J. Soc. Korea Ind. Syst. Eng Vol. 39, No. 1 : 123-129, March 2016
http://dx.doi.org/10.11627/jkise.2016.39.1.123
ISSN : 2005-0461(print)
ISSN : 2287-7975(online)

A Study on the Design of Tolerance for Process Parameter using Decision Tree and Loss Function

Yong-Jun Kim · Young-Bae Chung†

Department of Industrial and Management Engineering, Incheon National University

의사결정나무와 손실함수를 이용한 공정파라미터 허용차 설계에 관한 연구

김용준 · 정영배†

인천대학교 산업경영공학과

In the manufacturing industry fields, thousands of quality characteristics are measured in a day because the systems of process have been automated through the development of computer and improvement of techniques. Also, the process has been monitored in database in real time. Particularly, the data in the design step of the process have contributed to the product that customers have required through getting useful information from the data and reflecting them to the design of product. In this study, first, characteristics and variables affecting to them in the data of the design step of the process were analyzed by decision tree to find out the relation between explanatory and target variables. Second, the tolerance of continuous variables influencing on the target variable primarily was shown by the application of algorithm of decision tree, C4.5. Finally, the target variable, loss, was calculated by a loss function of Taguchi and analyzed. In this paper, the general method that the value of continuous explanatory variables has been used intactly not to be transformed to the discrete value and new method that the value of continuous explanatory variables was divided into 3 categories were compared. As a result, first, the tolerance obtained from the new method was more effective in decreasing the target variable, loss, than general method. In addition, the tolerance levels for the continuous explanatory variables to be chosen of the major variables were calculated. In further research, a systematic method using decision tree of data mining needs to be developed in order to categorize continuous variables under various scenarios of loss function.

Keywords : Continuous Variable, Decision Tree, Loss Function, Tolerance

1. 서 론

현재는 정보기술의 빠른 발전과 업무의 자동화를 촉진시켜 방대한 양의 데이터를 전자적으로 수집하고 보관하는 것이 가능하게 되었으며[3], 효율적인 의사결정을 위하여 대량의 데이터를 효과적으로 분석하여 정보화하려는 노력을 하고 있다. 이렇듯 날이 갈수록 새로워지고 발전하는 정보화 시대에서 경쟁사보다 앞서나가기 위해서는 데이터베이스의 데이터를 통하여 새로운 정보나 지식을 얻는 것은 필수적인 사안이 되고 있다. 제조업 분야에서는 컴퓨터의 발달과 기술력의 증대로 공정시스템이 자동화됨에 따라 하루에도 수천개의 품질 특성치들이 계측됨은 물론, 데이터베이스화되어 실시간으로 공정의 상태를 파악하고 있다. 특히 공정이나 제품 설계 단계의 데이터를 분석하여 유용한 정보를 얻어내어 제품 설계에 반영

Received 4 February 2016; Finally Revised 16 March 2016; Accepted 17 March 2016
† Corresponding Author : ybchung@incheon.ac.kr

함으로써 소비자 요구를 충족시키는 제품을 만드는 데 기여할 수 있다. 이러한 유용한 정보를 효과적으로 도출해 내기 위해서는 통계적 기법이 대표적으로 사용될 수 있지만 기존의 통계적 기법들은 한계점들이 나타나고 있다. 예를 들면, 기존의 통계적 기법들은 대량의 데이터를 분석하는데 소요되는 시간이 많이 걸린다는 단점도 있지만, 표본의 크기가 커지면 유의한 차이도 유의하지 않게 판정하는 통계적 오류를 가지고 있다는 문제이다[1]. 이러한 한계점을 해결하기 위한 방법으로 데이터 마이닝 기법이 활용되어지고 있다[4]. 특히 의사결정나무는 적용하고자 하는 데이터의 유형에 크게 상관이 없으며, 얻어진 규칙의 해석이 용이하기 때문에 목표변수에 영향을 미치는 주요 설명변수를 쉽게 찾아낼 수 있다는 장점이 있다[5].

