신뢰성응용연구
제16권, 제2호. pp.98-103, 2016

주성분 분석을 이용한 고객 공정의 불량률 예측 모형 개발

장윤희[1] · 손지욱[2] · 이동혁[2] · 오창석[2] · 이득중[2] · 장중순[1†]

[1]아주대학교 산업공학과, [2]LG display 품질센터

Development of Prediction Model using PCA for the Failure Rate at the Client's Manufacturing Process

Youn-Hee Jang[1] · Ji-Uk Son[2] · Dong-Hyuk Lee[2] · Chang-Suk Oh[2]
Duek-Jung Lee[2] · Joongsoon Jang[1†]

[1]Department of Industrial Engineering, Ajou University
[2]Quality Center, LG display

Purpose: The purpose of this paper is to get a meaningful information for improving manufacturing quality of the products before they are produced in client's manufacturing process.
Methods: A variety of data mining techniques have been being used for wide range of industries from process data in manufacturing factories for quality improvement. One application of those is to get meaningful information from process data in manufacturing factories for quality improvement. In this paper, the failure rate at client's manufacturing process is predicted by using the parameters of the characteristics of the product based on PCA (Principle Component Analysis) and regression analysis.
Results: Through a case study, we proposed the predicting methodology and regression model. The proposed model is verified through comparing the failure rates of actual data and the estimated value.
Conclusion: This study can provide the guidance for predicting the failure rate on the manufacturing process. And the manufacturers can prevent the defects by confirming the factor which affects the failure rate.

Keywords: Failure Rate Prediction, PCA(Principle Component Analysis), Regression Analysis

1. 서 론

생산 기술이 발전하고 제품 구조가 복잡해짐에 따라 제품의 품질을 향상시키거나 유지하기 위하여 하나의 특성치 보다는 여러 개의 연관된 특성치를 동시에 관리해야 하는 경우가 많다[1]. 기업들은 제조장비의 운영 데이터 등 제품의 제조현장에서 발생하는 다양한 공정 데이터를 수집하고, 다양한 분석기법을 이

† 교신저자 jsjang@ajou.ac.kr
2016년 4월 26일 접수; 2016년 5월 17일 수정본 접수; 2016년 5월 30일 게재 확정.

용하여 분석하고자 노력한다.

LCD TV는 제조사에서 패널을 제작하고 고객사에서 패널을 TV로 조립하여 소비자에게 판매하기 때문에 불량 발생에 두 기업이 유기적으로 연결되어 있다. 그러나 소기업의 제조현장에서는 수기데이터에 의존하여 공정을 관리하며, 많은 부분이 제대로 저장 및 분석되지 못하고 있다[2]. 또한 고객사라고 할지라도 타사의 데이터는 자체기술력으로써 기밀인 경우가 많아 공정 데이터 수집에 어려움이 있다. 따라서 제조사에서는 고객의 품질을 대표하는 지표로써 가장 쉽게 얻을 수 있는 고객 공정 불량률 데이터를 수집한다[3].

고객 공정 불량률 데이터는 각각 모델마다 기록되는데, 제조사가 판단할 때 기준 불량률을 넘는 이상 모델에 대해서는 설계변경 등의 사후조치를 진행한다. 그러나 이러한 이상 모델에 대해 사후조치를 반복할 경우 고객의 불만족 가중, 기업 이미지 하락에 대한 복구비용이 크게 발생한다[4]. 따라서 제품이 고객사에 전달되기 전, 사전 검사단계에서 모델의 불량률 예측을 진행하여 이상치를 나타내는 모델에 대해 검토를 진행할 필요가 있다.

공정관리 및 불량률 예측에 대한 연구는 공정능력지수에 관한 연구, 다변량 통합공정관리에 대한 연구 등 다수 진행되었다[1, 2]. 그러나 제조사가 가진 정보만을 이용하여 고객사의 공정 불량률을 예측한 사례는 극히 드물다. 기존에 유화윤, 김성범[3]은 고객 공정 불량률을 예측할 수 있는 인자를 선별하여 회귀분석을 진행하였다. 하지만 많은 변수를 모두 사용하여 다중회귀분석을 진행할 경우 설명변수들 사이의 높은 상관관계가 나타나며, 이는 다중공선성 문제를 일으킬 수 있다[5].

