J. Soc. Korea Ind. Syst. Eng Vol. 40, No. 2 : 129-136, June 2017
https://doi.org/10.11627/jkise.2017.40.2.129
ISSN : 2005-0461(print)
ISSN : 2287-7975(online)

Tolerance Computation for Process Parameter Considering Loss Cost : In Case of the Larger is better Characteristics

Yong-Jun Kim* · Geun-Sik Kim** · Hyung-Geun Park***†

*Department of Industrial Management, Gyeong-gi College of Science and Technology
**Department of Communication and Information, Inha University
***Department of Industrial Management, Shin Ansan University

손실 비용을 고려한 공정 파라미터 허용차 산출 : 망대 특성치의 경우

김용준* · 김근식** · 박형근***†

*경기과학기술대학교 산업경영과
**인하대학교 언론정보학과
***신안산대학교 산업경영과

Among the information technology and automation that have rapidly developed in the manufacturing industries recently, tens of thousands of quality variables are estimated and categorized in database every day. The former existing statistical methods, or variable selection and interpretation by experts, place limits on proper judgment. Accordingly, various data mining methods, including decision tree analysis, have been developed in recent years. Cart and C5.0 are representative algorithms for decision tree analysis, but these algorithms have limits in defining the tolerance of continuous explanatory variables. Also, target variables are restricted by the information that indicates only the quality of the products like the rate of defective products. Therefore it is essential to develop an algorithm that improves upon Cart and C5.0 and allows access to new quality information such as loss cost. In this study, a new algorithm was developed not only to find the major variables which minimize the target variable, loss cost, but also to overcome the limits of Cart and C5.0. The new algorithm is one that defines tolerance of variables systematically by adopting 3 categories of the continuous explanatory variables. The characteristics of larger-the-better was presumed in the environment of programming R to compare the performance among the new algorithm and existing ones, and 10 simulations were performed with 1,000 data sets for each variable. The performance of the new algorithm was verified through a mean test of loss cost. As a result of the verification show, the new algorithm found that the tolerance of continuous explanatory variables lowered loss cost more than existing ones in the larger is better characteristics. In a conclusion, the new algorithm could be used to find the tolerance of continuous explanatory variables to minimize the loss in the process taking into account the loss cost of the products.

Keywords : Continuous Variable, Decision Tree, Larger-is-Better Characteristics, Loss Cost, Tolerance

Received 29 May 2017; Finally Revised 22 June 2017;
Accepted 23 June 2017
† Corresponding Author : pitt690309@daum.net

1. 서 론

최근에는 정보기술이 발달함에 따라 기업에서 제품과 고객 등에 관한 방대한 규모의 데이터베이스를 구축하게 되었고, 효율적인 의사결정을 위하여 데이터베이스 내 대량의 데이터를 효과적으로 분석하여 정보화하려는 노력을 하고 있다[1]. 정보화 시대에 경쟁사보다 정보 획득의 속도에서 앞서기 위해서는 현재의 데이터나 정보 등을 통하여 또 다른 새로운 정보나 지식을 얻는 방법이 경쟁력을 확보할 수 있는 좋은 수단이라고 할 수 있다[4]. 제조업 분야에서는 컴퓨터의 발달과 기술력의 증대로 공정시스템이 자동화됨에 따라서 하루에도 수천개의 품질 특성치들이 계측됨은 물론 실시간으로 공정의 상태를 파악하여 관리하고 있다[6].

기업의 데이터베이스에서 유용한 정보를 효과적으로 도출해 내기 위해 데이터 마이닝 기법이 활용되어지고 있다. 품질관리 분야에서는 “품질 마이닝”이라는 용어가 생길만큼 최근에는 제조 공정에서 품질관리나 공정관리를 위해 많이 적용되어지고 있다[5]. 특히 의사결정나무 분석은 적용하고자 하는 데이터의 유형에 크게 상관이 없으며, 얻어진 규칙의 해석이 용이하기 때문에 많이 사용되는 데이터 마이닝 기법이다.