제조 산업에서 의사결정나무를 이용하여 데이터를 분석한 문헌은 다음과 같다. Milne et al.[7]은 제지공장에서 얻어진 데이터를 분석하였다. 무게, 습도, 함유량, 두께 등을 설명변수로 하고, 부적합품률을 목표변수로 하였다. Kim [5]은 공정 모니터링 데이터를 분석하여 목표 변수인 적합품률에 영향을 미치는 설명변수의 규격을 파악하는 연구를 하였다. Shin[10]은 제조공정에서 발생하는 품질의 문제점을 발견하고자 6개의 설명변수와 1개의 종속변수를 사용하였다. 종속변수인 품질에 영향을 주는 설명변수를 발견하여 제품의 설계 단계에서 이 요소들을 반영해야 한다고 주장하였다. 선행연구들을 살펴보면 첫째, 공정이나 제품설계 단계의 데이터를 이용하여 특성치와 설명변수들 간의 관계를 분석한 연구가 미미하다는 것이다. 둘째, 목표변수가 부적합품률, 품질 등의 정보에만 한정되었고, 품질의 변동에 따른 경제적 손실비용에 대한 정보는 연구되지 않았다.

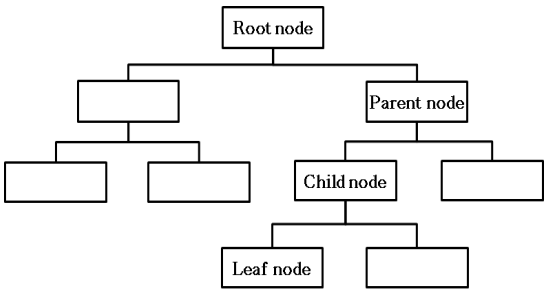
본 논문에서는 첫째, 제조 산업에서 공정이나 제품 설계 단계의 데이터를 통해 특성치와 그에 영향을 미치는 변수들의 관계를 의사결정분석을 통해 파악하였다. 둘째, 목표변수에 주로 영향을 미치는 연속형 설명변수의 구간을 도출하였다. 이를 위해 의사결정나무의 알고리즘 중 하나인 C4.5의 이론을 전처리 과정에서 적용하였다. 셋째, 손실 비용을 목표변수로 설정하여 설명 변수들과의 관계를 파악함으로써 품질의 변동에 따른 경제적 손실비용에 대한 정보를 파악하였다. 이를 위해 다구찌가 주장한 손실함수의 개념을 적용하였다.

본 연구의 나머지 부분은 다음과 같이 구성되어 있다. 제 2장은 이론적 배경으로 데이터 마이닝의 개요, 의사결정나무 분석의 개념과 알고리즘, 다구찌의 손실함수에 대해서 살펴보았다. 제 3장에서는 의사결정나무와 손실함수를 이용하여 설계 단계의 데이터를 분석하는 절차와 방법에 대하여 설명하였다. 제 4장에서는 설계 단계의 데이터 분석을 수치 예제를 통해 보여주었다. 마지막으로 제 5장에서는 의사결정분석과 손실함수를 이용한 설계 단계의 데이터 분석에 대한 결과를 요약하고, 이에 대한 시사점과 향후 연구 방향을 제시하였다.

2. 이론적 배경

2.1 의사결정나무 분석의 개요

Gartner Group은 데이터 마이닝을 통계적, 수학적 기법뿐만 아니라 패턴 인식 기법을 사용함으로써 대량의 데이터로부터 의미 있는 새로운 패턴이나 경향을 발견해내는 과정이라고 정의하였다[2]. 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 C4.5 알고리즘

C4.5는 Ross Quinlan이라는 호주 학자에 의해 수년 동안 연구되어 개발되어진 데이터 마이닝 이론 가운데 하나이다. 1986년 초기에는 ID3(iterative dichotomiser 3)라는 이름으로 몇 가지 상업적 상품을 기계 학습 분야에 적용하여 발표되었고, 이 분야에서 매우 효과적인 이론으로 소개되었다. C4.5 알고리즘은 설명변수에 대해서는 데이터의 유형과 상관없이 사용할 수 있지만 목표변수에 대해서는 연속형인 경우 범주화 시켜서 사용한다.