따라서 본 논문에서는 제조사 자체에서 가용한 변수를 고려하여 고객 공정에서 산출된 불량률에 미치는 영향을 확인할 수 있는 모형을 개발한다. 또한 기업에서 고객 공정 불량률 모니터링뿐만 아니라 사전분석을 통한 개선을 진행할 수 있도록 고객 공정 불량률 모형 개발 절차에 대해 설명한다. 또한 본 논문에서는 다중공선성 문제를 해결하기 위해 주성분 분석을 통해 주성분변수를 얻은 후, 그 변수를 회귀식에 대입하여 고객 공정불량률을 예측한다. 또한 예측 결과를 검증 모델에 대입하여 회귀모형의 적합성을 확인한다.

2. 고객 공정 불량률의 예측

고객 공정 불량률은 고객사에서 불량 판정된 제품 수(<Fig. 1>의 Defect2, 3, 4)를 제조사에서 출하된 제품 수로 나눈 것을 말한다.

본 논문에서는 고객 공정 불량률 예측을 <Fig. 2>와 같은 절차로 진행한다.

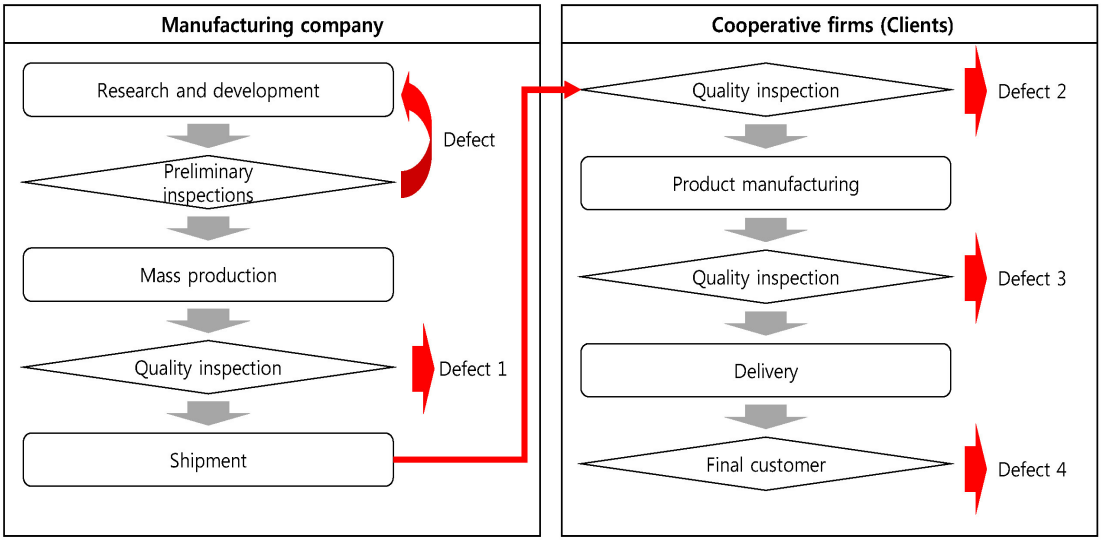

Fig. 1 Manufacturing process from manufacturer to client

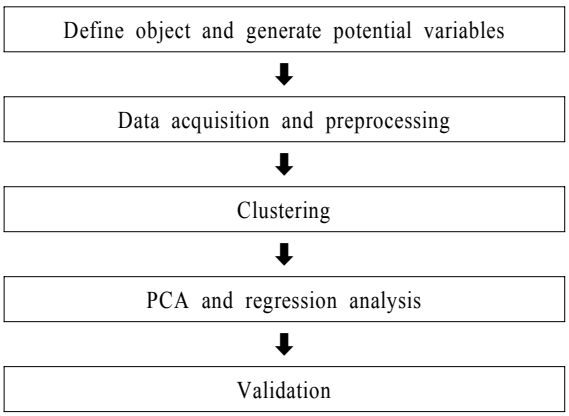

Fig. 2 Predicting procedure

2.1 대상 및 잠재변수 정의

제품을 이용하는 고객이나 제품 개발자의 Brain-storming을 통하여 대상의 고객 공정불량률에 영향을 미치는 잠재변수를 도출할 수 있다. 본 연구에서는 2014년 생산된 LCD TV 119개 모델을 대상으로 연구한다. LCD TV에 대한 잠재변수를 도출하기 위해 Brainstorming을 진행한 결과로 기구, 회로, 광학 등과 관련된 변수들로 총 70개 인자가 도출되었다.

