2023 제조 빅데이터 분석 경진대회
결과보고서

2023.10.03

팀명 : 박광진

Ⅰ. 분석과제 개요

1. 분석 대상

1) 짝맞춤 조립

본 과제의 목표는 튜브와 샤프트의 치수가 주어졌을 때, 최적의 ball size를 제안하는 것으로 제품의 정밀도를 향상시키기 위해 짝맞춤 조립방식을 적용하고 있다.

짝맞춤 조립방식은 여러 개의 핵심부품을 조립하여 완성되는 제품을 생산할 때 많이 사용되는 제조 기법으로 각 부품의 치수 정밀도를 낮추어 부품의 가격을 떨어뜨리는 대신 조립 시에 부품을 잘 조합하여 제품의 품질을 유지시키는 방법이다.

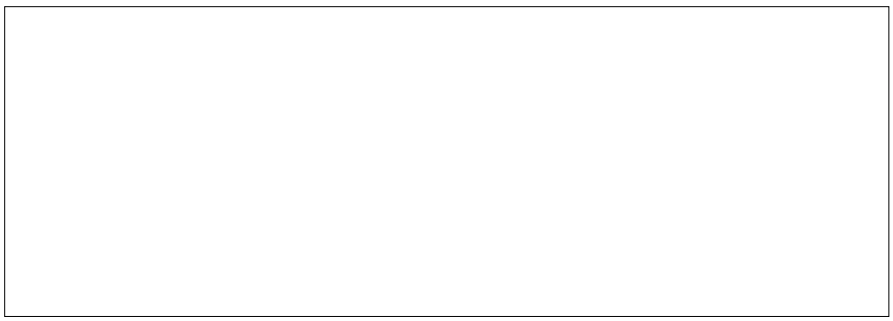

Figure 2. 부품의 구성

2) 치수분할방식

짝맞춤 조립의 대표적인 유형으로 치수분할방식이 있는데 이는 생산된 부품 치수에 의해 각기 여러 범주로 분류한 후 대응되는 범주에 속한 부품들끼리 짝을 맞추어 조립하는 방식이다. 치수분할방식은 대표적으로 등간격분할방식(erqual width partitioning)과 등면적분할방식(equal area partitioning)이 있다. 전자는 치수를 측정하여 범주로 분류할 때 각 범주의 치수 폭이 동일하도록 하는 방식이고 후자는 각 범주에 속할 기대 부품 수를 동일하도록 하는 방식이다.

2. 현상 및 문제점

[ ] 산업에서도 치수분할방식으로 이를 해결하고자 짝맞춤 조립표를 만들어 사용하고 있으나 정확도가 보장되지 않고 조립표를 확인하는 시간이 길어 거의 사용되지 않고 있다. [ ] 산업에서 제공받은 데이터와 비교했을 때에도 짝맞춤 조립표의 부품별 grade 기준이 실제 부품 치수와 맞지 않은 경우가 대부분이었으며 1296개의 데이터 중 짝맞춤 조립표에 의해 ball size가 지정되는 경우는 190개에 불과했다. 따라서 기존 짝맞춤 조립표의 분류 방식이 적절하다고 할 수 없으며 최적의 분류 방식을 찾고 이를 검증할 필요가 있다.

Ⅱ. 데이터 분석

1. 데이터 분석 방법 및 제반 가정

본 분석에서는 짝맞춤 조립표의 정확도를 높이기 위한 최적 분류 방법을 찾고자 큰 규모의 시뮬레이션 데이터를 생성하여 치수분할방식별로 범주 개수에 따라 정확도가 어떻게 변화하는지 정밀하게 비교하였다.

기존 짝맞춤 조립표에서 tube와 shaft의 치수가 over-ball-diameter(OBD) 방식으로 측정된 중간 부분의 치수(M 위치의 평균값)로 정의되어 제작되었기 때문에 본 분석에서도 동일하게 정의하고 사용하도록 하겠다.

시뮬레이션 데이터는 tube와 shaft의 치수 분포가 서로 독립이며 정규분포를 따른다고 가정하고 100,000개의 난수를 무작위로 추출하여 얻었다. 이때 random seed는 임의의 수(1906)로 고정하여 같은 데이터에 의해 정확도가 계산되도록 하였다.

시뮬레이션 과정에서 정확도 계산에 필요한 최적 ball size는 기존 짝맞춤 조립표에 의해 얻어진 선형회귀식이 정확하다는 가정 하에 계산되었으며 이에 대한 자세한 설명은 2.2)에서 하도록 하겠다. 베어링 볼은 0.002mm 간격으로 생산되며 소수점 아래 셋째 자리가 짝수라 가정하고 ball grade는 소수점 아래 셋째 자리에서 짝수를 갖도록 변환(홀수인 경우 올림, 짝수인 경우 버림) 해주었다. 또한 예측한 ball grade가 최적 ball grade로부터 ±0.002mm 이내에 포함되는 경우 양품이라고 간주하였다.

평가 지표는 예측 ball grade와 최적 ball size의 mse, 최적 ball grade로 분류된 비율인 1st accuracy, 양품으로 분류된 비율인 2nd accuracy 총 3가지를 사용하였다. 시뮬레이션에서 각 부품별 범주 개수는 2~20개로 제한하였고 2nd accuracy가 90%가 넘으면서 범주의 개수가 가장 적은 분류 방식을 최적 분류 방법이라 정의하였다.