C4.5 알고리즘은 모든 가능한 분리 방법에서 이득비율

(gain ratio)을 계산하고, 그 이득비율이 최대가 되는 값을 최종 분리 값으로 선택하여 분리하는 방법을 사용한다. 이득비율을 계산하기 위해서 정보량(information), 엔트로피 계수(entropy index), 분리정보(split information), 이득 표준(gain criterion)에 대한 이해가 선행되어야 한다. 먼저 정보량은 전달되는 메시지의 확률 P에 좌우되는 값으로 $-\log_2(P)$로 계산되는 값이다. 예를 들어 각기 다른 4개의 메시지가 있다고 가정하였을 때, $\log_2(4)=2$가 된다. 이 때 정보량은 2가 되고 각기 다른 4개의 메시지를 확인하는데 2bits가 필요하다는 의미이다. 엔트로피 계수는 정보량을 일반화시켜 부르는 말이다. 예를 들어 $P=(p_1, p_2, \cdots, p_n)$라는 확률분포가 있다고 가정하면 이와 같은 분포에서 전달되는 정보를 엔트로피 P라고 부르게 된다. 엔트로피 P는 다음과 같이 수식 (1)로 표현할 수 있다.

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

이득 표준은 데이터의 분할을 통해 감소한 정보량의 크기를 나타내다. 분리정보는 데이터가 여러 개의 부분집합으로 분할될 때 추가적으로 발생하는 정보량을 의미한다. 이득 비율은 이득 표준과 분리정보의 비로 정의되는 값이다[6].

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 말하여 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 져야하는 현대적 개념을 나타낸다[9].

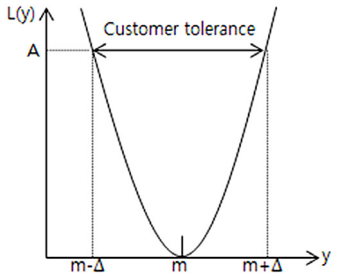
다구찌는 특성치가 연속적인 양의 값을 갖는다고 가정하였을 때 품질특성치를 망목특성(nominal-is-best characteristics), 망소특성(smaller-is-better characteristics), 망대특성(lager-is-better characteristics)의 3가지 경우로 분류하였다. 먼저 망목특성은 특정한 목표치가 주어져 있는 경우로서 특성치의 값이 주어진 목표치에 가까우면 가까울수록 좋은 경우의 특성이다. 예를 들어 길이, 두께 등이 있다. 망소특성은 특성치의 값이 음의 값을 갖지 않으며 이상적인 목표치가 0인 경우의 특성이다. 예를 들어 소음, 불순물함량 등이 있다. 마지막으로 망대특성은 특성치의 값은 음의 값을 갖지 않으며 이상적인 목표치가 무한대인 경우의 특성이다. 예를 들어 강도, 수율 등이 있다.

2.2.1 망목특성의 손실함수

측정치가 y이고 목표치가 m인 경우에 손실함수 $L(y)$를 다음 식 (2)와 같이 정의한다.