2.2 변수별 데이터 수집 및 전처리

도출된 잠재변수에 대한 데이터를 수집하고 분석을 위해 데이터 전처리를 실시한다. 본 연구에서는 아래와 같은 전처리 방법을 진행하여 불량률에 미치는 영향이 적은 인자와 모델을 제거하였으며, 개발자가 예측 모형에 필요하다고 판단하는 변수는 제거하지 않고 추측치 삽입 등의 방법을 진행하였다.

2.2.1 이상치 제거

이상치는 데이터의 특성에 대한 해석을 바꿀 수 있기 때문에 데이터 전처리 과정에서 삭제해야 한다. 잠재변수와 Y(불량율) 간의 산점도 그래프(Scatter Plot)를 통해 대다수 결과와 다르게 경향이 나타나는 데이터를 삭제한다.

2.2.2 상관분석을 통한 변수 제거

상관계수가 높은 공정변수는 데이터 간의 설명을 중복해서 하기 때문에 데이터 특성에 대한 해석을 바꿀 수 있다. 상관관계가 높은 두 변수 중 하나를 제거한다.

2.2.3 결측치가 많거나 분산이 매우 작은 변수 제거

결측치가 많은 변수는 불량률과 관계와 변수들 간의 관계를 파악하기 어렵고 데이터 특성을 왜곡시킬 수 있다. 따라서 결측치가 많은 변수는 데이터 분석에서 삭제하는 것이 좋다. 또한 분산이 작은 변수는 불량에 대한 설명력이 작으므로 삭제하는 것이 전체 데이터 특성 파악을 쉽게 한다.

2.3 변수 클러스터링

본 연구에서는 데이터의 전처리 과정을 거친 후, 총 31개 인자, 46개 모델에 대한 데이터분석을 진행한다. 그러나 앞서 전처리 과정을 거쳤음에도 불구하고 데이터에 결측치가 존재하며 모델 데이터가 인자의 수에 비해 적어 분석에 어려움이 존재한다. 따라서 전문가 협의과정을 통해 기술적으로 관련이 깊은 변수를 분류하여 클러스터를 구성하였다.

- Cluster1(C1): 패널사이즈 및 화질 관련 10인자
- Cluster2(C2): TFT(Thin Film Transistor) 관련 11인자
- Cluster3(C3): Bezel 관련 8인자
- Cluster4(C4): 구동주파수 1인자
- Cluster5(C5): Lamp 전류 1인자

2.4 주성분 분석 및 회귀분석

주성분 분석이란 서로 상관성이 있는 다변량 데이터를 내포된 정보의 손실을 적게 하여 저차원의 데이터로 축약하는 다변량 비모수적 방법이다. 즉, 복잡한 데이터 세트를 단순화하기 위한 통계기법이다. 주성

분 분석을 진행한 결과, 각 Cluster의 고유값, 비율 및 누적비율은 <Table 1>과 같다. 산출된 주성분을 모두 이용할 경우, 차원의 축소라는 연구의 목적과 부합하지 않으므로, 주성분 변수를 선택하여 차원을 축소하면서 최대한의 변량을 포함시켜야한다.

주성분 변수를 선택하는 방법은 3가지가 있다. 첫 번째로는 카이저 기준이라고 하는 주성분 변수의 고유치가 1 이상인 변수를 선택하는 방법이다. 두 번째로, 누적기여율을 이용하는 누적고유치 비율기준이 있다. 일반적으로 누적기여율이 80% 수준에 도달할 때의 변수를 선택하는 방법이다. 세 번째는 전체 변량 중 가장 많은 부분을 설명하는 주성분인 제1 주성분만을 이용하는 방법이다[6]. 주성분을 선택하기 위하여 첫 번째 방법이었던 카이저 기준을 이용한다. Cluster 3은 PC2와 PC3의 고유값과 비율에 큰 차이가 없으므로 PC3까지 주성분 변수로 선택하였다.

