Journal of the Korean Institute of Industrial Engineers
Vol. 42, No. 5, pp. 327-336, October 2016.
ISSN 1225-0988 | EISSN 2234-6457
http://dx.doi.org/10.7232/JKIIE.2016.42.5.327

<Application Research>

LCD 디스플레이 산업에서 데이터마이닝 알고리즘을 이용한 고객 불량률 예측

유화윤 · 김성범†

고려대학교 산업경영공학과

Prediction of Customer Failure Rate Using Data Mining in the LCD Industry

Hwa Youn You · Seoung Bum Kim

School of Industrial Management Engineering, Korea University

Prediction of customer failure rates plays an important role for establishing appropriate management policies and improving the profitability for industries. For these reasons, many LCD (Liquid crystal display) manufacturing industries have attempted to construct prediction models for customer failure rates. However, most traditional models are based on the parametric approaches requiring the assumption that the data follow a certain probability distribution. To address the limitation posed by the distributional assumption underpinning traditional models, we propose using parameter-free data mining models for predicting customer failure rates. In addition, we use various information associated with product attributes and field return for more comprehensive analysis. The effectiveness and applicability of the proposed method were demonstrated with a real dataset from one of the leading LCD companies in South Korea.

Keywords: Failure rates, Field Return Data, Life Testing, Data Mining, LCD

1. 서 론

디스플레이 산업이란, TV, 컴퓨터, 모바일 등의 화면구성에 관련된 산업을 말한다. 디스플레이 산업은 우리나라 주력 산업 가운데 하나로서 지난 10년간 한국 기업들이 세계 시장점유율 1위를 차지해왔다(Seo, 2014). 최근 중국의 기술 추격과 일본의 신 기술 개발을 통한 반격 속에서도 한국 디스플레이 기업들은 세계 최대 디스플레이 생산국의 지위를 지키고 있으며, 지속적인 투자로 일자리 창출에도 기여하고 있다(Park, 2015). 디스플레이의 주요 품목으로는 CRT(Cathode ray tube), LCD(Liquid crystal display), PDP(Plasma display panel), OLED(Organic light emitting diode) 등을 들 수 있고, 그 중 LCD는 디스플레이의 대표 품목으로서 디스플레이 산업의 성장동력으로 자리매김하였다(Moon, 2012). OLED의 경우, 2012~2013년에 각각 전년대비 70% 이상의 고속성장을 거듭하며 새로운 성장동력으로 가능성이 대두되고 있다(Seo, 2014).

LCD 디스플레이 산업은 액정 화면을 통해서 전기적인 신호를 영상으로 표시해주는 전자 장치인 LCD를 제조 및 판매하는 산업이다. LCD 디스플레이 산업의 주요 특징은 수 백 개의 기업들이 하나의 산업에 속해 있어서 전후방 산업 연관 효과가 큰 산업이라는 점이다. LCD 디스플레이 산업에서는 유리기판, 드라이버 집적회로, 컬러필터, 광원 등 원자재와 제조장비를 전방산업체들로부터 제공받아서 LCD를 생산한다. 동시에 TV, 모니터, 노트북, 모바일 등 응용기기를 생산하는 후방

† 연락저자 : 김성범 교수, 02841 서울특별시 성북구 안암로145 고려대학교 산업경영공학과, Tel : 02-3290-3397, Fax : 02-929-5888, E-mail : sbkim1@korea.ac.kr
2016년 1월 28일 접수; 2016년 5월 11일 수정본 접수; 2016년 5월 31일 게재 확정.

산업체에 LCD를 부품으로 제공하는 부품 산업이기도 하다(Bae, 2004). 즉, LCD 디스플레이는 전방산업체로부터 원자재를 제공받아 조립한 LCD 패널을 의미하고, LCD 디스플레이의 고객은 응용기기를 생산하는 후방산업체와 제품을 최종 구매하는 소비자를 모두 포함한다.

세계 LCD 디스플레이 시장의 성장률은 2012년 12.9%, 2013년 0.7%, 2014년 10%, 2015년 6.3%로 전체 성장세가 둔화되었고, 국내 LCD 기업의 세계 시장 점유율은 현재까지 세계 1위를 지키고 있지만 2년 연속 하락하면서 매출증가율이 답보상태에 있는 실정이다(Seo, 2014). 이에 따라 국내 LCD 디스플레이 기업들은 다방면으로 수익성을 증가시키기 위한 방안을 모색하고 있으며, 고객 불량률 예측에 관한 연구는 이러한 수익성 개선 방안 가운데 하나가 될 수 있다.

고객 불량률이란 제품을 고객에 판매한 이후 서비스 센터에 접수된 불량 건수를 판매 수량으로 나눈 비율을 의미하며, 고객인지품질(Perceived quality) 수준을 평가 할 수 있는 지표 가운데 하나이다. 고객인지품질은 고객이 제품을 구매한 후 인지한 품질 수준으로서 고객 만족도지수를 평가하는 모델에서 그 의미를 알 수 있다. <Figure 1>은 국가고객 만족지수(National customer satisfaction index, 이하 NCSI) 모델로서, 고객인지품질 수준은 고객기대 수준과 함께 고객 만족도에 영향을 미치는 주요 요인으로 고객 만족도를 파악하기 위해서 반드시 평가되어야 하는 항목이다(Yoon *et al*., 2003). 또한, 고객 만족도는 고객충성도로 연결되는데, 이는 고객의 재 구매 가능성을 나타내므로 궁극적으로 고객 만족도가 기업의 수익성에 긍정적인 영향을 미치는 것으로 알려져 있다(Lee, 2006). 따라서, 기업에서는 수익성과 관련 있는 고객 만족도를 평가하는 방법 가운데 하나로서 고객 불량률을 예측하여 경영계획을 수립 할 때 중요한 자료로 활용하고 있다.