제조 공정에서 데이터 마이닝 적용에 대한 연구는 두 가지 방향으로 나누어질 수 있다. 하나는 데이터 마이닝 기법을 제조 공정에 적용하는 사례 연구이며, 다른 하나는 데이터 마이닝 기법들의 성능을 비교하거나 제조 공정에 적합한 알고리즘의 개발과 관련된 연구이다. 첫 번째 경우 국내 문헌으로는 Woo et al.[13]는 의사결정나무 분석을 이용하여 제품 불량 발생에 영향을 미치는 주요 공정변수를 파악하고, 생성된 양품 및 불량 발생 관련한 공정 변수들의 규칙들을 해석하여 양품을 안정적으로 생산할 수 있는 공정변수를 선택하였다. Shin[10]은 의사결정나무 기법을 이용하여 하드 디스크 제조 공정상에서 발생하는 품질의 문제점을 발견하고자 하였으며, 품질에 영향을 주는 요소를 추출하였다. 국외 문헌으로 Li et al.[7]은 전통적인 중국의 제약 생산 공정에서 클러스터 분석인 데이터 마이닝과 시각화 기술을 이용하여 공정에 영향을 미치는 요소들을 찾아내고 분석하는 방법을 제안하였다. Ronowicz et al.[9]는 의사결정나무를 이용하여 총알 생산에 영향을 미치는 변수들과 주요 특성치들의 연관관계를 분석하여 총알의 구형에 영향을 미치는 공정 변수들을 찾아내고 분석하였다.

두 번째 경우 국내 문헌으로는 Sim and Kim[12]은 PCB 제조공정에서 제품의 수율을 향상시키기 위해서 데이터 마이닝 기법을 이용하여 저수율의 원인이 되는 불량요인을 파악하였으며, 불량요인에 영향을 미치는 중요 공정 및 설비를 찾는 방법을 제안하였다. Jung and Lee[3]은 의사결정나무 학습 기법을 이용하여 다변량의 공정 관리 절차를 소개하고, 이변량일 경우 모의실험을 통하여 그 효율성을 검증하였다. 국외 문헌으로는 Sim et al.[11]은 PCB 생산공정에서 고장 규칙 마이닝을 이용하여 기계의 순서와 불량의 유형을 고려한 최적 공정절차를 제안하였다. Chien et al.[2]은 반도체 공정에서 자동적으로 수집된 많은 양의 데이터를 이용하여 실험계획법 데이터 마이닝을 주장하였다. Purr et al.[8]은 자동차 차체를 생산하는 금형 공정에서 품질관리를 위한 데이터를 얻기 위해 데이터 마이닝을 이용하였다.

기존 문헌들을 살펴본 결과 첫째, 제조 산업에서 제품 설계 단계의 데이터를 이용한 제품의 특성치와 설명변수들 간의 관계를 분석한 연구들이 미미하다는 것이다. 둘째, 데이터 마이닝 기법을 이용하는 경우 특성치에 영향을 미치는 주요 변수를 찾는 것에 집중하여 변수 허용차에 관한 부분은 연구가 부족하였다. 셋째, 목표변수가 적합품률이나 부적합품률 등의 품질 자체의 정보에만 한정되어 품질 비용과 같이 경제적 손실에 관한 부분은 접근하지 못하였다는 것이다.

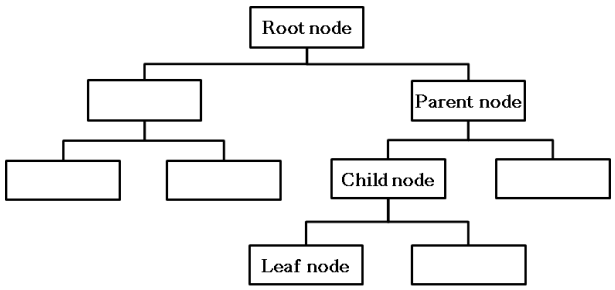
본 연구에서는 이러한 한계점을 극복하고자 제품 설계 단계의 데이터를 이용하여 경제적 손실비용을 목표변수로 설정하여 이를 최소화하는 설명변수들의 최적 허용 범위를 구하는 알고리즘을 개발하였다. 이 알고리즘에 의해 다음과 같은 해결방안을 제시하고자 한다. 첫째, 통계적 기법들로 처리하기 힘든 많은 데이터를 데이터 마이닝 기법을 이용하여 주요 설명변수를 선택함으로써 관련 전문가의 주관적인 판단에 의존하는 것을 해소할 수 있다. 둘째, 경제적 손실비용에 영향을 미치는 연속형 설명변수들의 허용 범위를 구함으로써 최적 공정조건을 유지하는데 도움이 될 수 있다. 셋째, 손실비용을 목표변수로 설정함으로써 기업의 관리자로 하여금 제품의 재무적인 문제에 접근할 수 있도록 하였다.