2. 데이터 분석 절차와 내용

1) simulation data 생성

시뮬레이션 데이터를 생성하기 위해 태림 산업에서 제공받은 실제 제품 치수 데이터를 이용해 tube와 shaft 치수 분포를 확인해 보았다. figure 2에서 대각 영역은 Kernel Density Estimation(KDE)[1]을 이용해 데이터로부터 tube와 shaft 치수의 분포를 추정한 KDE plot, 비대각 영역은 tube와 shaft 치수의 산점도이다. KDE plot을 통해 tube와 shaft 치수가 정규분포와 유사한 형태를 띄고 있음을 확인할 수 있다.

따라서 시뮬레이션 데이터는 실제 제품 치수 데이터의 평균과 표준편차를 반영하여 tube 치수는 $N($[redacted]$)$[2], shaft 치수는 $N($[redacted]$)$[3]를 따르는 난수 100,000개를 무작위로 추출하여 생성하였다.

---

1) smoothing 기법을 이용해 관측된 데이터의 분포로부터 모집단 분포 특성을 추정하는 방법
2) 평균이 [redacted], 표준편차가 [redacted]인 정규분포
3) 평균이 [redacted], 표준편차가 [redacted]인 정규분포

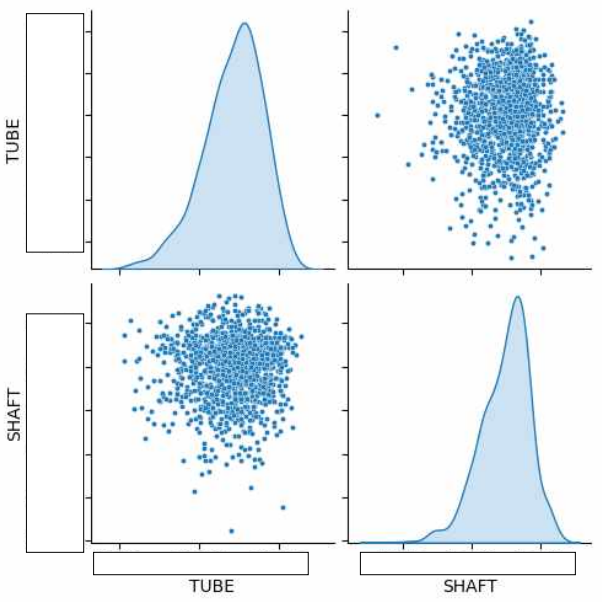

Figure 17. Tube와 Shaft의 치수 분포

2) 최적 ball size 계산

기존 짝맞춤 조립표(figure 3)를 통해 shaft 치수가 0.004mm 증가하면 0.002-0.003mm 작은 ball size를 사용하고 tube 치수가 0.01mm 증가하면 0.002-0.003mm 큰 ball size를 사용하는 규칙성을 확인할 수 있다. 이에 ball size가 tube 치수, shaft 치수와 선형 관계를 갖는다고 판단하고 기존 짝맞춤 조립표를 만족하는 임의의 데이터를 생성하여 선형 회귀를 통해 그 관계를 모델링하였다.

Figure 19. 기존 짝맞춤 조립표

모델링에 필요한 데이터는 기존 짝맞춤 조립표에서의 최솟값과 최댓값을 갖고 범위 내에서 균일하게 분포되도록, 즉 tube 치수는 $U($ [ ] , [ ] $)$[4], shaft 치수는 $U($ [ ] , [ ] $)$[5]를 따르는 난수 100,000개를 생성하였다. ball size는 기존 짝맞춤 조립표에 의해 지정한 후 결측치가 있는 행(짝맞춤 조립표에 의해 ball size가 지정되지 못한 경우)은 제거하여 선형회귀 모형에 적합시켰다. 그 결과는 figure 4에서 확인할 수 있으며 그에 따른 회귀식은 아래와 같다.

$$Ball\ size = 12.0385 + 0.2072 \times Tube\ size - 0.4534 \times Shaft\ size$$

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.948
Model:	OLS	**Adj. R-squared:**	0.948
Method:	Least Squares	**F-statistic:**	4.474e+05
Date:	Thu, 14 Sep 2023	**Prob (F-statistic):**	0.00
Time:	10:34:28	**Log-Likelihood:**	2.5651e+05
No. Observations:	49051	**AIC:**	-5.130e+05
Df Residuals:	49048	**BIC:**	-5.130e+05
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	12.0385	0.015	790.348	0.000	12.009	12.068
Tmean	0.2072	0.001	300.404	0.000	0.206	0.209
Smean	-0.4534	0.000	-933.245	0.000	-0.454	-0.452

Omnibus:	912.151	**Durbin-Watson:**	1.996
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	579.440
Skew:	0.128	**Prob(JB):**	1.50e-126
Kurtosis:	2.533	**Cond. No.**	7.78e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.78e+04. This might indicate that there are strong multicollinearity or other numerical problems.

Figure 28. OLS Regression Results

3-1) 등간격분할방식

등간격분할방식은 figure 5[6]에서 보는 바와 같이 부품 A와 B를 치수 기준으로 분류되는 범주의 폭이 동일하도록 분할하는 방식이다.