선택된 주성분변수의 계수에 절대값을 취하여 값이 클수록 주성분 변수에 미치는 영향력이 큰 인자로 볼 수 있다. 각각의 주성분에 영향력이 크다고 볼 수 있는 인자를 아래 <Table 3>과 같이 나타내었다. 결과를 살펴보면 길이와 활성영역이C1의 PC1에 미치는 영향이 컸다. 따라서 C1의 PC1은 화면의 사이즈와 관련된 변수로 정의할 수 있다. 또한 길이와 활성영역은 음의계수를 나타내어 이 값들이 클수록C1_PC1은 작은 값을 나타낸다. 나머지 변수에 대해서는 기업의 정보보안관계로 상세 정보 언급은 생략하고 각각의 주성분에 대해 영향력이 컸던 변수명을 기입하였다.

주성분 분석의 결과로 도출된 계수를 이용하여 각 변수별 Score를 계산한 후 회귀분석을 이용하여 불량률을 예측한다. 회귀분석은 한 개 이상의 독립변수들의 선형 함수식으로 한 개의 종속변수를 표현하는 통계적인 분석방법이며, 그 결과로 산출된 회귀방정식은 아래 <Table 3>과 같다. 회귀방정식의 R-square가 0.826이므로 회귀모델이 잘 적합 되었다고 판정할 수 있다. 회귀분석의 결과는 주성분 분석의 결과와 마찬가지로 각 계수의 절대값이 클수록 불량률에 큰 영향을 미친다. 예를 들면, 앞서 주성분 분석에서 C1_PC1은 길이와 관련된 음의 계수를 가진 변수였으므로 회귀분석의 결과에서는 TV의 화면 사이즈가 클수록 고객 공정불량률이 큰 것으로 알 수 있었다. 그러나 회귀방정식 전체에서 C1_PC1의 계수는 다른 변수들의 계수에 비해 절대값이 크지 않아 고객 공정 불량률에

Table 1 The eigenvalue matrix of the principle components

	Cluster 1			Cluster 2			Cluster 3		
Variable	Eigenvalue	Proportion	Cumulative	Eigenvalue	Proportion	Cumulative	Eigenvalue	Proportion	Cumulative
PC1	5.0574	0.506	0.506	2.9279	0.266	0.266	6.2175	0.777	0.777
PC2	1.6584	0.166	0.672	2.4488	0.223	0.489	1.1511	0.116	0.893
PC3	1.1527	0.115	0.787	1.7717	0.161	0.65	**0.9392**	**0.095**	**0.988**
PC4	**1.0181**	**0.092**	**0.879**	**1.4063**	**0.128**	**0.778**	0.0748	0.009	0.998

Table 2 The factors of the variables

	Cluster1(C1)	Cluster2(C2)	Cluster3(C3)
PC1	Length(inch) Active Area(H) Active Area(V)	Gate Cu thickness(Å) Gate CD(㎛) SD Cu thickness(Å)	On Bezel(L) On Bezel(R) On Bezel(U)
PC2	Wx Wy	TFT Width TFT Length	On Bezel(D) Off Bezel(D)
PC3	Resolution(H) Resolution(V) Wx Wy	Cell Gap(㎛) SD MoTi thickness(Å)	On Bezel(U) Off Bezel(D)
PC4	2D luminance(nit) CR	SD MoTi thickness(Å) Pol transmissivity	-

Table 3 The results of regression analysis

Y = ppm = 411 - 256 C1_PC1 - 442 C1_PC2 - 348 C1_PC3 - 306 C1_PC4 + 470 C2_PC1 + 399 C2_PC2 + 49.8 C2_PC3 - 89.0 C2_PC4 + 124 C3_PC1 + 290 C3_PC2 + 548 C3_PC3 + 117 std_C4 + 345 std_C5

S = 177.169 R-sq = 92.9% R-sq(adj) = 82.6%

Table 4 Data for validation

Num. of Sample	Actual Data(ppm)	Estimated Value(ppm)	Error(Actual data - Estimated value)
1	1788.452	1581.227	207.225
2	1243.523	1423.163	-179.64
3	821.6387	766.7062	54.9325
4	290.7822	240.791	49.9912
5	332.9315	636.6473	-303.716

큰 영향을 미치는 인자는 아님을 알 수 있었다.