본 연구의 나머지 부분은 다음과 같이 구성되어 있다. 제 2장은 이론적 배경으로 데이터 마이닝의 개요, 의사결정나무 분석의 개요와 알고리즘, 다구찌의 손실함수에 대해 살펴보았다. 제 3장에서는 품질 마이닝 기법인 의사결정나무를 이용하여 제품 설계 단계의 데이터를 분석하는 알고리즘의 절차와 방법에 대하여 설명하였다. 제 4장에서는 알고리즘 검증을 위한 시뮬레이션에 대해 설명하였다. 마지막으로 제 5장에서는 3장과 4장을 통해 도출된 결과를 요약하고, 이에 대한 시사점과 향후 연구 방향을 제시하였다.

2. 이론적 배경

2.1 의사결정나무 분석

데이터 마이닝의 대표적인 기법인 의사결정나무는 하나의 나무구조로 이루어져 있으며, 각각의 구성요소는 <Figure 1>과 같다.



<Figure 1> The Structure of Decision Tree

의사결정나무는 <Figure 1>과 같이 어떠한 분할 기준에 의하여 뿌리 마디에서 마지막 잎 마디를 거치게 된다. 의사결정나무 분석에서 목표변수가 이산형인 경우에는 분류나무를 구성한다고 말하며, 연속형인 경우 회귀나무를 구성한다고 말한다. 이산형 목표변수의 경우 카이제곱 통계량, 지니 지수, 엔트로피 지수 등을 기준으로 하여 분할되며, 연속형 목표변수의 경우 분산분석의 F-검정값이나 분산의 감소량 등을 기준으로 분할된다.

2.1.1 C5.0 알고리즘

C5.0은 의사결정나무 알고리즘 중 하나로 Ross Quinlan이라는 호주 학자에 의해 개발된 이론이다. 이 알고리즘은 모든 가능한 분리 기준 값에서 이득 비율(gain ratio)을 계산하고, 그 이득 비율이 최대가 되는 값을 최종 분리 기준 값으로 선택하는 방법이다. 정보량(information), 엔트로피 계수(entropy index), 분리 정보(split information), 이득 표준(gain criterion)에 의해 이득 비율이 계산된다. 정보량은 전달되는 메시지의 확률(P)에 좌우되는 값으로 $-\log_2(P)$로 계산되는 값이다. 엔트로피 계수는 정보량을 일반화시켜 부르는 말로 $P=(p_1, p_2, \cdots, p_n)$라는 확률분포를 가정하면 전달되는 정보를 엔트로피 P라고 부르게 되며, 엔트로피 P는 다음과 같이 식 (1)로 표현할 수 있다.

$$I(P) = -(p_1 \times \log_2(p_1) + \cdots + p_n \times \log_2(p_n)) \quad (1)$$

분리 정보는 데이터가 여러 개의 부분집합으로 분할될 때 추가적으로 발생하는 정보량을 의미하고, 이득 표준은 데이터의 분할을 통해 감소한 정보량의 크기를 나타낸다. 최종적으로 이득 비율은 이득 표준과 분리 정보의 비로 정의된다.

2.1.2 CART 알고리즘

CART 알고리즘은 지니 지수(gini index)를 이용하여 이지분리(binary split)를 수행하는 알고리즘이다. 이지분리는 부모마디로부터 자식마디가 형성될 때 2개의 자식마디만이 형성되는 것을 의미한다. 지니 지수는 마디의 순수함 정도를 재는 측도로 종속변수가 가지는 값 중에서 어느 한 쪽의 값으로만 구성되어 있는 마디일수록 순수하다고 판단한다.

2.2 다구찌의 손실함수

다구찌는 품질을 "제품이 출하되어 사용되어질 때 성능 특성치의 변동으로 인해 사회에 끼치는 유·무형의 총 손실이며 다만 기능 그 자체에 따른 손실은 제외된다."라고 정의하였다. 간단히 말하면 품질을 제품의 사회적 손실로 정의하였다. 사회적 손실의 개념은 원하는 품질 특성의 제품을 만들지 못함으로 인해 발생되는 손실을 생산자 혹은 소비자 중 어느 한쪽이 책임을 지는 것이 아니라 사회가 그 책임을 지는 현대적 개념을 말한다[12].