$$L(y) = k(y-m)^2 \quad (2)$$

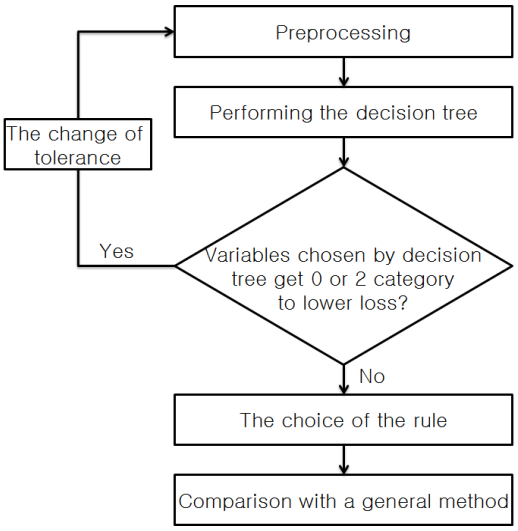
여기서 k는 상수이며 A/Δ^2으로 구할 수 있다. Δ는 목표치인 m으로부터의 허용한계를 의미하며, A는 소비자 허용한계인 $m \pm \Delta$에서 소비자의 손실을 의미한다. 이를 그림으로 표현하면 다음 그림(<Figure 2>)과 같다.



<Figure 2> Loss Function for Nominal-is-Best Characteristics

3. 연구 설계

본 연구는 의사결정나무와 다구찌 손실함수를 이용하여 제조 산업의 설계단계에서 발생하는 연속형 데이터를 분석하여 특성치가 망목 특성일 경우 손실비용을 최소화하는 설명변수의 최적 허용차를 설계하고자 하는 것이다. 이를 위한 일련의 과정을 그림으로 나타내면 다음 <Figure 3>과 같다.



<Figure 3> The Process of Analysis

3.1 전 처리

전 처리 과정에서는 목표변수인 손실비용을 계산하여 소비자 허용한계인 $m \pm \triangle$에서 소비자의 손실보다 큰 데이터의 범주와 작은 데이터의 범주로 구분하여 주고, 연속형의 값을 갖는 설명변수 모두를 3개의 구간으로 범주화하는 과정이다. 이 때 각 설명변수는 2번째 구간을 기준으로 하여 이보다 작은 값을 갖는 데이터의 구간을 '0', 큰 값을 갖는 데이터의 구간을 '2'라 표시하였다. 각 설명변수의 최적 구간이라 생각되는 2번째 구간은 '1'로 표시하였다. 이를 위한 전 처리 과정의 세부 단계는 다음과 같이 진행한다.

단계 1 : 연속형 설명변수와 망목특성을 갖는 제품의 특성치 데이터를 생성한다.

단계 2 : 목표변수인 손실비용을 계산하기 위해 특정한 목표치가 주어져 있는 경우, 즉 망목특성의 손실함수인 $L(y) = \frac{\triangle^2}{A}(y-m)^2$(단, m은 목표치, $\triangle$는 허용차, A는 소비자 허용한계인 $m \pm \triangle$에서 소비자의 손실)을 이용한다. 이 때 손실비용이 소비자 허용한계인 $m \pm \triangle$에서 소비자의 손실보다 작은 데이터의 범주를 '범주 1'로 큰 데이터의 범주를 '범주 2'로 구분하여 준다.

단계 3 : 각 설명변수 데이터를 훈련용 데이터 세트와 검증용 데이터 세트로 나눈다.

단계 4 : 단계 3에서 얻은 훈련용 데이터 세트에서 각 설명변수의 데이터를 크기 순서대로 정렬한다.

단계 5 : C4.5 알고리즘의 개념을 이용하여 이득 비율이 최대화되도록 각 설명변수를 3개의 구간으로 범주화한다. C4.5 알고리즘의 전개에 필요한 기호는 D : 데이터의 집합, C_j : 목표 변수의 j번째 범주, $|D|$: D에 속한 총 개체 수, $P(D,j)$: D에서 목표 변수의 j번째 범주에 속하는 개체의 비율이다. C4.5 알고리즘을 통해 연속형 데이터를 범주화하는 과정은 다음과 같다.

① 데이터의 집합 D에서 목표 변수의 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.