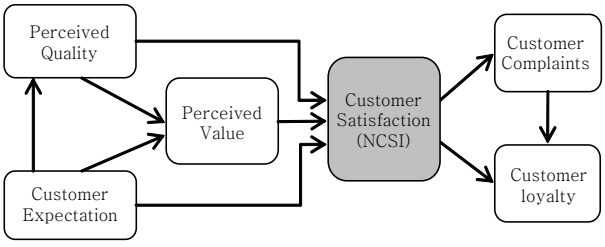

Figure 1. Overview of the NCSI model

그 뿐만 아니라, 고객 불량률 예측은 서비스 재고를 관리하기 위해서도 필수적이며, 손실비용과 밀접한 관련이 있기 때문에 기업의 이익을 증가시키기 위해서 반드시 필요하다고 하겠다. 우선, 제품을 판매한 기업에서는 고객이 서비스 센터에 접수한 불량품에 대해 교체 또는 수리 등의 서비스를 제공해야 하고, 이를 위해서는 사전에 서비스 재고를 비축해야 한다. 이때, 비축 재고 수량을 결정하기 위해서는 고객으로부터 반입될 불량 수를 사전에 파악해야 하므로 서비스 재고관리를 위해서 고객불량률 예측은 필수적인 활동이다. 또한, 고객 불량률 예측 성능에 따라서 서비스 재고에 대한 손실비용이 증감하므로, 손실비용을 줄이는 방법으로서 고객 불량률의 예측 성능을 높이려는 연구가 중요하다고 할 수 있다. 간단히 설명하면, 기업에서는 서비스 재고의 잔여수량에 대해서 일반적인 판가보다 낮은 가격으로 판매하고, 부족수량에 대해서는 고객에게 교체나 수리 대신 판가를 지불하는 과정에서 손실비용이 발생한다. 즉, 실제 발생할 고객 불량률에 근접하게 예측할수록 손실비용을 줄일 수 있으므로 기업의 이익을 증가시키기 위해서 고객 불량률을 정확히 예측하는 것이 매우 중요하다.

고객 불량률 예측 연구는 1950년대부터 1990년대에 신뢰성 공학 분야에서 활발히 연구된 고장률 예측 방법론과 2000년대 이후 신뢰성 공학 외 다른 분야에서 연구된 방법론으로 나눌 수 있다. 먼저 신뢰성 공학 분야의 예측 방법론은 크게 3가지로 설명 할 수 있다. 그 가운데 2가지는 수명분포함수를 이용하는 방법으로서 가속수명시험에 의한 예측과 시장반입데이터를 이용한 예측이고, 이외에 신뢰성 부품 규격을 이용한 방법이 있다(Shin *et al*., 2011).

첫 번째로 가속수명시험에 의한 수명분포함수 추정 방법은 제품을 가혹한 조건에서 시험한 다음, 가속 모형을 이용해서 가속수명시험 도중 고장이 발생한 시간 정보를 일반 사용 조건에서의 고장 시간으로 환산하여 제품의 수명을 예측하는 방법이다. 이때 가속 모형은 제품의 수명과 스트레스의 관계로서 수명분포의 모수를 스트레스 변수들의 함수로 표현하므로, 가속수명시험 계획의 수립과 분석이 매우 중요하다. 따라서 가속수명시험을 실시하기 전에 가속 인자와 가속 계수를 결정하기 위한 예비 시험을 실시해야 한다(Lee and Chung, 2009). 이 방법은 전자제품의 대표적인 신뢰성 평가 방법으로 제품을 판매하기 이전에 신뢰성을 검증하고 보증 할 수 있는 장점이 있다(Yum, 2014). 그러나 몇 가지 제약이 따르므로 본 논문에서는 제약을 중점적으로 서술하였다. 먼저, 여러 부품을 조립한 시스템의 고장률을 예측하기 위해서는 부품단위의 고장률을 계산한 다음 각 부품들의 특성을 찾아서 조합된 제품의 신뢰도를 추정하는데, 부품마다 고장 형태를 가속시키는 스트레스의 종류가 다르기 때문에 각 부품의 가속인자를 알맞게 선정해야 한다(Shin *et al*., 2011). 또한, 고가 제품의 경우 많은 샘플을 시험하는 것은 비용 및 시간 측면에서 부담이 크기 때문에 대개 한정된 시간 동안 소량의 샘플을 평가한다(Lee, 2005). 이 경우 실험 시간 동안 발생하는 불량 수가 극히 적으므로 가속수명시험 결과로 시장 불량률을 예측하는 데 한계가 있다.

두 번째로 시장반입데이터에 의한 수명분포함수 추정 방법은 일정기간 동안 서비스센터에 반입된 고객의 불량 수와 해당 기간 동안 불량으로 기록되지 않은 제품 수 정보를 바탕으로 수명분포함수를 추정하여 제품의 수명을 평가한다. 대표적인 수명분포함수는 와이블 분포, 지수분포, 감마 분포, 대수정규분포로 다양하지만 일반적으로 모수에 따라서 증가 불량률, 감소불량률, 상수 불량률을 나타낼 수 있는 와이블을 사용한다(Wang and Chu, 2011). 이 방법은 시장반입데이터를 이용하여 시간에 따라 발생하는 고장확률분포를 추정하고 이를 이용해 미래 고장확률을 예측하는 방법이므로, 분포 추정을 위한 시장반입데이터의 반입기간에 따라서 예측 시점이 달라질 수 있다. 무엇보다, 시장반입데이터에 포함된 시간은 제품의 사용시간뿐 아니라 기업에서 고객으로 운송되는데 걸린 시간, 고객이 사용하는 동안 전원을 꺼둔 시간, 불량이 발생해서 서비스 센터로 반입되는데 걸린 시간을 포함하기 때문에 LCD 시장반입데이터가 수명분포함수를 따르기 어렵다는 한계점이 있다(Wilson *et al*., 2009).

마지막으로 신뢰성 부품 규격을 이용한 예측은 MIL-HDBK-217F와 Telcordia SR-332 등 라이브러리를 이용하는 방법이다. 이는 가속수명시험보다 신속히 신뢰도를 평가 할 수 있는 장점이 있지만 부품 규격에 맞춰 전체 시스템의 신뢰도를 예측하는 방식으로서 현실과 예측 값의 차이가 크고, 규격에 제약이 많은 단점이 있다(Lee, 2005).

2000년대부터는 신뢰성 공학 분야에서 기존 예측 방법론을 활용한 연구나 진단 기술을 개발하는 연구가 주로 진행되었고(Yum *et al*., 2014), 고장률 예측에 관한 연구는 신뢰성 공학 외 기타 분야에서 연구사례를 찾아 볼 수 있다. 최근 몇 가지 연구에서 데이터마이닝 기법을 이용한 예측 방법론이 소개되었는데 대표적인 연구들은 다음과 같다. Murray *et al*.(2005)는 서포트벡터머신(Support vector machine) 또는 순위합과 나이브베이즈(Naïve bayes)를 결합한 알고리즘을 이용해서 하드드라이브의 고장률을 예측하였고, Nagappan *et al*.(2006)은 다중회귀분석을 이용해서 소프트웨어의 고장률을 예측하였으며, Hwang *et al*.(2008)은 의사결정나무를 이용해서 주상변압기의 고장여부를 예측하였다.