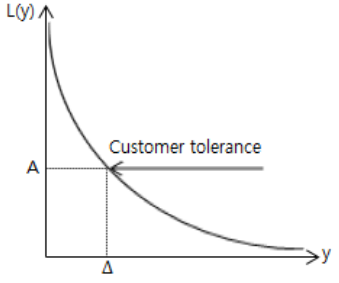
다구찌는 특성치가 연속적인 양의 값을 갖는다고 가정하였을 때 품질특성치를 망목특성(nominal-is-best characteristics), 망소특성(smaller-is-better characteristics), 망대특성(larger-is-better characteristics)의 3가지 경우로 분류하였다. 망목특성은 특정한 목표치가 주어져 있는 경우로서 길이나 두께 등과 같이 특성치의 값이 주어진 목표치에 가까우면 가까울수록 좋은 경우의 특성이다. 망소특성은 소음이나 불순물함량 등과 같이 특성치의 값이 음의 값을 갖지 않으며 이상적인 목표치가 0인 경우의 특성이다. 망대특성은 강도와 수율 등과 같이 특성치의 값이 음의 값을 갖지 않으며 이상적인 목표치가 무한대인 경우의 특성이다.

2.2.1 망대특성의 손실함수

망대 특성치의 경우 제품 특성치 x를 $1/x$로 변환하여 목표치 $m=\infty$으로 하는 경우로서 손실함수 $L(y)$를 다음 식 (2)와 같이 정의한다.

$$L(y) = k(\frac{1}{y} - 0)^2 = \frac{k}{y^2} \quad (2)$$

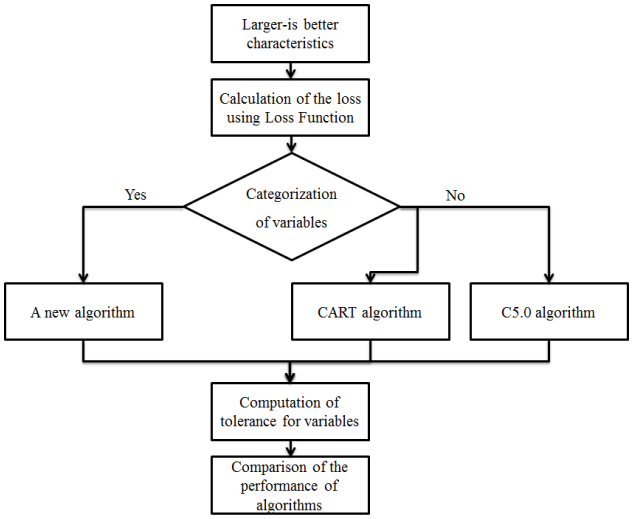
여기서 k는 상수이며 $A\triangle^2$으로 구할 수 있다. $\triangle$는 허용한계를 의미하며, A는 소비자 허용한계인 $\triangle$에서 소비자의 손실을 의미한다. 이를 그림으로 표현하면 다음 그림 <Figure 2>와 같다.



<Figure 2> Loss Function for Larger-is Better Characteristics

3. 연구 설계

본 연구에서는 제품의 설계 단계에서 제품 특성치와 이에 영향을 미치는 연속형 설명변수의 정보를 이용하여 목표변수인 경제적 손실비용을 계산하였다. 이 때 다구찌의 손실함수를 적용하였으며, 망대 특성치에 대하여 손실 비용을 계산하였다. 다음으로 새로운 알고리즘을 개발하여 손실비용에 영향을 미치는 설명변수를 찾아내는 동시에 이를 최소로 하는 설명변수의 최적 범위를 구하였다. 마지막으로 새로운 알고리즘의 성능을 입증하기 위하여 기존의 다양한 문헌에서 제품의 품질관리를 위해 사용하는 의사결정나무 분석 알고리즘을 그 비교 대상으로 삼았다. 성능 비교의 대상이 되는 알고리즘은 CART 알고리즘과 C5.0 알고리즘으로 본 논문에서 제안한 알고리즘과 성능 비교를 진행하였다. 성능 비교를 위한 일련의 절차를 그림으로 나타내면 다음 <Figure 3>과 같다.



<Figure 3> The Process of Analysis

3.1 경제적 손실비용의 계산

망대 특성치의 경우 손실함수 $L(y) = A\triangle^2(\frac{1}{y^2})$(단, $\triangle$는 허용차, A는 소비자 허용한계인 $\triangle$에서 소비자 손실)를 이용하였다. y는 수명, 효율, 강도 등과 같이 이상적인 목표치가 무한대인 경우 특성치가 될 수 있다.

3.2 알고리즘의 적용

단계 1 : 제품의 특성치에 영향을 미치는 각 연속형 설명변수 데이터에 대해 훈련용 데이터 세트와 검증용 데이터 세트로 각각 구분한다. 일반적으로 비율은 70:30 비율이 적당하나, 데이터가 전체적으로 부족한 경우 80:20 비율로 나누게 되면 그 성능이 향상된다.

단계 2 : 단계 1에서 발생된 훈련용 데이터 세트를 이용하여 각 연속형 설명변수의 데이터를 크기 순서대로 정렬한다.