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

3.2 의사결정나무 수행

의사결정나무 수행 과정에서는 전 처리 과정에서 범주화한 각 설명변수들과 목표변수의 인과관계를 분석하여 이들 변수간의 관계를 파악하는 과정이다. 이 때 구성된 의사결정나무의 결과는 다음과 같이 두 가지 경우가 존재할 수 있다.

① 목표변수인 손실비용의 '범주 1'에 영향을 미치는 설명변수의 구간이 최적 구간이라 생각되는 '1'일 경우 바람직한 형태이므로 그 때의 설명변수는 구간 '1'을 그대로 선택한다.

② 목표변수인 손실비용의 '범주 1'에 영향을 미치는 설명변수의 구간이 '0'이거나 '2'일 경우 설명변수 구간의 조정이 필요하게 된다. 이 때 '0'이거나 '2'인 범주를 다시 전처리 과정의 단계 5를 거쳐 3개 구간으로 재 범주화한 후 의사결정분석을 수행한다. 재 범주화된 설명변수가 목표변수인 손실비용의 '범주 1'에 영향을 미치는 설명변수로 앞서와 동일하게 선택된 경우 그 때 구간을 선택한다. 만약 다른 변수가 목표변수에 영향을 미친다고 선택된다면 앞서 선택된 변수와 함께 두 가지 변수 모두 목표변수에 영향을 미친다고 판단한다.

3.3 결과의 비교

마지막으로 각 설명변수를 이산형의 값으로 범주화시키지 않고 연속형의 값 그대로 의사결정나무 분석을 수

행하는 일반적인 방법과 앞서 주장한 전처리 과정에서 각 설명변수를 3개의 구간으로 나누어 의사결정나무 분석을 수행하였을 때의 결과를 비교하였다.

4. 수치 예제

4.1 전 처리

단계 1 : 연속형 설명변수 4개 각 온도(x_1), 압력(x_2), 시간(x_3), 습도(x_4) 데이터를 일정 구간에서 200개 데이터를 랜덤하게 발생시킨다. 또한 실제 데이터를 가정하기 위해 망목특성을 갖는 제품의 길이(y) 데이터를 임의의 회귀식 24.41+(0.72×x_1)+(0.23×x_2)+(0.10×x_3)-(0.43×x_4)을 통하여 생성한다. 다음 <Table 1>은 발생된 변수의 처음 5개 데이터를 나타낸 것이다.

<Table 1> The Sample of Simulation Data

	x_1(℃)	x_2(Pa)	x_3(hr)	x_4(%)	y(cm)
1	170.9	31.3	3.2	54.7	131.5
2	143.8	42.4	1.1	56.4	113.5
3	120.0	40.3	1.2	20.2	111.5
4	176.7	38.9	2.6	54.9	137.3
5	151.3	40.7	1.4	30.1	129.9

단계 2 : 목표치(m) = 123, 허용차(△) = 9, 허용한계에서 손실비용(A) = 3,000이라 가정한 후, 각 군에서 손실비용을 구한다. 이 후 허용한계에서 손실비용(A)보다 작은 범주를 '범주 1'로, 큰 범주를 '범주 2'로 나타낸다. 다음 <Table 2>는 단계 2를 수행한 후 처음 5개 데이터의 결과를 나타낸 표이다.

<Table 2> The Categorization for Loss

	A	Category
1	2645	Category 1
2	3315	Category 2
3	4869	Category 2
4	7533	Category 2
5	1770	Category 1

단계 3 : 총 200개의 데이터 중에서 훈련용 데이터로 약 70%인 134개 데이터, 검증용 데이터로 약 30%인 66개 데이터로 나눈다.

단계 4 : 단계 3에서 얻은 훈련용 데이터 세트에서 각각의 설명변수가 갖는 데이터를 크기 순서대로 정렬한다. 다음 <Table 3>은 x_1 설명변수에 대하여 크기 순서대로 정렬하여 처음 5개 데이터를 나타내었다.