그러나 이러한 데이터마이닝 기반의 연구들은 LCD 외 다른 산업 분야에서 연구된 것이며 LCD의 고객 불량률 예측 방법론은 현재까지 가속수명시험 또는 시장반입데이터를 이용해서 수명분포함수를 추정하는 방법 외에 관련 문헌을 찾아보기 힘들다. 기타 방법으로는 국내 LCD 디스플레이 기업에서 자체 도메인 지식을 기반으로 고안한 예측 방법이 있지만 이 역시 개선의 여지가 많다고 하겠다. 대표적으로 평균 불량 증가율을 이용한 방법이 있는데, 이는 제품 크기에 따라 그룹을 나누고 과거 축적 된 시장반입데이터를 이용해서 그룹마다 시간에 따른 고객 불량률의 평균 증가율을 구한 다음 일정기간 동안 발생한 고객 불량률에 그룹의 평균 증가율을 곱하여 새로운 제품에 대한 고객 불량률을 예측하는 방법이다. 예를 들어, 어떤 그룹의 과거데이터를 분석한 결과, 보증기간 동안 반입된 고객 불량률이 초기 6개월 동안 반입된 고객 불량률 대비 평균 3배 증가했다면, 해당 군집에 속하는 새로운 제품의 보증기간 동안 발생할 고객 불량률은 초기 6개월 동안 반입된 고객 불량률에 3배를 하는 방식이다. 이 방법은 분포를 가정하지 않고 방대한 양의 데이터를 사용 할 수 있는 장점이 있지만, 단순히 크기 정보만을 사용하여 그룹을 나누며 시간에 따라 증가하는 고객 불량률의 평균 기울기를 모든 제품에 적용하는 방법이므로 복잡한 속성을 지닌 제품의 고객 불량률을 예측하기에는 한계가 있다. 이처럼 LCD 디스플레이 산업의 고객 불량률 예측 방법은 개선 할 여지가 많음에도 불구하고 관련 연구가 부족한 실정이다.

따라서 본 연구에서는 신뢰성 공학 기반 예측 방법의 한계점을 보안하기 위하여 데이터의 분포를 가정하지 않는 데이터마이닝 알고리즘을 이용하고, 독립변수로서 신뢰성 공학 분야에서 사용되어 온 시장반입데이터에 제품 속성 정보를 추가로 사용하여 LCD 산업의 고객 불량률 예측력을 높이고자 하였다. 본 논문의 저자들이 조사한 바에 의하면 LCD 산업에서는 고객 불량률 예측에 제품 속성 정보를 사용하거나 데이터마이닝 알고리즘을 기반으로 한 연구가 없었으므로 본 연구가 관련 연구에 대한 촉매제가 되길 기대해 본다. 본 연구에서 소개한 데이터마이닝 기반 방법론의 우수성은 기존 예측 방법론의 대표 방법으로서 국내 LCD 기업에서 사용하고 있는 시장반입데이터를 이용한 수명분포함수 방법 및 평균 불량 증가율 방법과 예측 성능을 비교하여 입증하였다. 예측력 향상은 고객 불량률 예측 결과에 대한 신뢰도를 높이므로 경영계획 수립 시 더 중요한 역할을 할 수 있으며 서비스 관련 손실비용을 절감시킴으로써 기업 이익을 증가시키는데 기여할 수 있는 점에서 의의가 있다.

본 논문의 구성은 다음과 같다. 이어지는 제 2장에서는 데이터마이닝 기반의 고객 불량률 예측 방법론에 대해 서술하였다. 제 3장에서는 실제 LCD 디스플레이 데이터를 이용하여 데이터마이닝 기반 예측 방법과 기존 예측 방법론의 예측 성능을 비교하였고, 또한 예측력에 따른 손실비용을 계산하여 데이터마이닝 기반 예측 방법의 효과를 입증하였다. 마지막으로 제 4장에서는 본 연구의 결론 및 기대효과와 함께 한계점에 대해 논의하고 향후 연구 방향을 모색하였다.

2. 데이터마이닝 기반 고객 불량률 예측

기존 LCD 디스플레이 기업의 고객 불량률 예측 방법은 가속 수명시험 및 시장반입데이터를 이용하여 수명분포를 추정하는 방법과 평균 불량 증가율을 이용한 방법으로 나눌 수 있고, 각 방법론은 여러 가지 한계점이 있음을 앞서 설명하였다. 본 연구에서는 예측력을 높이기 위해 데이터의 특정 확률분포를 가정하지 않으며 다양한 데이터 형태를 처리할 수 있는 데이터마이닝 기법을 활용하였다.

2.1 데이터 구성

먼저, 종속변수는 연속형 변수로서 보증기간 동안 발생하는 총 고객 불량률이다. 이는 보증기간 동안 기록한 시장반입데이터를 이용해서 계산하며 방법은 식 (1)과 같다.

$$\text{Dependent variable} = \frac{\sum_{i=1}^{n} F_i}{\text{monthly sales}} \times 10^{-6} \quad (1)$$

F_i는 판매 이후 i-1개월에서 i개월 동안 반입된 불량 수이고 n은 보증기간을 나타낸다. LCD 서비스 센터에서 기록한 시장반입데이터는 경과 월 단위로 발생한 불량 수를 의미하므로, 종속변수는 보증기간 동안 경과 월 별 발생한 불량 수의 합을 제품의 월 판매량으로 나눈 비율로 산출된다. 종속 변수의 단위는 일반적으로 ppm(part per million)을 사용한다.

독립 변수는 총 18개로 시장반입데이터 6개와 제품 속성을 대표 할 수 있는 변수 12개로 구성하였다. 독립변수로 사용하는 시장반입데이터는 판매 직후 초기 일정기간 동안 발생한 고객불량률 정보로서, 본 논문에서는 6개월 동안 발생한 고객 불량률 정보를 이용하였다. 6개월은 일반적으로 LCD 기업에서 시장반입데이터를 토대로 수명분포함수 방법이나 평균 불량 증가율 방법을 이용할 때 사용하는 최소한의 정보이다. 이는 운송 및 유통에 걸리는 시간을 고려했을 때 안정적인 데이터를 확보하기까지 통상적으로 최소 6개월이 걸리기 때문이다. 본 연구에서는 제품을 판매한 이후 6개월 동안 발생한 고객불량률을 경과 월에 따라 6개의 독립변수로 구분 지어 사용하였다. 이를 구체적으로 나타내면, 판매 이후로 1개월동안 발생한 불량률, 판매 이후로 1개월에서 2개월 사이에 발생한 불량률, 2개월에서 3개월 사이에 발생한 불량률, 3개월에서 4개월 사이에 발생한 불량률, 4개월에서 5개월 사이에 발생한 불량률, 5개월에서 6개월 사이에 발생한 불량률이다.