단계 3 : 이득 비율이 최대화되도록 각 설명변수를 3개의 구간으로 범주화한다. C5.0 알고리즘의 절차를 응용한 연속형 설명변수의 범주화는 다음과 같이 5단계를 거쳐 이루어진다. 범주화 과정에 필요한 기호는 D : 데이터의 집합, C_j : 목표변수의 j번째 범주, $|D|$: D에 속한 총 개체 수, $P(D, j)$: D에서 목표변수의 j번째 범주에 속하는 개체의 비율이다.

① 데이터의 집합 D에서 목표변수의 j번째 범주 C_j에 속하는 개체를 구별하기 위한 정보량을 나타내는 엔트로피계수 $Info(D)$를 다음 식 (3)과 같이 계산한다.

$$Info(D) = -\sum_{j=1}^{k} P(D, j) \times \log_2(P(D, j)) \quad (3)$$

② 분리 기준값에 의해 데이터 D는 3개의 부분집합으로 분할된다. 이 때 얻어지는 정보량은 각 부분집합에서 정보량의 가중평균이 된다. 이를 모든 분리 기준 값에서 식 (4)와 같이 계산한다.

$$Info_X(D) = \sum_{i=1}^{3} \frac{|D_i|}{|D|} \times Info(D_i) \quad (4)$$

③ 정보량의 감소를 나타내는 정보량의 이득 표준을 식 (5)와 같이 계산한다.

$$Gain(D, X) = Info(D) - Info_X(D) \quad (5)$$

④ 데이터 D가 n개의 부분집합으로 분할될 때 추가적으로 발생하는 정보량인 분리정보를 다음과 같이 식 (6)과 같이 계산한다.

$$Split(D, X) = -\sum_{i=1}^{3} \frac{|D_i|}{D} \times \log_2(\frac{|D_i|}{|D|}) \quad (6)$$

⑤ $Gain(D, X)$를 $Split(D, X)$로 나눈 값인 $Gain\ \ ratio$를 식 (7)과 같이 계산한 후, 모든 경우에서 분리 기준값이 가장 큰 경우를 선택한다.

$$Gain\ ratio(D, X) = \frac{Gain(D, X)}{Split(D, X)} \quad (7)$$

3.3 주요 변수의 허용차 산출

목표변수인 손실비용에 영향을 미치는 연속형 설명변수를 찾기 위해 의사결정나무 분석을 실시하였다. 이를 통해 목표변수에 주로 영향을 미치는 설명변수를 찾아내었으며, 소비자 허용한계인 △에서 소비자의 손실보다 작은 범주를 만족시키는 설명변수의 구간을 파악하였다.

3.4 알고리즘의 성능 비교

알고리즘의 성능을 비교하기 위하여 본 논문에서 제시한 알고리즘, CART와 C5.0 알고리즘으로 도출한 소비자 허용한계인 △에서 소비자의 손실보다 작은 범주를 만족시키는 주요 설명변수의 허용차를 비교하였다. 각 알고리즘에 의해 구해진 주요 설명변수의 허용차에서 손실비용을 최소화하는 알고리즘을 더 우수한 성능이 있다고 판단하였다.

4. 시뮬레이션

연속형 설명변수의 데이터는 일정한 구간에서 랜덤하게 각 변수마다 1,000개 데이터가 생성되었으며, 손실비용을 구하기 위한 제품 특성치 데이터는 임의의 공정 상황을 고려하여 다중 선형 회귀식 $24.41+0.72x_1+0.23x_2+0.10x_3-0.43x_4$에 의하여 생성되었다. 특정한 상황에 대한 적용을 배제하기 위해 동일 난수를 가정하지 않은 10개 세트를 생성하였다. 각 연속형 설명변수의 구간은 다음 <Table 1>과 같다.

<Table 1>에서 연속형 설명 변수의 데이터를 랜덤하게 발생하였으며, 망대 특성치는 강도를 가정하였다. 프로그래밍 R에서 랜덤하게 구현된 4개의 연속형 설명변수에 따른 특성치 데이터(y) 5개를 나타내면 <Table 2>와 같다.

실험은 Intel(R) Core(TM) i5-2400 CPU, 3.10GHz, RAM 4.00GB 환경에서 수행되었으며, 프로그래밍 R을 이용하여 분석을 수행하였다. 실험은 1,000개 데이터 세트에 대해 10회를 반복하여 실험하였다.

