존 문헌 연구들은 이상감지 및 분류 문제에서 데이터 불균형을 해결하기 위해  (불량이나 불합격의 소수 데이터의 해결) 오버샘플링 기법을 적용하고 특징을 추출하여 불량의 분류 모델을 제안하였다.

그리고 대부분의 결측치는 비율에 따라 처리 대상에서 제외하거나 남아있는 결측치는 평균/중앙값 대체하는 방법을 사용하였다.

장비나 센서의 오류로부터 발생하는 결측치를 고려하지 않은 오버샘플링을 통해 분류모델을 구축하는 것은 잠재적으로 가치 있는 변수를 탈락시키는 한계점을 지닌다.

(오버샘플링 하기 전에 중요한 변수를 선별해야 한다)

본 논문에서는 신경망으로 구성된 생성모델인 GAN을 사용하고, 학습알고리즘에서 결측치에 대한 처리과정을 접목하여 미세한 결측치의 보정을 통해 오버샘플링 기반의 제품 이상을 분류하는 효율적인 방법을 모색한다.

3. GAN을 활용한 결측치 보정 및 생성을 통한 제조공정 이상탐지 프레임워크

3.1 GAN을 이용한 결측치의 보정을 동반한 소수 클래스의 오버샘플링

1. 데이터 전처리

전처리 과정으로 결측치의 대체 및 데이터의 정규화를 수행한다.

데이터 분포에 영향을 미치지 않는 상수 값 형태의 데이터 속성 값을 제거

기존 연구들에서는 결측치가 샘플 수의 45~60%을 넘으면 제거[17.19.22]

(저희는 50%)

1. GAN을 활용한 결측치 보정 및 데이터 생성

GAN 은 생성자 G(Generator)와 구별자 D(Discriminator)의 대립 구조 학습을 통해 훈련데이터와 유사한 자료를 만들어내는 생성모델이다. 이 때 D와 G는 미분가능한 함수로 각각이 특성을 파라미터로 같은 멀티레이어 퍼셉트론이다.

* 역전파 관련 원리는… 잘 모르겠는데… 구현하는 방법 등은 다른 문서에서 잘 나와있습니다…

4. 실험 및 분류 성능 결과

4.1 가상데이터로 학습한 분류기의 성능

이 연구에서 사용한 데이터는 SECOM데이터 셋이다. 개체 수 1567개 특징벡터 591개이며 각 개체의 판별은 합격과 불합격 두 가지로 나뉜다. 본 연구에서는 전 처리 과정에서 상수값을 가지는 특징벡터 117개를 제거하고 총 474개의 특징만을 고려하였다.

이 데이터는 합격데이터 1463개, 불합격 데이터 104개로 데이터의 불균형 문제를 가지고 있으며, 이러한 불균형 문제 및 결측치를 위하여 본 연구에서 제안한 GAN을 통한 데이터 생성 방법을 적용하였다.

하이퍼 파라미터인 학습률은 0.00001 모멘텀계수는 0.5 구별자 D에서의 활성화 함수는 계수 2를 가진 relu함수 미니배치는 4로 설정하였다.

결측치의 미세 보정을 거친 생성모델이 목표 결과에 더 수렴하고, 실제 목표 값에 가깝게 결과를 내는 것을 알 수 있다.

4.2 기존 SECOM데이터를 활용한 기존 연구들과의 비교

SECOM활용 데이터 활용 연구에서 가장 좋은 분석결과를 보여주었던 군집 기반의 특징추출 기법인 MeanDiff방법과 의사결정트리를 결합한 기법이다. 비교 대상 방법에서는 결측치 55% 이상의 속성은 제거하였고, 특징 168개만을 선택하고 소수 그룹의 샘플 수를 복제하여 오버샘플링 한 후 알고리즘에 적용하였다

표3. 각 분류모델의 혼동행렬

표4. 분류모델의 성능평가 비교

성능평가를 보았을 때, TPR (True Positive Rate)은

Logistic Regression (1.0) == Decision Tree (1.0) > k-NN (0.983) > **Proposed Model (0.952)** > Naïve Bayes (0.746)으로 기존 방법에 대해 약간 낮은 수치를 보여주지만

**FPR (False Positive Rate)는 0, Precision은 1, F-Score 0.976** 이라는 압도적인 수치를 보여주었다.

<군집화 및 특성도를 이용한 결측치 대체 방법 – 요약>

여기서 제시한 방법인 MCMC의 방법은 변동계수가 크며, 연관성이 높은 경우, 결측 비율이 증가함에 따라 기존 방법에 비해 제안 방법이 조금 더 우수함을 보였고, 변동 계수가 작으며 연관성이 높은 경우에는 MCMC가 우수하나 결측 비율이 증가함에 따라 우수성이 점점 떨어졌다.

🡪 “연관성이 높은 경우” 기 연구에 비해 우수함

🡪 변동계수가 큰 경우, 결측 비율이 증가할수록 우수함이 더 잘 나타났다.

몬테카를로 모의실험을 통해 결측치를 대체했고, RRA, SSP, NRMSE를 이용하여 MCMC와 기존 방법들의 성능을 비교하였다. 특정한 조건이 성립되는 경우 우수한 부분이 있으나, 초기 시점의 데이터가 매우 중요하게 작용하는 점을 고려하면 이번 SECOM에서는 사용하기 조금 까다로울 것 같습니다.

<Weighted k-Nearest Neighbors를 이용한 결측치 대치 – 요약>

이 연구에서 제안하는 가중 k-최근접 이웃 대치법은 숫자형으로 이루어진 자료에서 유사성 거리를 반영하여 계산된 가중치를 이용하고, 결측치를 가중 평균값으로 대치하는 방법이다.

이는 k개의 최근접 이웃들에 대한 거리에 반비례하여 거리가 가까운 최근접 이웃들에 대해서는 큰 가중치를 부여하고, 상대적으로 거리가 먼 최근접 이웃들에 대해서는 작은 가중치를 부여하는 것이다.

가까운 이웃들을 찾기 위해서 거리를 계산하고, 그 거리에 따른 가중치는 커널 함수를 이용하여 계산하며, 유사성 측정을 위한 거리함수는 Euclidean distance와 Hellinger distance를 사용함,

본 논문에서는 KNN 대치법의 단점을 보완하기 위해 WKNN 대치법을 제안하였다. 이 대치법은 거리와 커널 함수에 따라 4개로 나누고, KNN은 거리에 따라 2가지로 나눠서 총 여섯 방법을 실험을 통해 비교해 보았다.

결론적으로 최근접 이웃의 개수, 결측비율, 표본에 크기에 상관없이 **KNN 방법보다 WKNN방법이 더 뛰어나다는 것**을 확인할 수 있었다.