제품 속성을 나타내는 12개의 변수는 4개의 연속형 변수와 8개의 범주형 변수(이중 3개가 이진변수)로 나뉜다. 연속형 변수는 제품 화면 크기, 주요 배선 두께, 해상도, 판매수량이고, 범주형 변수는 생산 월, FAB 공정, 모듈 공정, 후방산업체, 제품 설계 구조이며, 이진 변수는 전극 재료, 광원, 초도 양산 여부(양산 승인 이후 3개월 이내 생산 제품 여부)이다.

2.2 데이터 전처리

데이터마이닝 알고리즘을 적용하기에 앞서 전처리 과정으로서 데이터를 변환 및 정제하여 사용하였다. 먼저 데이터 변환 방법으로 연속형 변수의 경우 Z-score로 표준화하고, 범주형 변수는 Dummy 변수로 변환하였으며, 이진 변수는 0 또는 1로 변경하였다. 또한, 데이터 정제 방법으로는 6개월간 반입된 고객 불량률 변수 6개 가운데 5개 이상의 변수가 0으로 결측치가 많거나, 또는 종속변수가 극단 이상치에 해당하는 관측치의 경우를 제거함으로써 데이터의 품질을 높이고 추후 구축하는 예측 모델의 신뢰성을 높이고자 하였다(Bae *et al*., 2007). 앞에서 서술한 데이터의 구성과 전처리 방법을 요약하면 <Table 1>과 같다.

2.3 데이터마이닝 예측모델

전처리 과정을 거친 후 데이터마이닝 알고리즘을 이용하여 예측 모델을 구축하였다. 본 연구에서는 다양한 데이터마이닝 기반 방법론 가운데 예측 문제에 널리 사용되어 효과가 검증된 라쏘(Least absolute shrinkage and selection operator : LASSO), 인공신경망, 의사결정나무, 랜덤포레스트, 서포트벡터머신을 사용하였다. 사용된 예측 기법들은 다음과 같이 간단히 설명하며, 보다 자세한 설명은 참고 문헌으로 대신한다.

먼저, 라쏘는 다중선형회귀와 같은 선형 모델링 방법이지만, 회귀계수 추정 시 다중선형회귀의 목적식에 제약식을 도입하여 예측 모형을 구축하는 기법이다. 라쏘의 회귀계수 추정식은 식 (2)와 같다.

$$\hat{\beta} = argmin(RSS + \lambda \sum_{i=1}^{p} |\beta_i|) \quad (2)$$

β는 회귀 계수이고, RSS(Residual sum of square)는 잔차 제곱합이며, λ는 조절 모수이다. 즉, 라쏘는 잔차 제곱 합과 제약식을 모두 최소화하는 회귀계수를 구하며, 이때 λ를 0부터 1까지 조절하여 모델의 성능을 개선 할 수 있다. 이 기법은 회귀계수의 크기를 줄여줄 뿐만 아니라 주요 변수를 선택함으로써 모델의 해석력과 예측 정확도를 높여주는 장점이 있다. 특히, 고차원 회귀 문제를 해결하는 데 적합한 기법으로 알려져 있다(Tibshirani, 1996).

인공신경망(Rosenblatt, 1958)은 패턴인식이나 데이터마이닝 등 다양한 분야에 응용되고 있는 기법으로 회귀 문제뿐만 아

Table 1. Structure and conversion of dependent and independent variables

No.	Division	Name	Type		Preprocess
1	Dependent variable	The sum of field failure rates occurred during the warranty period	Numerical		z-score Normalization
2	Independent variables	Field failure rate during the first month after sales			
3		Field failure rate during the second month after sales			
4		Field failure rate during the third month after sales			
5		Field failure rate during the fourth month after sales			
6	Independent variables	Field failure rate during the fifth month after sales	Numerical		z-score Normalization
7		Field failure rate during the sixth month after sales			
8		Screen size			
9		Main pad thickness			
10		Resolution			
11		Sales			
12		Month of production	Categorical	12ea	Dummy variables
13		FAB Line		3ea	
14		Module Line		3ea	
15		Customer brand		8ea	
16		Design structure		3ea	
17		Electrode material	Binary		0 or 1
18		Light source			
19		Whether the initial production or not			

니라 분류 문제 해결을 위한 예측 알고리즘이다. <Figure 2>에서 보여주듯이, 입력층, 은닉층, 출력층으로 구성되며, 뉴런으로 불리는 단위처리개체로 이루어져 있다.

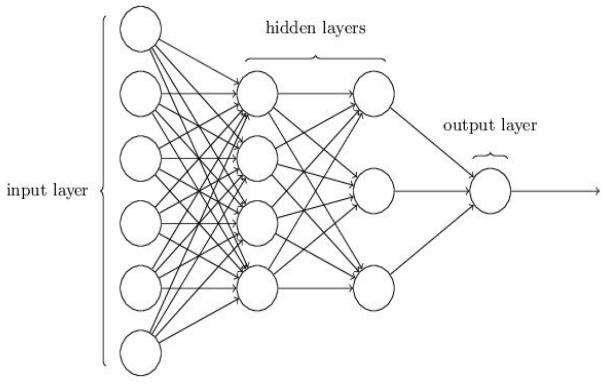

Figure 2. The structure of artificial neural network

각 뉴런은 입력 신호에 대한 연산을 수행한 후 연결고리를 통하여 다음 뉴런에 정보를 전달하는 처리개체이다. 연결고리란 하나의 뉴런으로부터 다음 뉴런으로의 정보 흐름을 의미한다. 가중치는 연결고리를 통하여 할당되며, 결과적으로 가중치가 고려된 입력신호 값이 다음 뉴런에 전달된다(Park *et al*., 2005).

인공신경망의 특징은 데이터가 복잡하고 비선형적인 경우에도 패턴을 추출할 수 있는 점이다(Kim, 2004). 그러나, 모델이 제시하는 결과에 대해서 원인을 명쾌하게 설명하지 못하고 과도하게 학습을 진행 할 경우 전역 최적해가 아닌 국부 최적해(Local minima)가 선택 될 가능성이 있다(Han, 2009).

의사결정나무는 의사결정 규칙을 도표화하여 관심대상이나 집단을 몇 개의 소집단으로 분류하거나 예측하는 분석방법이다. <Figure 3>과 같이 각 속성의 분할 기준을 포함하는 마디, 마디와 마디를 연결해주는 가지, 그리고 최종 결정된 속성을 의미하는 잎 마디로 구성된다.

