韓國IT서비스學會誌
第10卷 第3號
2011年 9月

철강 연주공정에서 데이터마이닝을 이용한 품질제어 방법에 관한 연구*

김재경** · 권택성** · 최일영** · 김혜경*** · 김민용**

A Study on Quality Control Using Data Mining in Steel Continuous Casting Process*

Jae Kyeong Kim** · Taeck Sung Kwon** · Il Young Choi** ·
Hyea Kyeong Kim*** · Min-Yong Kim**

■ Abstract ■

The smelting and the continuous casting of steel are important processes that determine the quality of steel products. Especially most of quality defects occur during solidification of the steel continuous casting process. Although quality control techniques such as six sigma, SQC, and TQM can be applied to the continuous casting process for improving quality of steel products, these techniques don't provide real-time analysis to identify the causes of defect occurrence.

To solve problems, we have developed a detection model using decision tree which identified abnormal transactions to have a coarse grain structure. And we have compared the proposed model with models using neural network and logistic regression. Experiments on steel data showed that the performance of the proposed model was higher than those of neural network model and logistic regression model. Thus, we expect that the suggested model will be helpful to control the quality of steel products in real-time in the continuous casting process.

Keyword : Data Mining, Decision Tree, Neural Network, Logistic Regression, Continuous Casting

논문투고일 : 2011년 06월 13일 논문수정완료일 : 2011년 08월 25일 논문게재확정일 : 2011년 09월 16일
* 본 연구는 2010년도 경희대학교 협동연구지원에 의한 결과임(KHU-20100847).
** 경희대학교 경영대학 & 경영연구원
*** 경희대학교 경영대학 & 경영연구원, 교신저자

1. 서 론

최근 경영환경의 급격한 변화로 품질이 중요한 경영전략으로 작용함에 따라 직접적으로 품질에 영향을 미치는 연구와 이를 고려한 품질 예측이 요구되고 있다. 이와 같은 경영환경에서 경쟁우위를 얻기 위하여 많은 제조 기업들은 생산 공정에서 생성되는 대량의 데이터들을 이용한 통계적 품질관리(statistical quality control), 식스 시그마(six sigma), 전사적 품질관리(total quality management)등의 기법을 통해 생산현장의 품질관리를 하고 있다[1].

철강기업 또한 제선공정, 제강공정, 연주공정, 압연공정에서 생성되는 측정 데이터를 수집하여 식스 시그마 기법등에 활용하고 있다. 그러나 이러한 품질관리 기법은 생산 공정에서 불량을 일으키는 요인파악과 같은 데이터들의 심도 있는 분석을 제공하지 않기 때문에 유용한 지식을 도출하기에는 한계가 있다. 특히, 철강제품의 품질결함 대부분이 연주공정에서 발생하지만 기존의 품질관리 기법으로는 연주공정을 실시간으로 분석할 수 없을 뿐만 아니라 제품의 품질결함 유·무를 연주공정이 끝난 후에도 확인하지 못하고 압연공정이 끝난 후에 단지 추정 가능하다는 문제점이 있다.

본 연구에서는 철강공정 중 연주공정에서 발생하는 품질 결함을 파악하고 실시간으로 연주공정을 관리하기 위한 방법으로 철강 생산공정에서 발생하는 대용량의 데이터로부터 데이터마이닝 기법중 하나인 의사결정나무(decision tree)를 이용한 연주공정상의 이상트랜잭션 검출모형(abnormal transaction detection model)을 제시하였다. 또한 제시한 모형의 유용성을 검증하기 위하여 신경망(neural network)과 로지스틱 회귀분석(logistic regression)으로 비교 분석하였다.

본 연구 결과를 통해 연주공정에서의 실시간 품질관리 및 사전 품질관리가 유용할 것으로 기대되며, 더 나아가서는 공정 중 이상 원인을 해결하는 과정에서 소요되는 시간과 비용의 감소를 통해 공정의 생산성을 향상시킬 수 있을 것으로 기대된다.