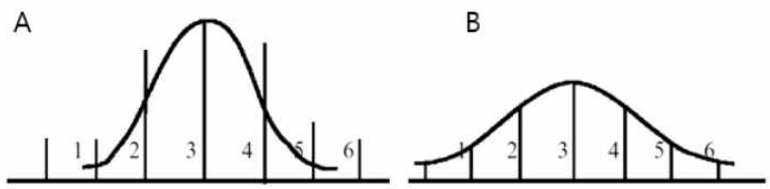

Figure 29. Equal width partitioning for components A and B

(ⅰ) 범주 경계값 설정

부품의 분포는 정규분포를 따른다고 가정하였으므로 값의 범위가 $(-\infty, \infty)$가 되기 때문에 범주를 분할하기 위한 상·하한 지점을 정해야 한다(상·하한 지점을 넘어가는 경우 가장

4) 최솟값이 [ ] 최댓값이 [ ] 인 균일분포
5) 최솟값이 [ ] 최댓값이 [ ] 인 균일분포
6) 권혁무, 이영준, 이민구, 홍성훈. (2017).

말단에 있는 범주에 속하도록 하였다). 상·하한 지점이 말단으로 갈수록 범주 분할 대상이 되는 구간이 길어져 이를 넘어가는 확률이 줄어들기 때문에 대푯값 설정에 반영되지 못할 가능성이 줄어들지만 범주의 폭이 넓어지면서 전체적인 정확도는 떨어질 수 있다는 trade-off가 존재한다. 따라서 상·하한 지점을 $F^{-1}(0.9999)\&F^{-1}(0.0001)$[7], $F^{-1}(0.999)\&F^{-1}(0.001)$, $\mu \pm 3\times\sigma^2$, 즉 구간 밖에 속할 확률이 상·하한 각각 0.0001, 0.001, 0.0013이 되는 지점으로 바꾸어가며 시뮬레이션을 진행하였다.

(ii) ball grade 예측

범주를 분할하고 난 후에는 각 범주에 해당하는 대푯값을 지정하여 **2.2)**에서 적합된 회귀식을 통해 ball size를 예측하고 이를 0.002mm 간격의 ball grade로 변환시킨다. 이때 대푯값으로는 평균값과 중앙값을 사용하여 정확도를 확인하였다.

(iii) 예측 성능 평가

대푯값을 중앙값으로 하고 구간 밖에 속할 확률을 각 0.0001, 0.001, 0.0013으로 변화시켜가며 정확도를 계산한 결과는 figure 6를 통해 확인할 수 있다. 범주 개수를 20 이하로 제한하고 구간 밖에 속할 확률이 순서대로 0.0013, 0.001, 0.0001일 때 높은 정확도를 보였다.