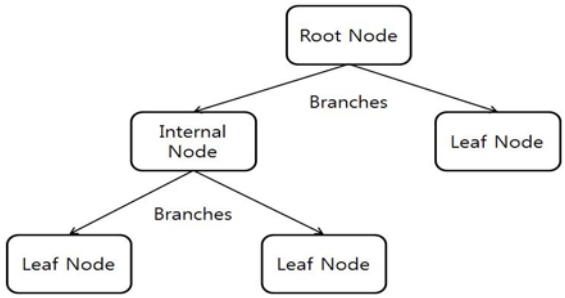

Figure 3. The structure of decision tree

의사결정나무는 분석 과정을 나무 구조에 의해서 표현하므로 분석 과정과 결과를 쉽게 이해할 수 있는 특징이 있다(Curram

and Mingers, 1994). 또한, 독립 변수 형태가 범주형과 연속형이 혼합된 경우에도 데이터의 변환 없이 적용할 수 있는 장점이 있다(Hastie *et al.*, 2001). 본 연구에서 사용한 의사결정나무 알고리즘은 CART(Classification and regression tree)로 의사결정나무의 알고리즘 가운데 가장 잘 알려진 방법론 가운데 하나이며(Linoff and Berry, 2011), 지니 지수(Gini Index) 또는 분산의 감소량을 사용하여 나무 가지를 이진 분리하는 방법이다(Breiman *et al.*, 1984).

랜덤포레스트(Breiman, 2001)는 의사결정나무의 메타학습형태로서, 다수의 의사결정나무를 형성하고 각각의 예측 값을 조합하는 대표적인 앙상블 방법 가운데 하나이다. 랜덤포레스트는 훈련 과정에서 무작위로 추출된 훈련 데이터로 많은 수의 의사결정나무를 형성하여 다양한 패턴을 포괄하므로, 새로운 데이터에 대해서도 높은 수준의 정확도를 보이는 것으로 알려져 있다. 또한, 연속형 및 범주형 변수가 혼합된 데이터에도 별도의 변환 없이 사용 할 수 있고 변수간 스케일이 다른 경우에도 사용할 수 있는 장점이 있다(Deng and Runger, 2013). 무엇보다도 모델 구축 시 사용자가 결정해야 하는 독립변수의 수가 적고 그 값에 덜 민감하며 변수의 중요도를 제공하므로 사용자 입장에서 모델 구축이 용이하다(Liaw and Wiener, 2002).

서포트벡터머신(Vapnik, 1995)은 문서분류, 영상인식, 문자인식 등 여러 분야에서 우수한 성능을 보여주는 대표적인 데이터마이닝 기법 중 하나이다. 이 기법은 분류와 예측에 모두 응용할 수 있는 알고리즘으로서 크게 분류를 위한 SVC(Support vector classification)과 예측을 위한 SVR(Support vector regression)으로 구분하며(Sholkopf *et al.*, 2002), 본 논문에서는 예측 문제를 해결하기 위해서 SVR을 사용하였다. 기본적으로 SVC는 다차원 상의 점들로 표현되는 데이터를 두 개의 그룹으로 분할 하는데 마진을 최대화하는 초평면(Hyperplane)을 구하고, 이 초평면을 결정함수로 사용하여 미지의 데이터가 속할 그룹을 예측한다(Park, 2006). SVR은 SVC를 회귀분석에 사용할 수 있도록 확장한 개념으로서, 실제값과 예측값의 오차가 최고 ϵ 이내이면서 마진을 최대화 하는 함수를 구하여 미지 데이터의 종속변수를 예측한다. 이와 같은 일반화 능력을 최대화 하기 위해서 <Figure 4>와 같이 ϵ-무감각 손실함수(ϵ-intensive loss function)를 사용한다. 이는 실제값으로부터 일정거리 이내의 오차를 무시하기 때문에, ϵ-튜브 안에 있는 데이터는 무시하고 밖에 있는 데이터의 오차만 슬랙변수(ξ)로 측정한다.

서포트벡터머신의 주요 특징은 데이터가 입력공간에서 선형으로 설명되지 않는 경우, 데이터를 고차원 특징 공간의 점으로 변환시킨 다음 학습을 수행하는 것인데, 이때 특징공간에서 나타나는 함수 $\phi_{ij} = \phi(x_i)^T\phi(x_j)$는 커널함수 K($x_i$, x_j)를 사용하여 계산 할 수 있다(Alpaydin, 2014). 또한, 서포트벡터머신은 구조적인 위험 최소화(Structural risk minimization)에 기초하기 때문에 과대적합문제에서 비교적 자유로우며 볼록 함수를 최소화 하는 학습을 진행하기 때문에 전역 최적해를 구할 수 있다는 장점이 있다(Han, 2009). 커널 함수는 대표적으로 Linear 함수, Polynomial 함수, Gaussian RBF(Radial Basis Function), 및 Laplacian 함수 등이 있다(Vapnik, 1995). 데이터의 특성에 따라 커널함수의 성능이 다르므로 본 연구에서는 4가지 커널 함수를 모두 사용하여 예측 성능을 비교하였다.

3. 실험 계획 및 결과

3.1 실험 계획

본 연구에서 제안하는 고객 불량률 예측 모델의 효용성을 입증하기 위해 국내 LCD 디스플레이 기업의 실제 데이터를 적용하였다. 특별히 모니터에 사용된 LCD 제품에 대해서 2011년부터 2013년까지 수집한 데이터를 사용하였으며 전처리 이후 관측치 수는 773개이다. 학습데이터는 2011년부터 2012년까지 수집한 464개의 관측치로서 데이터마이닝 알고리즘 별 예측 모델 구축에 사용하였고, 평가 데이터는 2013년 수집한 309개의 데이터로서 구축된 모델의 예측 성능을 평가하는데 사용하였다. 모델 구축을 위한 데이터 학습 시, 샘플링에 따른 Bias를 없애기 위해서 10-fold Cross Validation을 수행하였다. 이