2.5 결과 검증

회귀모델이 실제로 고객 공정 불량률 예측에 사용될 수 있는지 확인하기 위해, 분석에 넣지 않았던 검증데이터와 예측 값이 유사한지 확인하였다. 취득 가능하였던 검증데이터는 총 5개였으며, <Table 4>는 검증데이터와 앞의 회귀모델을 이용해 예측한 예측값이다.

위의 검증데이터와 예측데이터의 차이에 대해 1-sample t-test를 진행한 결과는 <Table 5>와 같으며, P-value = 0.727로 유의하였다. 따라서 귀무가설($\mu = 0$)을 선택하며, 검증데이터와 예측데이터의 차는 없다고 볼 수 있다.

Table 5 Test of $\mu = 0$ vs $\mu \neq 0$

Variable	Error(Actual data - Estimated value)
N	5
Mean	-34.2
StDev	204.4
SE Mean	91.4
95% CI	(-288.0, 219.5)
T	-0.37
P	0.727

3. 결 론

본 연구에서는 고객 공정 불량률을 예측하는 절차를 수립하고, 주성분 분석과 회귀분석을 이용하여 예측 모델을 개발하였다. 구축된 예측모델의 유효성을 검증하기 위해 실제 발생한 불량률과 예측된 불량률의 오차에 대해 검정을 실시하였고, 그 결과에 유의차는 존재하지 않았다. 이를 통해 고객사의 공정 데이터를 이용하는 것이 아니라 제조사의 품질 특성치를 이용하여 고객사에서 조립한 제품에 대한 불량률을 예측하는 연구에 의미가 있음을 알 수 있다.

고객 공정 불량률을 예측할 수 있는 절차를 제시하여 한 가지 제품뿐만 아니라 여러 제품에 예측 모델을 구축할 수 있을 것으로 보인다. 또한 예측 모델을 제시함으로써 제품의 제조사에서 고객의 공정으로 전달되기 이전에 제품의 품질에 큰 영향을 미치는 인자에 대해 확인하고 조치를 취하여 저품질 비용을 낮출 수 있을 것으로 기대된다.

그러나 본 연구에서는 검증데이터를 충분히 확보하지 못했다는 한계가 존재하기 때문에 구축된 회귀모델을 검증하기 위한 다양한 평가 데이터가 확보되어야 할 것으로 보인다. 또한, 고객 공정 불량률과 관련한 새로운 잠재인자를 회귀모델에 반영하기 위해서는 새로운 회귀 모델을 구축해야 한다는 단점이 존재한다. 따라서 향후에는 새로운 잠재인자를 간단하게 포함할 수 있도록 추가적인 연구가 필요할 것으로 생각된다.

References

[1] Cho, G. Y. and Park, J. S. (2013). "Parameter estimation in a readjustment procedure in the multivariate integrated process control". Journal of the Korean Data and Information Science Society, Vol. 24, No. 6, pp. 1275-1283.

[2] Kim, J. S. and Cho, W. S. (2015). "Data analysis of 4M data in small and medium enterprises". Journal of the Korean Data and Information Science Society, Vol. 26, No. 5, pp. 1117-1128.

[3] Yu, H. Y. and Kim, S. B. (2015). "Prediction of the Client's Defect Rate at the LCD Display Industry using the Data Mining Technique". Proceedings of The Korean Institute of Industrial Engineers, pp. 1351-1359.

[4] Andersen, B. and Moen, R. M. (1999). "Integrating benchmarking and poor quality cost measurement for assisting the quality management work". Benchmarking: An International Journal, Vol. 6, No. 4, pp. 291-301.

[5] Kwun, S. H. (2008). "Multivariate Analysis and Application". Freeacademy.

[6] Lee, J. Y. (2009). "Measuring the Accuracy of Proxy Variables with PCA". Sungkyunkwan University.