<Table 3> x_1 Data Aligned in Order

	x_1(℃)	x_2(Pa)	x_3(hr)	x_4(%)	y(cm)
1	100.8	41.5	1.2	58.0	81.7
2	101.0	34.3	3.1	46.1	85.5
3	101.7	47.1	2.6	26.7	97.3
4	101.9	47.6	2.3	36.0	93.5
5	101.9	44.4	1.4	44.3	89.1

단계 5 : C4.5 알고리즘의 개념을 이용하여 이득 비율이 최대화되도록 각 설명변수를 3개의 구간으로 범주화한다. 이를 위해 엔트로피계수(*Info$_X$(D)*), 변수 X의 정보량((*Info$_X$(D)*), 이득표준(*Gain(D, X)*), 분리정보(*Split*(D, X), 이득비율(*Gain ratio*(D, X)) 가 각 변수에서 구해진다. 변수 x_1의 경우 다음 식 (8)~식 (12)에 의해 그 값들이 계산된다.

$$Info(D) = -\sum_{j=1}^{k} P(D,j)\times\log_2(P(D,j)) = -(\frac{n_i}{N}\times\log_2(\frac{n_i}{N})+\frac{n_d}{N}\times\log_2(\frac{n_d}{n})) \quad (8)$$

$$Info_{x_1}(D) = \sum_{i=1}^{n}\frac{|D_i|}{|D|}\times Info(D_i) = \frac{F}{N}\times-(\frac{f_i}{F}\times\log_2(\frac{f_i}{F})+\frac{f_d}{F}\times\log_2(\frac{f_d}{F})) + \frac{S}{N}\times-(\frac{s_i}{S}\times\log_2(\frac{s_i}{S})+\frac{s_d}{S}\times\log_2(\frac{s_d}{S})) + \frac{T}{N}\times-(\frac{t_i}{T}\times\log_2(\frac{t_i}{T})+\frac{t_d}{T}\times\log_2(\frac{t_d}{T})) \quad (9)$$

$$Gain(D,X) = Info(D) - Info_{x_1}(D) \quad (10)$$

$$Split(D,X) = -\sum_{i=1}^{n}\frac{|D_i|}{D}\times\log_2(\frac{|D_i|}{|D|}) = -\frac{F}{N}\times\log_2(\frac{F}{N})-\frac{S}{N}\times\log_2(\frac{S}{N}) - \frac{T}{N}\times\log_2(\frac{T}{N}) \quad (11)$$

$$Gain\ ratio(D,X) = \frac{Gain(D,X)}{Split(D,X)} \quad (12)$$

(단, N = 전체 데이터 수, n_i = 전체 데이터 중 범주 1의 수, n_d = 전체 데이터 중 범주 2의 수, F = x_1변수의 구간 '0'의 데이터 수, f_i = x_1변수의 구간 '0'의 범주 1의 수, f_d = x_1변수의 구간 '0'의 범주 2의 수, S = x_1변수의 구간 '1'의 데이터 수, s_i = x_1변수의 구간 '1'의 범주 1의 수, s_d = x_1변수의 구간 '1'의 범주 2의 수, T = x_1변수의 구간 '2'의 데이터 수, t_i = x_1변수의 구간 '2'의 범주 1의 수, t_d = x_1변수의 구간 '2'의 범주 2의 수)

각 설명변수의 경계 값을 결정할 때 x_i와 x_{i+1}의 중간 값인 $(x_i+x_{i+1})/2$를 경계값으로 사용하지 않고, 데이터 세트에 실제로 존재하는 x_i를 경계값으로 사용할 수 있다[8]. 각 설명변수에서 이득 비율이 최고가 되는 3개의 구간을 구하면 다음 <Table 4>와 같다.