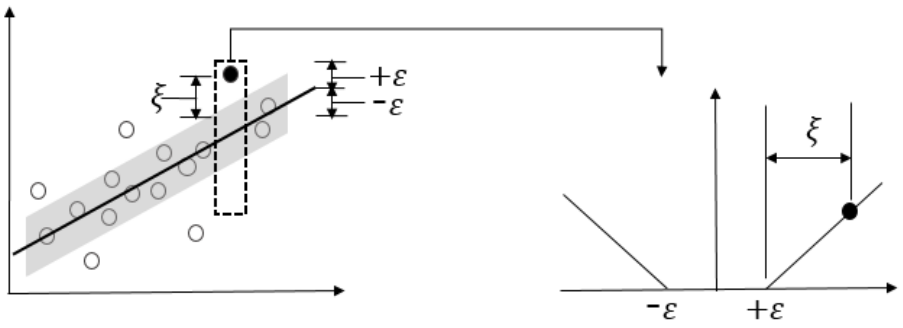

Figure 4. Linear regression function with epsilon intensive band and ϵ-intensive loss function

때 데이터마이닝 알고리즘 마다 가장 우수한 예측력을 보이는 매개변수를 최적의 매개변수 선정하여 예측 모델을 구축하였고, 모든 실험은 R에서 제공하는 패키지를 사용하였다. <Table 2>는 데이터마이닝 알고리즘 마다 예측 모델을 구축할 때 사용한 R패키지와 최적의 매개변수들을 정리한 표이다.

본 논문에서는 데이터마이닝 기반 예측 모델의 효과를 입증하기 위해서 두 가지 관점에서 기존 방법론의 예측 모델과 데이터마이닝 기반 예측 모델을 비교하였다. 첫 번째는, 모델의 예측력을 비교하였는데, 이때 예측 성능을 평가하는 척도는 여러 가지가 있으나 본 연구에서는 절대평균오차(이하 MAE)를 사용하였고 식 (3)과 같이 계산 할 수 있다.

$$\text{MAE} = 1/n \sum_{i=1}^{n} |y_i - \hat{y_1}| \quad (3)$$

여기서 $y_i - \hat{y_1}$는 실제 고객 불량률과 예측 고객 불량률의 차이를 의미하고, MAE가 낮을수록 우수한 예측력을 의미한다.

두 번째는, 예측 성능에 따른 손실비용을 실제 데이터의 가격 정보를 이용하여 비교함으로써 예측 성능 향상이 기업의 이익에 미치는 효과를 구체적으로 나타내었다. 서비스 재고에 관한 손실비용은 보증기간 동안 발생하는 실제 고객 불량률과 예측 고객 불량률이 동일 할 때 가장 적게 발생한다, 반면, 실제값과 예측값이 다른 경우에는 오차에 해당하는 만큼 손실비용이 발생한다. 손실비용에 대해서는 서론에서 언급하였으나 구체적으로 설명하면 다음과 같다. 기업에서는 예측된 고객불량 수만큼 서비스 재고를 준비하기 때문에 이에 대한 기본적인 손실비용은 예측 고객 불량률에 해당하는 수량과 원가의 곱으로 계산할 수 있다. 그러나, 예측값이 실제값 보다 크면 오차에 해당하는 수량만큼 재고가 남으므로 원가보다 낮은 가격으로 판매하고, 반대의 경우에는 재고가 부족하여 고객에게 원가보다 높은 판가를 보상하므로 오차만큼 손실비용이 증가할 수 있다. 이처럼 서비스 재고 관리 시 고객불량률 예측 오차에 따라 추가로 발생하는 비용을 본 논문에서 손실비용이라고 표현하였으며 이를 계산하는 방법은 식 (4)와 같다.

$$\text{Loss cost} = \begin{cases} m \times (a-b) & if\ y_i - \hat{y_1} < 0 \\ m \times (c-a) & otherwise \end{cases} \quad (4)$$

먼저, 각 관측치에 대해서 실제값과 예측값의 차이를 계산한다. 차이 값은 ppm로 나타낸 불량률이므로 판매 수량을 곱해서 불량 수(m)으로 변환한다. 예측값이 실제값 보다 큰 관측치는 m에 해당하는 수량만큼 재고가 남으므로, 남은 재고 수량(m)에 원가(a)와 저가(b)의 금액 차를 곱한다. 반면, 예측값이 실제값 보다 작은 관측치는 m에 해당하는 수량만큼 재고가 부족하므로, 부족한 수량(m)에 판가(c)와 원가(a)의 금액 차를 곱한다. 손실비용을 계산하는 데 필요한 제품 가격은 실험 데이터의 실제 가격 정보를 사용하였다.

Table 2. R Package and parameters of prediction models

Algorithm		R Package	Parameters
Support Vector Machines	Linear	e1701	Epsilon : 0.1 Cost : 4
	Gaussian RBF	e1701	Epsilon : 0.1 Cost : 4 Gamma : 0.0227
	Laplacian	kernlab	Epsilon : 0.02 Cost : 4 Sigma : 0.03
	Polynomial	kernlab	Epsilon : 0.02 Cost : 4 Scale : 5 Offset : 1 Degree : 1
Lasso		glmnet	Lambda : 2.63
Random Forests		randomForest	Number of trees : 10 Number of variables : 300
Decision Trees		rpart	Complexity parameter : 0.03
Artificial neural networks		neuralnet	Number of hidden layers : 2 Number of hidden neurons : 50, 10 Threshold : 0.7

3.2 실험 결과

<Table 3>은 모델 별 예측 성능을 비교한 표로서, 평가데이터에 데이터마이닝 기반 예측 모델과 기존 방법론의 모델에 대한 예측 성능 결과를 보여주고 있다. 기존 방법론으로는 시장반입데이터를 이용한 수명분포 기반 방법과 평균 불량 증가율을 이용한 방법을 비교하였다. 표에서 보여주듯이, 시장반입데이터를 이용한 수명분포 기반 방법의 경우 예측 성능이 MAE 기준 16,045ppm이고, 평균 불량 증가율 이용 방법은 예측 성능이 1,404ppm이었다. 수명분포함수 기반 방법을 사용한 경우 평균 불량 증가율 이용 방법을 사용한 경우 보다 예측 성능이 크게 좋지 않았는데, 이는 실험 데이터가 가정한 분포를 따르지 않기 때문으로 판단된다. 반면, 데이터마이닝 기반 예측 모델을 사용 했을 때는 서포트벡터머신의 경우 커널에 따라 817ppm, 824ppm, 845ppm, 847ppm이었고, 라쏘 889ppm, 랜덤포레스트 928ppm, 의사결정나무 1,057ppm, 인공신경망 1,290ppm의 예측 성능을 나타냈다. 모든 데이터마이닝 기반 예측 모델이 기존 방법 보다 우수한 성능을 보였고, 데이터마이닝 모델 중에는 서포트벡터머신의 예측력이 가장 우수했다. 서포트벡터머신의 경우, 본 연구에서 사용한 4가지 커널함수 중에는 Laplacian 함수를 사용한 경우에 가장 높은 예측력을 나타냈으나 다른 커널함수에 비해 그 차이는 크지 않았다.