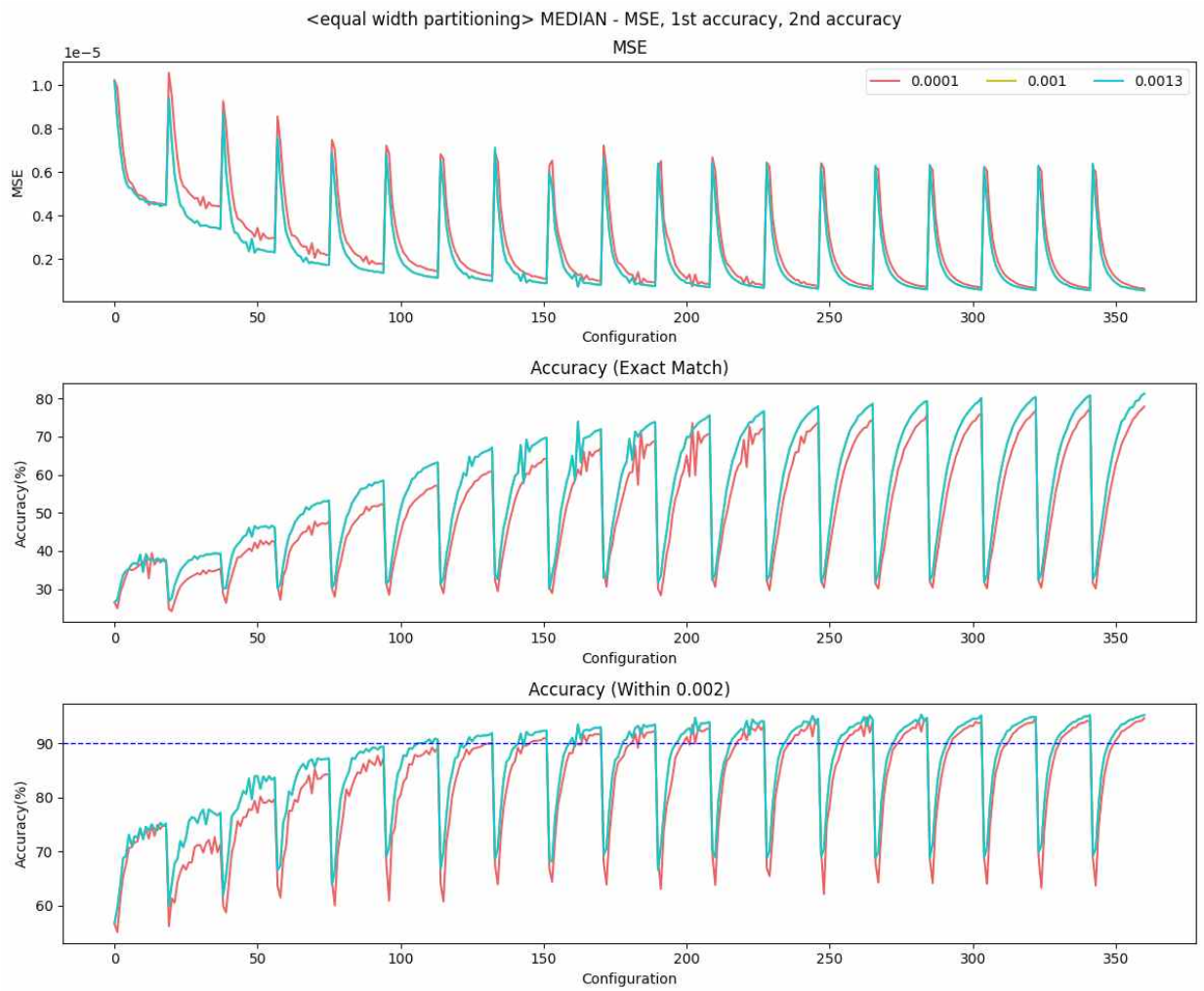

Figure 30. equal width partitioning with median

7) F는 역누적분포함수

대푯값을 평균값으로 하고 구간 밖에 속할 확률이 각 0.0001, 0.001, 0.0013으로 변화시켜 가며 정확도를 계산한 결과는 figure 7을 통해 확인할 수 있다. 중앙값과 마찬가지로 범주 개수를 20 이하로 제한하고 구간 밖에 속할 확률이 순서대로 0.0013, 0.001, 0.0001일 때 높은 정확도를 보였다.