<Table 1> The Section of Explanatory Variables

Variable	Section
temperature	100℃~200℃
pascal	30Pa~50Pa
hour	0.5Hr~3.5Hr
humidity	20%~60%

<Table 2> Simulation Data(Seed Number 30)

	x_1(℃)	x_2(Pa)	x_3(Hr)	x_4(%)	y
1	109.87	46.27	0.99	34.09	99.60
2	148.82	41.47	3.26	44.55	122.27
3	136.40	47.14	2.62	31.74	120.08
4	142.06	37.60	2.43	42.87	117.15
5	130.10	32.69	2.26	29.37	113.19

4.1 경제적 손실비용의 계산

단계 1 : 강도의 목표치(m) = ∞kgf로 가정하며, 허용차(△) = 100kgf, 허용한계에서 손실비용(A) = 3,000원이라 가정하여 목표변수인 손실비용을 계산하였다. 각 제품의 군에서 손실함수를 이용하여 손실비용을 계산하고, 허용차(△)에서 손실비용(A) = 3,000원보다 작은 범주를 '범주 1'로, 큰 범주를 '범주 2'로 구분하였다. 다음 <Table 3>은 단계 1을 수행한 첫 5개 데이터의 결과를 나타낸 것이다.

<Table 3> The Category of the Loss

	Loss	Category
1	3,024	2
2	2,006	1
3	2,080	1
4	2,185	1
5	2,341	1

4.2 알고리즘의 적용

단계 1 : 각 연속형 설명변수 데이터에 대하여 훈련용과 검증용 데이터 세트로 각각 구분하였다. 변수별 1,000개의 데이터 세트에서 훈련용 70%와 검증용 30% 데이터로 분류하였다. 이 때, 데이터는 중복이 가능하도록 설정하였다.

단계 2 : 단계 1에서 얻은 훈련용 데이터 세트에서 각 연속형 설명변수의 데이터를 크기 순서대로 정렬하였다. <Table 4>는 x_1 설명변수에 대하여 크기 순서대로 정렬한 첫 5개 데이터를 나타내었다.

<Table 4> Sorting of x_1 Variable

	x_1(℃)	x_2(Pa)	x_3(Hr)	x_4(%)	y(kgf)
1	100.01	38.98	0.61	46.44	85.48
2	100.07	30.71	0.70	24.54	93.04
3	100.11	33.33	2.26	49.91	82.92
4	100.14	48.84	0.68	42.71	89.45
5	100.17	43.78	1.32	37.06	90.80

단계 3 : 이득 비율이 최대화되도록 각 설명변수를 3개의 구간으로 범주화한다. 식 (3)~식 (7)에 의해 각 값들이 계산되어, 각 설명변수에서 이득비율이 최대가 되는 3개의 구간을 계산하면 <Table 5>와 같다.

<Table 5> The Section of Each Variable

	section 1	section 2	section 3
x_1	x_1 < 108.983	108.983 ≤ x_1 ≤ 119.9939	x_1 > 119.9939
x_2	x_2 < 35.84046	35.84046 ≤ x_2 ≤ 42.42317	x_2 > 42.42317
x_3	x_3 < 1.217878	1.217878 ≤ x_3 ≤ 2.319019	x_3 > 2.319019
x_4	x_4 < 35.1606	35.1606 ≤ x_4 ≤ 50.60412	x_4 > 50.60412

4.3 주요 변수의 허용차 산출

목표변수인 손실비용에 주로 영향을 미치는 설명변수를 찾기 위하여 의사결정나무 분석을 실시하였다. 이 때, 소비자 허용한계(△)에서 소비자의 손실보다 작은 데이터의 범주를 '범주 1'로 하여 이를 만족시키는 설명변수의 허용차를 파악하였다. 의사결정나무 분석을 수행한 결과 <Table 6>과 같다.

<Table 6> The Result of Decision Tree

x_1 in (2) : category 1(767/8) x_1 in (0, 1) : mis classification ratio : 2.5%		
truth \ prediction	category 1	category 2
category 1	808	13(mis)
category 2	12(mis)	167

<Table 6>을 통해 x_1 변수가 손실비용에 영향을 주는 주요 변수라는 사실을 알 수 있으며, x_1이 구간 x_1 > 119.9939를 만족할 때 목표변수의 '범주 1'을 만족시킴을 알 수 있다.

4.4 알고리즘의 성능 비교

본 논문에서 제시한 알고리즘, CART, C5.0 알고리즘의 성능을 비교하였다. 소비자 허용한계(△)에서 소비자의 손실보다 작은 범주인 '범주 1'을 만족시키는 주요 설명변수의 허용차에서 손실비용을 비교하였다.