Table 3. Prediction results from various models

Division	Algorithm		MAE(ppm)
Traditional model	Lifetime distribution based on field return data		16,045
	Average increasing rate of failures based on field return data		1,404
Data mining model	Support Vector Machines	Laplacian	817
		Gaussian RBF	824
		Linear	845
		Polynomial	847
	Lasso		889
	Random forests		928
	Decision trees		1,057
	Artificial neural networks		1,290

데이터마이닝 기반 모델의 예측 성능이 기존 기법 대비 얼만큼 향상 됐는지 평가하기 위해서 기존 예측 방법론 가운데 더 나은 성능을 보인 평균 불량 증가율 이용 방법과 비교하였다. 기존 방법 대비 향상 정도로 표현한다면 서포트벡터머신의 경우 커널 함수에 관계없이 MAE가 40% 이상 감소하였고, 라쏘 37%, 랜덤포레스트 34%, 의사결정나무 25%, 인공신경망 8%의 감소 효과가 있다고 볼 수 있다.

데이터마이닝 기반 방법의 예측 성능 향상이 통계적으로 유의미한지 분석하기 위하여 기존 방법론과 데이터마이닝 기반 방법들 가운데 각각 가장 좋은 성능을 나타낸 예측 모델을 평가데이터에 적용하여 절대오차 값으로 대응표본 t 검정을 실시하였다. 기존 방법론에서는 평균 불량 증가율을 이용한 예측 모델을, 데이터마이닝 기반 방법에서는 Laplacian 커널을 사용한 서포트벡터머신 예측 모델을 적용하여 309개 관측치에 대한 절대오차 값을 각각 구하였다. 예측 모델의 절대오차 값으로 대응표본 t 검정을 실시한 결과는 <Table 4>와 같으며, $e_{traditional}$은 평균 불량 증가율을 이용한 모델을 적용했을 때 관측치 별 절대오차이고, $e_{data\ mining}$은 Laplacian 커널을 사용한 SVM 모델을 적용했을 때 관측치 별 절대오차이다. 그 결과 P-value가 2.484×10^{-13}으로 5% 유의수준에서 데이터마이닝 기반 예측 모델의 예측력이 기존 방법론의 예측 모델보다 우수한 것을 확인 할 수 있었다.

마지막으로, 예측 성능 향상이 기업 이익에 미치는 효과를 평가하기 위해서 <Table 5>에서 보여주듯이, 평가데이터집합에 적용했을 때 발생한 예측 오차를 토대로 손실비용을 비교하였다. 이때 사용한 예측 모델은 유의성 검정과 마찬가지로 기존 예측 방법론과 데이터마이닝 기반 방법들에서 각각 가장 우수한 성능을 나타낸 평균 불량 증가율을 이용한 예측 모델과 Laplacian 커널을 이용한 서포트벡터머신 예측 모델을 대표로 사용하였다. 계산식은 제 3.1절의 식 (4)와 같다.

평균 불량 증가율을 이용한 예측 모델의 경우에는 손실비용이 229,554(USD)이고 Laplacian 커널을 이용한 서포트벡터머신 예측 모델의 경우에는 손실비용이 116,150(USD)으로 나타났다. 즉, 데이터마이닝 기반 예측 모델을 사용할 경우 기존 방법의 예측 모델 보다 손실비용이 113,404(USD)만큼 감소하였는데, 이는 49%의 절감 효과가 있음을 보여준다. 이는 보증기간 동안 고객불량률의 예측력 향상이 기업의 손실비용을 줄이는데 직접적인 영향을 미칠 뿐만 아니라 손실 감소 효과가 크므로, 본 연구가 기업의 이익 증가시키는데 크게 기여 할 수 있음을 증명한 것이다.

4. 결 론

통상적으로 LCD 디스플레이 산업에서는 보증기간 동안 발생하는 고객 불량률을 예측하기 위해서 수명분포함수나 평균 불량 증가율을 이용하였다. 하지만 이러한 방법은 알려진 확률분포를 가정하거나 기초 통계 분석에 그친 수준이고, 예측에 사용한 변수 역시 시장반입데이터만을 이용하기 때문에 예측

Table 4. Performance comparison between the traditional and data mining models by paired t-tests

	Paired Differences					t	df	P-value
	Mean	Std.Deviation	Std.Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
Paired $e_{traditional} - e_{data\ mining}$	527.39	1210.81	68.88	391.85	662.92	7.66	308	2.484×10-13

Table 5. Loss cost of each best performance model of traditional and data mining models

Division	Method	Loss Cost (USD)	Reducing Cost (reducing rate)
Traditional model	Using of average rate of failures in every size of product	229,554	113,404 (49%)
Data mining model	Support vector machines with laplacian kernel	116,150	

정확도에 한계가 있다. 본 연구에서는 기존 예측 방법론의 한계점을 보완하기 위해 크게 두 가지를 개선하였다. 첫 번째, 예측 모델 구축 시 확률 분포를 가정하지 않는 비모수적 데이터마이닝 방법론을 이용하였다. 두 번째, 시장반입데이터에 기존 예측 방법론에서 사용하지 않은 제품 속성 정보 12개를 추가로 사용하였다.

본 논문의 저자들이 조사한 바에 따르면, LCD 디스플레이 산업에서 제품 속성 정보를 사용하거나 데이터마이닝 알고리즘을 이용하여 고객 불량률을 예측한 사례가 없었으므로 본 연구가 최초의 응용 논문이라는 점에서 의의가 있다고 할 수 있다.

연구의 효용성을 검증하기 위해 실제 LCD 데이터를 이용하여 실험하였는데, 기존 예측 방법론과 데이터마이닝 기반 방법들의 대표 방법으로서 시장반입데이터를 이용한 수명분포함수 기반 방법과 평균 불량 증가율을 이용한 방법을 함께 비교하였다. 실험 결과, 데이터마이닝 기반 예측 모델들이 기존 방법론의 예측 모델보다 우수한 예측력을 보였는데, 특히 서포트 벡터머신 알고리즘을 사용했을 때 평균 불량 증가율을 이용한 기존 방법보다 MAE가 40% 이상 감소하였다.