<Table 4> The Categorization of Explanatory Variables

	Section 1	Section 2	Section 3
x_1	x_1 < 130.902	130.902 ≤ x_1 ≤ 161.609	x_1 > 161.609
x_2	x_2 < 35.173	35.173 ≤ x_2 ≤ 41.418	x_2 > 41.418
x_3	x_3 < 1.355	1.355 ≤ x_3 ≤ 2.437	x_3 > 2.437
x_4	x_4 < 29.132	29.132 ≤ x_4 ≤ 40.596	x_4 > 40.596

4.2 의사결정나무 수행

의사결정나무 수행 과정에서는 전 처리 과정에서 범주화한 각 설명변수와 목표변수의 인과관계를 분석하여 이들 변수간의 관계를 파악하는 과정이다. 의사결정나무 분석을 수행한 결과는 다음 <Table 5>와 같다.

<Table 5> The Result of Decision Tree(new)

x_1 in (0, 2) : category 2(137/12) x_1 in (1) : category 1(63/12) Errors : 12.0%		
	forecast	
Category 1	51	12(error)
Category 2	12(error)	125

<Table 5>를 통해 x_1 변수가 목표 변수에 영향을 주는 주요 변수라는 사실을 알 수 있다. 또한, x_1이 '1'의 구간 즉, 130.902 ≤ x_1 ≤ 161.609를 만족할 때 목표변수의 '범주 1'을 만족시킴을 알 수 있다.

4.3 결과의 비교

마지막으로 각 설명변수를 이산형의 값으로 범주화시키지 않고 연속형의 값 그대로 의사결정나무 분석을 수행하는 일반적 방법과 각 설명변수를 3개의 구간으로 나누어 의사결정나무 분석을 수행하였을 때의 그 결과를 비교하였다. 각 설명변수를 연속형의 값 그대로 의사결정 분석을 수행한 결과는 다음 <Table 6>과 같다.

<Table 6> The Result of Decision Tree(Old)

x_1 ≤ 126.9766 : category 2(58) x_1 > 126.9766 : x_1 ≤ 167.61 : category 1(89/27) x_1 > 167.61 : category 2(53/1) Errors : 14.0%		
	forecast	
Category 1	62	1(error)
Category 2	27(error)	110

<Table 6>을 통해 x_1 변수가 목표변수에 영향을 주는 주요 변수라는 사실을 알 수 있다. 또한, x_1이 x_1 ≤ 167.61을 만족할 때 목표변수의 '범주 1'을 만족시킴을 알 수 있다. 두 가지 방법에서 '범주 1'로 분류한 x_1 변수의 구간에서 손실비용을 비교하기 위하여 t-검정을 실시한 결과는 다음 <Table 7>과 같다.

<Table 7> Comparison of Two Methods

Group	N	Mean	Std Dev	Std Error
old	60	2651.35	3442.95	444.48
new	45	1726.16	2226.35	331.89
Variances	Equal variance test		t-test	
	F	P-value	t	P-value (both sides)
Equal	4.414	.038	1.572	.119
Unequal			1.668	.098

t-검정은 두 집단간의 분산이 동일한 경우와 동일하지 않은 경우 두 가지로 나누어진다. <Table 7>의 결과 두 집단의 분산이 같다는 가설 H_0가 기각이 되므로 두 집단간의 분산은 동일하다고 볼 수 없다. 위의 결과에서 두 가지 방법의 손실비용 평균이 유의수준 5%에서 통계적 차이가 없지만 10%를 기준으로 삼는다면 p = 0.098로 통계적으로 유의한 차이가 있다. 따라서 기존의 의사결정분석보다 새롭게 제시된 의사결정분석에 의한 설명변수 허용차 구간이 목표 변수인 손실비용을 더 적게 한다는 것을 알 수 있다. 또한 새롭게 제시된 방법을 통해 목표 변수에 주로 영향을 미치는 변수가 아닌 연속형 설명변수에 대해서도 허용차를 구할 수 있다는 장점이 있다.