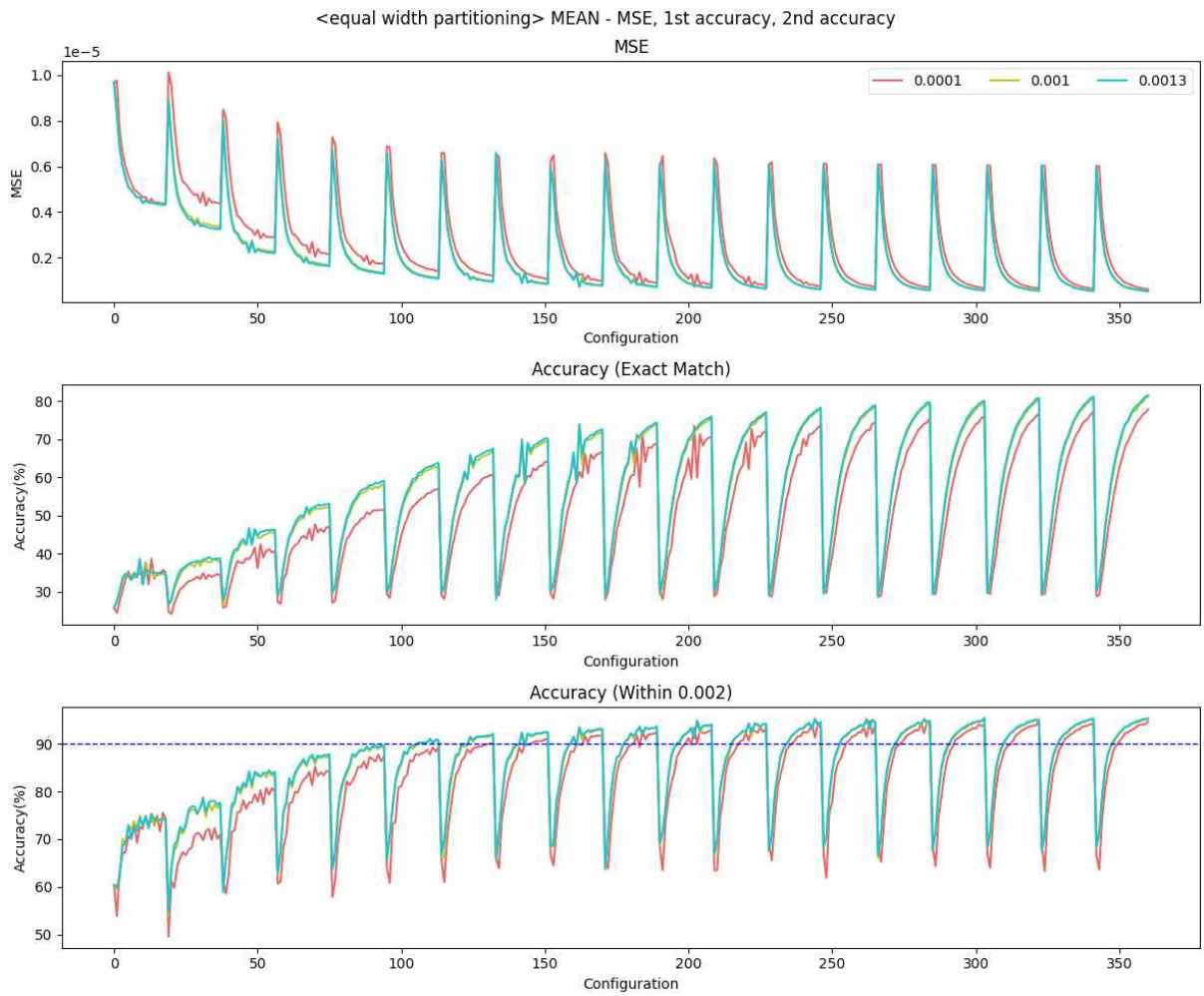

Figure 31. equal width partitioning with mean

중앙값과 평균값의 경우 모두 구간 밖에 속할 확률이 각 0.0013일 때 가장 높은 정확도를 보여 그 두 가지 경우를 비교해 보면 figure 8와 같다. 대체로 큰 차이를 보이지는 않지만 범주의 개수에 따라 약간의 차이를 보였다. 범주의 개수가 적을 때는 MSE, 1st accuracy, 2nd accuracy 모두 중앙값이 약간 더 좋은 성능을 보였지만 범주의 개수가 더 많을 때는 1st accuracy, 2nd accuracy는 평균값이 약간 더 좋은 성능을 보였다. 그리고 범주의 개수가 더 증가할수록 중앙값과 평균값 간의 차이가 거의 없어짐을 확인할 수 있다.

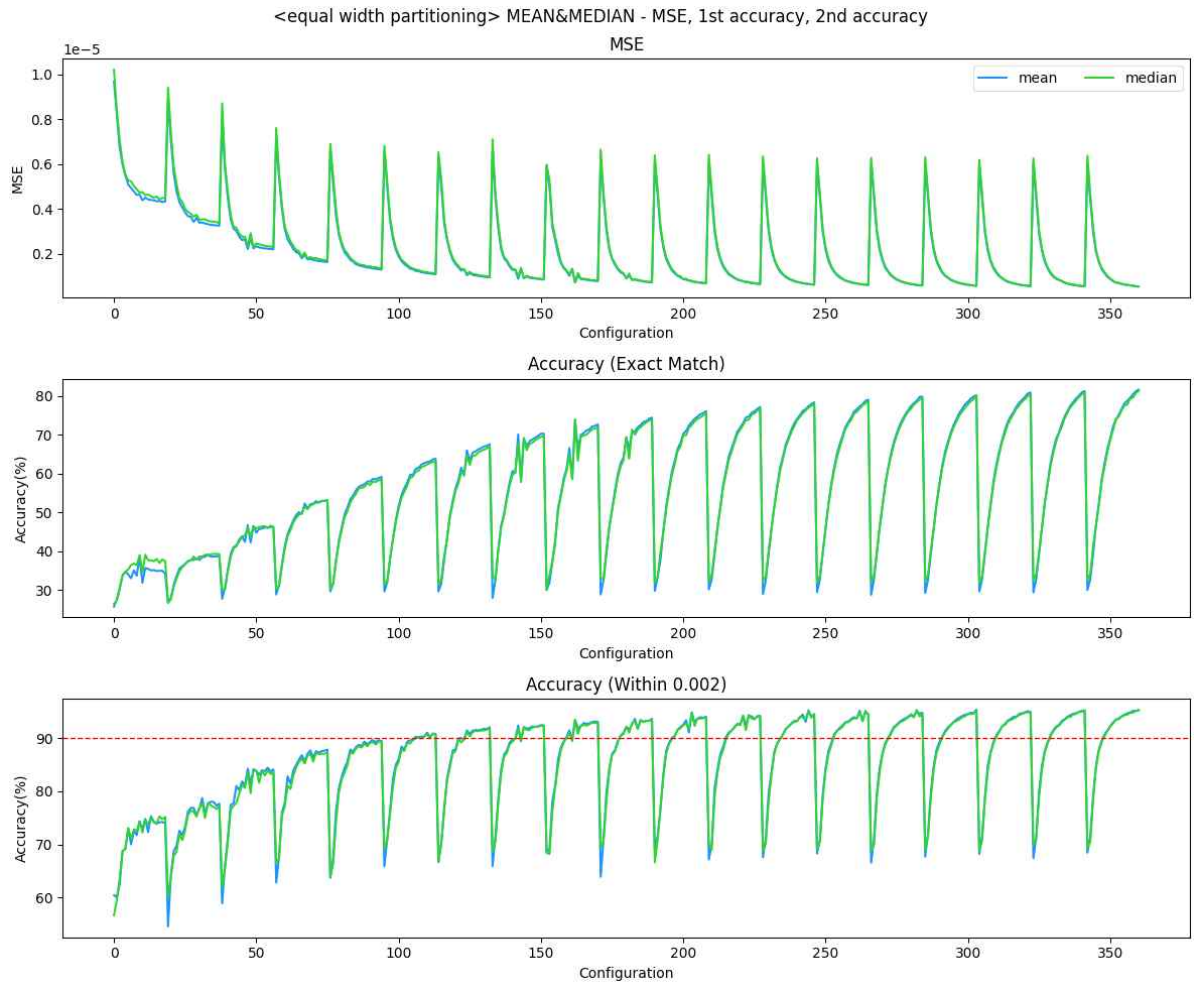

Figure 32. equal width partitioning with mean and median(0.0013)

등간격분할방식 결과 튜브의 범주가 10개, 샤프트의 범주가 8개일 때 2nd accuracy가 90% 이상(90.049%)이며 범주 개수를 최소로 가졌다.