CART 알고리즘 분석의 경우 일정 범위의 변수 허용차를 구하지 못하여 비교 대상에서 제외하였다. 따라서 C5.0 알고리즘의 분석 결과를 제안한 새로운 알고리즘과 비교하여 그 성능을 비교하였다. 목표변수의 '범주 1'로 분류한 x_1 변수의 허용차 구간에서 손실비용 차이를 비교하기 위하여 t-검정을 실시한 결과 <Table 7>과 같다.

<Table 7> The Comparison between C5.0 and the New Algorithm

classification	mean	standard error	t	p-value
C5.0	1811.21	20.975	-3.074	.002
new	1724.62	18.803		

<Table 7>의 결과 t = -3.074이며 두 알고리즘의 손실비용이 p-value = 0.002로 통계적으로 매우 유의한 차이가 있다. 따라서 새로운 알고리즘에서 구한 설명변수의 허용차가 C5.0 알고리즘에 비해 목표변수인 손실비용을 더 작게 한다는 것을 알 수 있다.

본 논문에서는 다양한 seed number를 부여하여 랜덤하게 데이터를 발생하였다. 새로운 알고리즘과 C5.0 알고리즘을 이용하여 얻어진 주요 연속형 설명변수는 다음 <Table 8>과 같다.

<Table 8> The Major Explanatory Variables

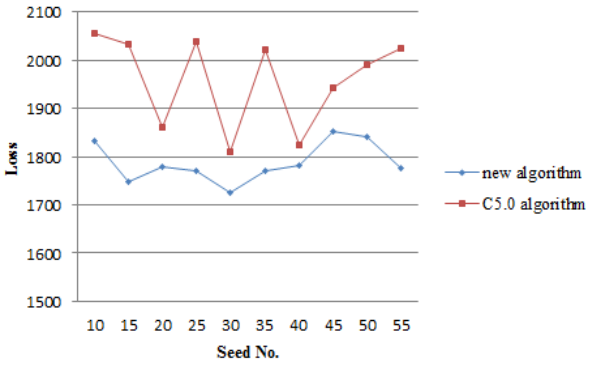
seed No.	10	15	20	25	30	35	40	45	50	55
variables	x_1	x_1	x_1	x_1	x_1	x_1	x_1	x_1	x_1	x_1

또한, 새로운 알고리즘과 C5.0 알고리즘에서 구한 주요 연속형 설명변수의 허용차에서 손실비용 평균을 구해보았다. 시뮬레이션은 각 seed number에서 1,000개 데이터를 생성하여 분석하였다. 10가지 경우에서 구한 손실비용의 t-검정 결과 p-value 값은 다음 <Table 9>와 같다.

<Table 9> The p-Value in Various Seed Numbers

seed No.	new algorithm	C5.0 algorithm	p-value
10	1832.93	2054.91	.000
15	1747.68	2031.83	.000
20	1777.64	1860.49	.000
25	1770.30	2039.42	.000
30	1724.62	1811.21	.000
35	1771.19	2020.72	.000
40	1782.16	1823.81	.176
45	1853.17	1941.83	.006
50	1841.93	1991.61	.000
55	1777.31	2023.21	.000

각 seed number 모든 경우에서 새로운 알고리즘에서 구한 주요 연속형 설명변수의 손실비용이 C5.0 알고리즘에 비해 유의하게 평균이 작다는 것을 알 수 있다. seed number별 계산된 손실비용의 평균을 그림으로 표현하면 <Figure 4>와 같다.



<Figure 4> Loss in Various Seed Numbers

5. 결 론

본 논문에서는 C5.0 알고리즘 이론을 응용한 연속형 설명변수의 3개 범주화를 생성하는 알고리즘을 제안하였다. 목표변수는 손실비용을 이용하였으며, 망대 특성치를 고려하였다. 새로운 알고리즘은 목표변수에 영향을 미치는 주요 변수를 파악하며, 최적의 허용차를 구하는데 초점을 맞추었다. 시뮬레이션을 통한 검증 결과, 새로운 알고리즘이 C5.0 알고리즘에 비해 손실비용을 더 작게 하는 연속형 설명 변수의 허용차를 구할 수 있음을 확인하였다.