5. 결 론

제조업 분야에서는 컴퓨터의 발달과 기술력의 증대로 공정시스템이 자동화됨에 따라 하루에도 수천 개의 품질 특성치들이 계측됨은 물론, 데이터베이스화되어 실시간

으로 공정의 상태를 파악하고 있다. 특히 공정이나 제품 설계 단계의 데이터를 분석하여 유용한 정보를 얻어내어 제품 설계에 반영함으로써 소비자 요구를 충족시키는 제품을 만드는 데 기여할 수 있다. 본 논문에서는 첫째, 제조 산업에서 설계 단계의 데이터를 통해 특성치와 그에 영향을 미치는 변수들의 관계를 의사결정분석을 통해 파악하였다. 둘째, 목표 변수에 주로 영향을 주는 연속형 설명변수의 최적 구간을 도출하였다. 이를 위해 전처리 과정에서 의사결정나무의 알고리즘 중 하나인 C4.5의 이론을 적용하였다. 셋째, 손실 비용을 목표변수로 설정하여 설명 변수들과의 관계를 파악함으로써 품질의 변동에 따른 경제적 손실비용에 대한 정보를 파악하였다. 이를 위해 다구찌 손실함수의 개념을 적용하였다. 본 연구는 목표변수에 영향을 미치는 연속형 설명변수의 최적 범위를 찾는데 기여할 수 있음을 파악하였다.

본 논문에서 설명변수들을 이산형의 값으로 범주화시키지 않고 연속형의 값 그대로 의사결정나무 분석을 수행하는 기존의 방법과 각 설명변수를 3개의 구간으로 나누어 의사결정나무 분석을 수행하는 새로운 방법을 비교하였다. 비교 결과 기존의 의사결정분석 방법보다 새롭게 제시된 의사결정분석에 의한 설명변수의 구간이 손실비용을 더 작게 한다는 것을 알 수 있었다. 또한 새롭게 제시된 방법을 통해 목표변수에 주로 영향을 미치는 변수에 대하여 최적 구간을 구할 수 있었다. 추후 연구에서는 연속형 설명 변수를 보다 체계적으로 범주화할 수 있는 방법론이 있는지 연구해볼 가치가 있다. 또한 망대나 망소특성을 가지고 있는 특성치의 손실비용을 계산하여 의사결정분석 결과를 검토하는 것이 필요하다.

References

[1] Adriaans, P. and Zantinge, D., Data Mining, Addison Wesley Longman, 1996.

[2] Berry, M.J. and Linoff, G., Data Mining Techniques for Marketing Sales and Customer Support, John Wiley and Sons Inc., 1997.

[3] Chung, Y.S. and Kang, C.W., Application of Data Mining for Improving and Predicting Defect Rate in Manufacturing Process, *Proceedings of the Society of Korea Industrial and Systems Engineering*, 2004, Fall Conference, pp. 121-123.

[4] Chung, Y.S. and Kang, C.W., Improvement and Detection of Process Variables Using Data Mining, *Proceedings of the Society of Korea Industrial and Systems Engineering*, 2005, Spring Conference, pp. 272-278.

[5] Kim, S.J., Analysis of process monitoring data using decision trees [master's thesis], [Daejeon, Korea] : KAIST, 2001.

[6] Lee, G.S., Comparison of Discretization Algorithms for Efficient Data Mining Modeling [master's thesis], [Cheongju, Korea] : Chungbuk National University, 2004.

[7] Milne, R., Drummond, M., and Renoux, P., Predicting paper making defects online using data mining, *Knowledge-Based Systems*, 1998, Vol. 11, pp. 331-338.

[8] Quinlan, J.R., C4.5 : Programs for machine learning, Morgan Kaufmann Publishers, 1993.

[9] Ree, S.B., Analysis of Quality Loss Function(QLF) of Taguchi, *The Korean Society for Quality Management*, 1997, Vol. 25, No. 3 pp. 119-130.

[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.

ORCID

Yong-Jun Kim | http://orcid.org/0000-0002-0144-4100

Young-Bae Chung | http://orcid.org/0000-0003-4259-6677