3-2) 등면적분할방식

등면적분할방식(equal area partitioning)은 각 범주에 속할 기대 부품 수를 동일하도록 하는 방식으로 등확률분할방식으로 이해할 수 있다. 등확률분할방식은 두 부품이 각 범주에 속할 확률이 동일하도록 분할하는 방식이다. figure 9에서 확률밀도함수를 각 구간별로 적분하면 확률밀도함수로 결정되는 해당 구간의 면적이 된다. 이것은 부품의 치수를 나타내는 확률변수가 그 구간에 속할 확률과 같다.

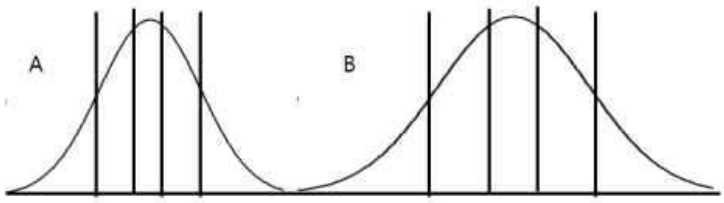

Figure 33. Equal area partitioning for components A and B

등면적분할방식은 등간격분할방식과 **(ⅰ) 범주 경곗값 설정**에서의 차이만 존재한다. 등간격분할방식이 값의 범위로 범주를 분할했다면 등면적분할방식은 정규분포를 따르는 두 부품에 대해 동일 면적을 가지게끔 범주를 분할한다. **(ⅱ) ball grade 예측**과 **(ⅲ) 예측 성능 평가**는 동일하게 시행된다.

예를 들어, 튜브의 범주를 2, 샤프트의 범주를 3으로 한다면 튜브는 figure 10과 같이 partition 1, partition 2로 나뉘고 샤프트는 figure 11과 같이 partition 1, partition 2, partition 3로 나뉜다. 튜브의 partition 1, partition 2의 중앙값을 X_1, X_2라 하고, 샤프트의 partition 1, partition 2, partition 3의 중앙값을 Y_1, Y_2, Y_3라 하자. X_1과 Y_1을 균일분포로 구한 회귀식에 넣어 ball grade를 구한다. 튜브의 partition 1과 샤프트의 partition 1에 해당하는 ball grade와 개별 치수에 의해 예측된 ball grade 간의 차이를 평가한다. X_1과 Y_2, Y_3에 대해 해당 절차를 반복하고, X_2와 Y_1, Y_2, Y_3에 대해 해당 절차를 반복하면서 2^{nd} accuracy를 구한다. 튜브와 샤프트 각각의 범주를 [2, 20]로 설정하여 총 361개의 2^{nd} accuracy를 구한 후 범주의 개수를 적게 하면서 90%의 정확도를 넘기는 계급 개수를 선택한다.

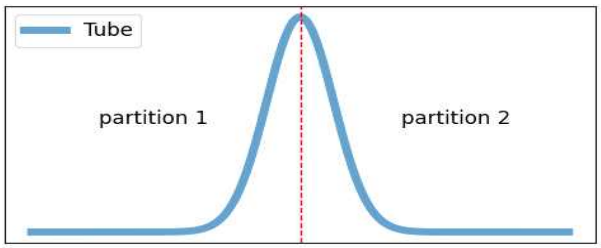

figure 34. equal area partitioning for Tube

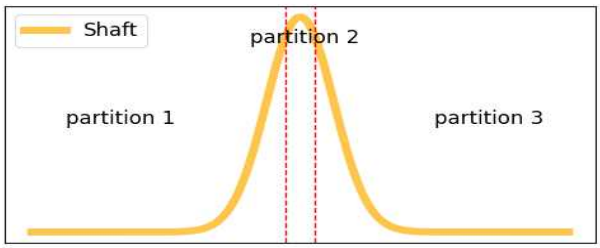

figure 35. equal area partitioning for Shaft

등면적분할방식 결과 튜브의 범주가 9개, 샤프트의 범주가 11개일 때 2^{nd} accuracy가 90% 이상(90.146%)이며 범주 개수를 최소로 가졌다.

각 범주의 대푯값을 평균과 중앙값으로 설정하였을 때의 각각의 평가 지표를 시각화하면 figure 12와 figure 13과 같다. 두 지표를 비교했을 때 큰 차이는 없었으나 중앙값을 사용했을 때 범주의 수를 더 줄이면서 2^{nd} accuracy는 90%를 넘기는 것을 확인할 수 있었다.