또한, 고객 불량률 예측은 서비스 재고를 관리하는데 필수적인 활동이고 서비스 관련 손실비용에 직접적인 영향을 끼치므로 실험데이터를 이용하여 손실비용을 비교하였다. 기존 예측 방법론과 데이터마이닝 기반 예측 방법론에서 각각 가장 높은 성능을 보인 예측 모델을 비교한 결과, 데이터마이닝 기반 예측 모델을 사용했을 때 손실 비용이 기존 방법 대비 49% 절감되는 효과를 보였다. 이처럼 예측 성능 향상은 그 자체로도 의미가 있지만, 서비스 관련 손실비용을 감소시킴으로써 기업의 이윤 창출에 직접적으로 기여 할 수 있는 점에서 본 연구가 더욱 가치 있다고 할 수 있겠다.

본 논문에서 제안한 방법의 한계점으로는 고객 불량률을 예측하는 시점이 제품을 판매한 이후 6개월이 지난 시점인 점이다. 이는 기존 예측 방법을 기준으로 보면 가장 앞선 시점에 해당하지만, 예측 시점을 더 앞 당길수록 기업 입장에서는 더 많은 이익과 가치를 창출 할 수 있을 것이다. 따라서, 고객 불량률의 예측 시점을 제품을 판매하기 이전으로 앞당기기 위한 시도를 할 예정이며 이 경우 제조 공정의 계측데이터와 개발 단계에서 평가된 신뢰성 데이터를 사용할 예정이다.

참고 문헌

Alpaydin, E. (2014), *Introduction to machine learning*, MIT press, Massachusetts, United States, **2**.

Bae, S. J. (2004), *An Analysis of Success Factors on LCD Display Industry in Korea*, Hannam University, Korea.

Bae, S.-M., Lee, H.-W., Lee, G.-A., Choi, S., and Park, H.-K. (2007), Enhancing Manufacturing Data Quality for Data Mining, *Journal of the Korean Society for Precision Engineering*, **6**, 795-796.

Breiman, L. (2001), Random forests, *Machine learning*, **45**(1), 5-32.

Breiman, L., Friedman, J., Stone, C.-J., and Olshen, R.-A. (1984), Classification and regression trees, *CRC press.*

Curram, S.-P. and Mingers, J. (1994), Neural networks, decision tree induction and discriminant analysis : An empirical comparison, *Journal of the Operational Research Society*, **45**(4), 440-450.

Deng, H. and Runger, G. (2013), Gene selection with guided regularized random forest, *Pattern Recognition*, **46**(12), 3483-3489.

Han, H.-Y. (2009), *Introduction of recognizing patterns : training with MATLAB*, Hanbit Media, Inc., Seoul, Korea, **2**.

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The elements of statistical learning*, Springer, Berlin, Germany.

Hwang, W.-H., Kim, J.-H., Jang, W.-S., Hong, J.-S., and Han, D.-S. (2008), Fault Pattern Analysis and Restoration Prediction Model Construction of Pole Transformer Using Data Mining Technique, *The Transactions of*

The Korean Institute of Electrical Engineers, **57**(9), 1507-1515.

Kim, K.-J. and Lee, W.-B. (2004), Stock market prediction using artificial neural networks with optimal feature transformation, *Neural Computing and Applications*, **13**(3), 255-260.

Lee, D. K. (2005), Reliability Prediction using Telcordia SR-332 in Electric Home Appliance, *Journal of Applied reliability*, **5**(4), 427-438.

Lee, S.-H. and Chung, H.-J. (2009), Study on Lifetime Estimation of TFT-LCD Modules by Temperature Stress, *Journal of Korean institute of information*, **7**(5), 1-7.

Lee, Y.-J. and Lee, C.-L. (2006), The impact on profitability and customer satisfaction and corporate values, *Journal of Korea marketing association*, **21**, 85-113.

Liaw, A. and Wiener, M. (2002), Classification and regression by random forest, *R news*, **2**(3), 18-22.

Linoff, G.-S. and Berry, M.-J. (2011), *Data mining techniques : for marketing, sales, and customer relationship management*, John Wiley and Sons.

Moon, D. K. (2012), *Issues and Countermeasures in the display industry*, Soonchunhyang university, Asan, Korea.

Murray, J.-F., Hughes, G.-F., and Kreutz-Delgado, K. (2005), Machine learning methods for predicting failures in hard drives : A multiple-instance application, *Journal of Machine Learning Research*, **6**, 783-816.

Nagappan, N., Ball, T., and Zeller, A. (2006), Mining metrics to predict component failures, *Proc. 28th international conference on Software engineering*, 452-461.

Park, C. K. (2006), Estimating Software Development Cost USING Support Vector Regression, *The Korean Operations Research and Management Science Society*, **23**(1), 75-91.

Park, H.-I., Kwon, G.-C., and Oh, S.-B. (2005), Prediction of Modulus of Subgrade Soils and Subbase Materials Based on Artificial Neural Network Model, *Journal of the Korean Society of Civil Engineers*, **25**(2C), 61-71.

Park, S. H. (2015), Analysis of industry environment of Korean display materials and policy trends, *Journal of science and technology policy institute*, **25**(2), 32-37.

Rosenblatt, F. (1958), The perceptron : a probabilistic model for information storage and organization in the brain, *Psychological review*, **65**(6), 386.

Seo, D. H. (2014), *e-KIET Issues and analysis*, Korea Institute for Industrial Economics and Trade, Seoul, Korea, **580**.

Shin, K.-Y., Ji, J.-G., Han, J.-H., Lee, D.-G., Son, Y.-J., and Lee, H.-S. K. (2011), Life Analysis and Reliability Prediction of Relays based on Life Prediction Method, *Journal of the korean society for railway*,1327-1335.

Sholkopf, B. and Smola, A. (2002), *Learning with Kernels*, MIT Press, Cambridge MA.

Tibshirani, R. (1996), Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society*, Series B (Methodological), **58**(1), 267-288.

Vapnik, V. (1995), *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, United States, **2**.

Wang, F.-K. and Chu, T.-P. (2011), The mean time between failures for an LCD panel, *Quality and Reliability Engineering International*, **27**(2), 203-208.

Wilson, S., Joyce, T., and Lisay, E. (2009), Reliability estimation from field return data, *Lifetime data analysis*, **15**(3), 397-410.

Yoon, J.-I., Seo, H.-S., and Yim, C.-S. (2003), A study on customer satisfaction evaluation framework for mobile services, *Journal of the Korean Institute of Industrial Engineers*, **15**, 170-174.

Yum, B.-J., Seo, S.-K., Yun, W.-Y., and Byun, J.-H. (2014), Trends and Future Directions of Quality Control and Reliability Engineering, *Journal of the Korean institute of industrial engineers*, **40**(6), 526-554.