연구결과를 통해 본 연구의 시사점은 다음과 같다. 첫째, 본 연구는 기존 연구에서 많이 다루어지지 않은 제품의 설계 단계에서 나오는 데이터를 분석하여 유의한 결과를 도출하였다는 점이다. 둘째, 제품의 품질을 파악함에 있어 적합품률, 부적합품률 등의 단순 품질 정보에서 벗어나 품질의 변동에 따른 경제적 손실비용에 대한 정보를 목표변수로 설정하여 기업의 관리자로 하여금 제품의 재무적인 정보에 접근할 수 있도록 하였다. 셋째, 연속형 설명변수의 범주화를 시도하여 CART와 같은 이진분리 알고리즘의 한계점을 극복할 수 있었으며, C5.0과 같이 주요 변수를 찾는 것에 중점을 두는 다지 분리 알고리즘보다 더 좋은 허용차를 구할 수 있었다는 점이다.

본 연구는 망대 특성을 가진 단일 특성치에 대한 경제적 손실 비용을 계산하고, 연속형 설명변수의 최적 허용차를 연구하였다는데 그 의미가 있으나 다음과 같은 추가적인 검토가 필요하다. 첫째, 임의적으로 발생한 시뮬레이션 데이터를 통해 알고리즘의 성능을 정확하게 판단하는 것은 무리가 있으므로 실제 현장의 데이터를 통한 분석이 필요할 수 있다. 둘째, 규격의 이론에 기초한 3개 범주가 아닌 체계적으로 범주를 선정하여 판단하는 방법도 검토 될 수 있다.

References

[1] Chang, N.S., Hong, S.W., and Chang, J.H., Data mining in critical information technology for a successful intelligence management, Daechung Media, 1999.

[2] Chien, C.F., Chang, K.H., and Wang, W.C., An empirical study of design-of-experiment data mining for yield-loss diagnosis for semiconductor manufacturing, *Journal of intelligent manufacturing*, 2014, Vol. 25, No. 5, pp. 961-972.

[3] Jung, K.Y. and Lee, J.H., Multivariate process control procedure using a decision tree learning technique, *Journal of the Korean Data & Information Science Society*, 2015, Vol. 26, No. 3, pp. 639-652.

[4] Kim, Y.J. and Chung, Y.B., A Study on the Design of Tolerance for Process Parameter using Decision Tree and Loss Function, *Journal of Society of Korea Industrial and Systems Engineering*, 2016, Vol. 39, No. 1, pp. 123-129.

[5] Lee, H.W. and Nam, H.S., A Quality Data Mining System in TFT-LCD Industry, *Journal of The Korean Society for Quality Management*, 2006, Vol. 34, No. 1, pp. 13-19.

[6] Lee, H.W., Nam, H.S., and Kang, J.C., A Study on Data Mining Application Problem in the TFT-LCD Industry, *Journal of the Korean Data & Information Science Society*, 2005, Vol. 16, No. 4, pp. 91-101.

[7] Li, Z., Kang, L.Y., and Fan, X.H., Data integration, data mining and visualization analysis of traditional Chinese medicine manufacturing process, *China Journal of chinese materia medica*, 2014, Vol. 39, No. 15, pp. 2992.

[8] Purr, S., Meinhardt, J., Lipp, A., Werner, A., Ostermair, M., and Gluck, B., Stamping Plant 4.0-Basics for the Application of Data Mining Methods in Manufacturing Car Body Parts, *Key Engineering Materials*, 2015, Vol. 639, pp. 21-32.

[9] Ronowicz, J., Thommes, M., Kleinebudde, P., and Krysinski, J., A data mining approach to optimize pellets manufacturing process based on a decision tree algorithm, *European Journal of Pharmaceutical Sciences*, 2015,

Vol. 73, No. 1, pp. 44-48.

[10] Shin, S.H., A study on the quality control in production processing by data-mining [master's thesis], [Chungju, Korea] : University of transportation, 2013.

[11] Sim, H., Cho, D., and Kim, C.O., A data mining approach to the causal analysis of product faults in multi-stage PCB manufacturing, *International Journal of Precision Engineering and Manufacturing*, 2014, Vol. 15, No. 8, pp. 1563-1573.

[12] Sim, H.S. and Kim, C.W., Data Engineering : Fault-Causing Process and Equipment Analysis of PCB Manufacturing Lines Using Data Mining Techniques, *Korea Information Processing Society Review*, 2015, Vol. 4, No. 2, pp. 65-70.

[13] Woo, B., Lee, J.N., and Lee, J.H., A Quality Management using Data Mining Techniques for Small and Medium Manufacturing Companies, *Korea Journal of Business Administration*, 2012, Vol. 25, No. 9, pp. 3579-3599.

ORCID

Yong-Jun Kim | http://orcid.org/0000-0002-0144-4100
Geun-Sik Kim | http://orcid.org/0000-0002-9663-9620
Hyung-Geun Park | http://orcid.org/0000-0001-7955-9170