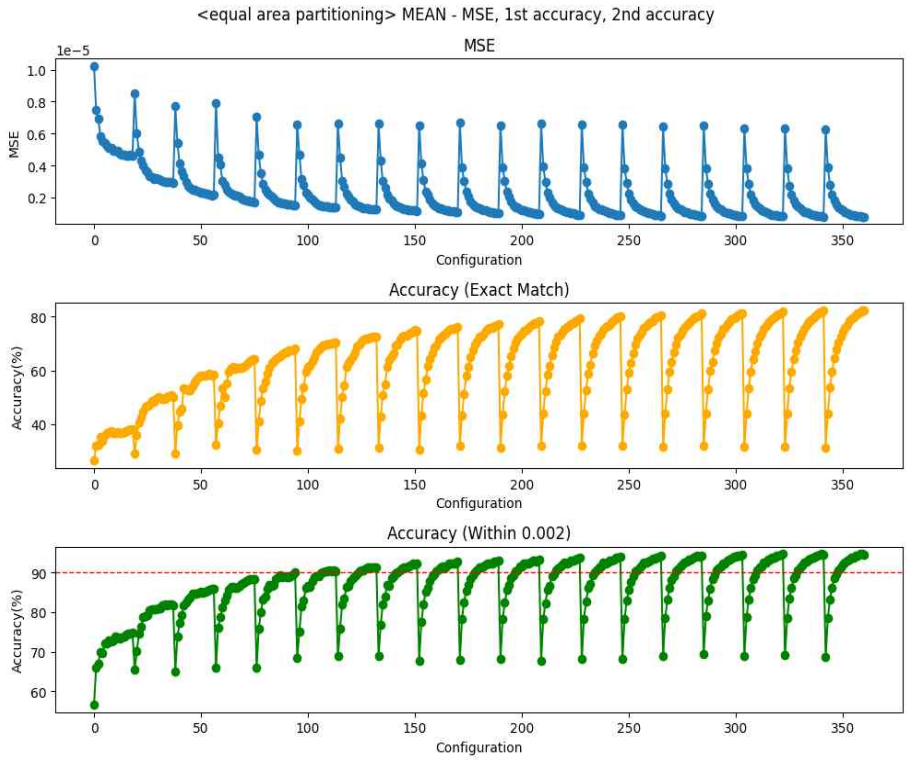

Figure 36. equal area partitioning with mean

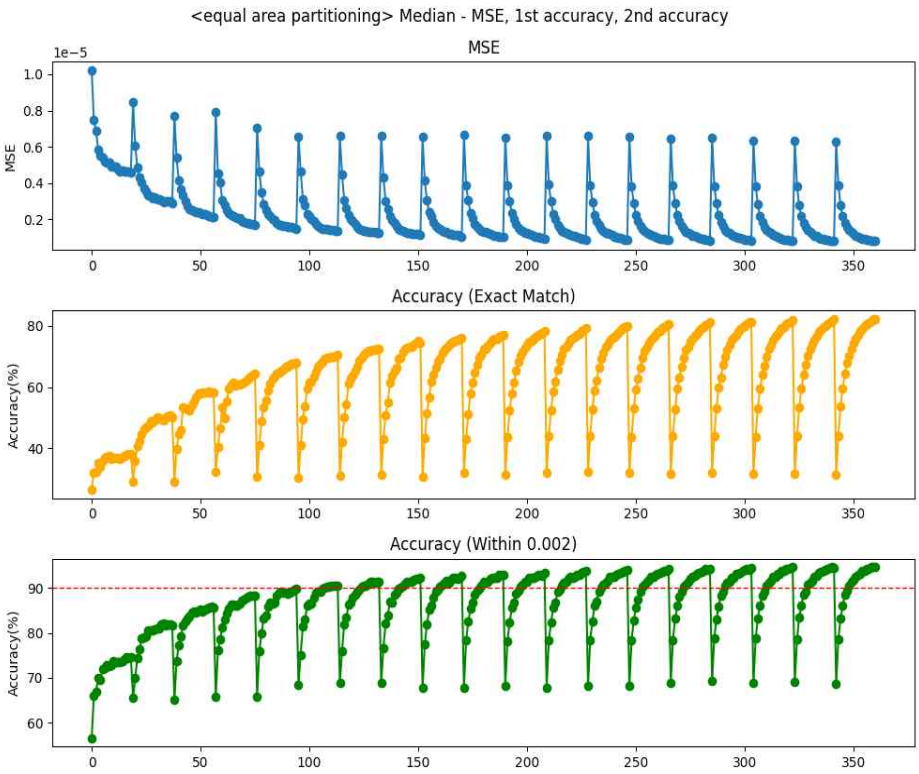

figure 37. equal width partitioning with medain

3. 데이터 분석 결과

시뮬레이션 결과 상·하한 지점을 $\mu \pm 3 \times \sigma^2$로 하고, 평균값을 대푯값으로 사용하는 등간격분할방식에 의해 튜브 8개, 샤프트 10개로 분류하는 것이 최적 분류 방법이라는 결론을 내렸다.

1-1) 정규분포 근사를 이용한 신뢰구간

각 예측은 정답이거나(1) 오답인(0) 이진 결정이기 때문에 분류 과정은 이항 분포를 따른다. 또한 표본 크기가 100,000으로 대표본이므로 정규분포 근사를 이용해 분류 정확도의 신뢰구간을 다음과 같이 계산할 수 있다.

$$accuracy \pm z\sqrt{\frac{accuracy(1-accuracy)}{n}}$$

따라서 정규분포 근사를 이용한 1st accuracy, 2nd accuracy의 95% 신뢰구간은 순서대로 (0.6125,0.6185), (0.8986, 0.9023)이다.

1-2) 몬테카를로 시뮬레이션을 이용한 신뢰구간

자료를 발생시키고 통계량을 계산하는 과정을 반복적으로 시행하는 몬테카를로 시뮬레이션을 이용해 경험적인 신뢰구간을 구할 수 있다. random seed를 바꾸어가며 10만 개의 표본을 반복적으로 발생시켜 2의 내용을 100번 반복 시행하였다.

그 결과 몬테카를로 시뮬레이션을 이용한 mse, 1st accuracy, 2nd accuracy의 95% 신뢰구간은 순서대로 (0.000001,0.000001), (0.6112, 0.6165), (0.8967, 0.9004)으로 도출되었다.

정규분포 근사와 몬테카를로 시뮬레이션을 이용한 신뢰구간 모두 0.9를 넘지는 않았지만 0.9 근처에서 형성되어 정확도가 0.9와 가깝다고 할 수 있다.

2) 짝맞춤 분할표

위 과정을 통해 제안하고자 하는 짝맞춤 분할표는 다음과 같다.

GRADE	SIZE	MARKING
A		A
B		B
C		C
D		D
E		E
F		F
G		G
H		H
I		I
J		J

Table 1. GRADE 기준표(TUBE)

GRADE	SIZE	MARKING
A		A
B		B
C		C
D		D
E		E
F		F
G		G
H		H

Table 2. GRADE 기준표(SHAFT)

Tgroup	Sgroup	2nd (-0.002)	1st	2nd (+0.002)
A	A			
	B			
	C			
	D			
	E			
	F			
	G			
	H			
B	A			
	B			
	C			
	D			
	E			
	F			
	G			
	H			
C	A			
	B			
	C			
	D			
	E			
	F			
	G			
	H			
D	A			
	B			
	C			
	D			
	E			
	F			
	G			
	H			
E	A			
	B			
	C			

	D			
	E			
	F			
	G			
	H			
F	A			
	B			
	C			
	D			
	E			
	F			
	G			
	H			
G	A			
	B			
	C			
	D			
	E			
	F			
	G			
	H			
H	A			
	B			
	C			
	D			
	E			
	F			
	G			
	H			
I	A			
	B			
	C			
	D			
	E			
	F			
	G			
	H			
J	A			
	B			
	C			
	D			
	E			
	F			
	G			
	H			

Table 3 개선된 짝맞춤 조립표

Ⅲ. 분석 의의 및 한계

1) 공정 개선 방향 및 기대효과

현재 공정 과정을 개선하기 위해 먼저 납품되는 tube와 shaft의 치수를 재고 등급별로 분류한다. 그리고 등급별로 분류된 부품함에서 tube와 shaft를 랜덤하게 선택한 뒤 짝맞춤표에서 제시하는 최적의 볼사이즈에 맞게 tube와 shaft, ball을 조립한다. 이렇게 하면 tube와 shaft의 범주의 수가 많아지더라도 작업자가 짝맞춤표를 한 번만 보고 조립을 할 수 있어서 시간 절약을 할 수 있다. 또한 제작한 짝맞춤표를 활용하여 신규 작업자도 적은 횟수로 최적의 조합을 찾아낼 수 있을 것이다. 그리고 짝맞춤표를 이용해 부품 조립 공정을 자동화한다면 인건비의 부담과 인적 오류에 대한 위험을 줄이고, 결과적으로 부품 조립 정확도를 높일 수 있을 것이다.

2) 분석 한계 및 향후 개선 방향

본 분석은 짝맞춤 조립표에 의해 얻어진 선형 관계가 정확하며 최적 ball size가 tube 치수, shaft 치수와 선형 관계를 갖는다는 가정 하에 진행되었다. 실제 양품 데이터를 이용해 최적 ball size를 모델링 하는 것이 바람직하나 제공받은 데이터에서 부품 치수들과 ball size의 연관관계를 확인하지 못해 차선책인 기존 짝맞춤표를 이용했다는 점에서 아쉬움이 남는다. 추후에 기계에 의한 제품 검사 결과(ex. 슬라이딩 부의 힘 크기) 등의 정보가 보충된다면 최적 ball size에 대한 정의를 보완할 수 있고 노이즈가 제거되어 질적으로 향상된 데이터를 이용해 더 정확한 예측과 시뮬레이션이 가능할 것이라 기대된다.

그리고 본 분석에서는 최적 분류 방법을 결정하기 위해 2^{nd} accuracy가 90%가 넘으면서 부품 범주의 수를 최소로 하는 것을 기준으로 설정하였다. 범주의 수가 클수록 부품 분류 비용과 관련된 행정 비용이 증가하기 때문에 범주의 수에 제한을 두는 것은 적절하지만 90%의 2^{nd} accuracy는 임의로 설정한 기준일 뿐이다. 만약 범주 수가 증가함에 따라 증가하게 되는 분류 비용과 정확도와 품질의 향상이 가져올 이익을 안다면 이를 함수 형태로 정의하여 총이익을 최대화시키도록 최적 분류 방법을 결정할 수 있을 것이다.

이러한 한계와 개선 가능성을 고려하면서, 향후 추가 데이터 및 분석 방법을 활용하여 더 정확하고 효과적인 공정 개선을 통해 [redacted]산업이 K-스마트 등대 공장으로써 디지털 전환 시대에 선도적인 역할을 할 수 있을 것으로 기대된다.

Ⅳ. 테스트 데이터 예측치

No	TubeNo	ShaftNo	BallSize(예측치)
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			
19			
20			
21			
22			
23			
24			
25			
26			
27			
28			
29			
30			
31			
32			
33			
34			
35			
36			
37			
38			
39			
40			
41			
42			
43			
44			
45			
46			
47			
48			
49			
50			

References

Pugh, G.A. 1986. “Partitioning for selective assembly.” Computers and Industrial Engineering Conference Proceedings, 175-179.

권혁무, 이영준, 이민구, 홍성훈. (2017). 선택조립방식의 효율성에 대한 시뮬레이션 검토. 품질경영학회지, 45(4), 829-846.