본 논문의 구성은 다음과 같다. 제 2장에서는 의사결정나무 분석, 신경망 분석, 로지스틱 회귀분석의 특징에 대해서 설명하였으며, 본 연구의 분석 주제로 철강 생산공정 중의 하나인 연주공정에 대해서 살펴보았다. 제 3장에서는 본 연구의 분석절차에 대해서 설명하였으며, 제 4장에서는 본 연구에서 제시한 방법론을 실험적으로 분석하였다. 마지막으로 제 5장에서는 본 연구의 결론과 추후연구에 대하여 논의하였다.

2. 이론적 배경

2.1 의사결정나무

의사결정나무(decision tree)는 의사결정규칙을 나무구조로 도표화하여 분류와 예측을 하는데 효과적으로 사용되며, 예측 정확도가 다른 분류모델보다 높거나 동등하여 많이 사용하는 분석방법이다[6, 10]. 의사결정나무의 장점은 분류 또는 예측의 과정이 나무구조에 의한 추론규칙에 의해서 표현되기 때문에 모델 구축과정이 단순하며 결과에 대한 해석이 용이하다는 것이다. 하지만 모형을 구축하는데 사용되는 표본의 크기에 민감하여, 정확한 모형을 만들기 위해서는 서로 상이한 값을 갖는 레코드들을 많이 포함하는 데이터가 필요하다는 단점이 있다.

지금까지 의사결정나무 기법을 이용한 다양한 연구가 진행되어 왔다. 임세헌 외[8]는 온라인 자동차 보험을 계약한 고객의 데이터로 의사 결정나무를 이용해 고객이탈을 예측하였으며, 장남식[9]은 카드사, 이동 통신사, 보험사의 고객 인적 데이터 및 거래 데이터로 의사결정나무 기법을 이용해 해지 모형을 구축하였다. 또한 변성규 외[3]는 제조공정에서 발생하는 대용량 공정데이터를 이용하여 불량항목별 예측을 위한 분석절차와 모형을 제시하였다. 이처럼 기존의 의사결정나무 기법의 연구는 금융, 서비스업, 제조업 등 다양한 연구가 이루어

졌지만, 철강분야에서의 연구는 미흡한 상태이다.

따라서 본 연구에서는 연주공정상의 이상트랜잭션 검출모형을 생성하기 위해서 C5.0 알고리즘을 이용하였다. 특히, C5.0 알고리즘은 CHAID, CART, C4.5 알고리즘에 비해 가장 정확한 분류를 만들어 주는 알고리즘으로 알려져 있으며 최근 의사결정나무 분석에서 많이 이용되고 있다[15].

2.2 신경망

신경망(neural network)은 인간 두뇌의 기본단위인 뉴런의 생리학적 모델을 모방한 개념으로 외부로부터 입력을 받아들이는 노드(node)와 외부로 출력을 담당하는 노드가 있고 이들 사이에 은닉 노드가 존재하면서 과거에 수집된 데이터로부터 반복적인 학습과정을 거쳐 데이터에 내재되어 있는 패턴을 찾아내는 모델링 기법이다[5, 11]. 신경망은 입력층, 은닉층, 출력층으로 구성되어 있다. 입력층은 각 입력변수에 대응되는 노드로 구성되어 있으며 노드의 수는 입력변수의 개수와 같다. 은닉층은 입력층으로부터 전달되는 변수 값들의 선형결합을 비선형함수로 처리하여 출력층 또는 다른 은닉층에 전달하는 역할을 한다. 마지막으로 출력층은 반응변수에 대응되는 노드로 예측값이 생성되는 곳이다.

신경망의 장점은 입력과 출력마디에 이산형, 연속형 변수들을 모두 사용할 수 있고 입력변수들의 비선형 조합을 통해 제공하기 때문에 예측력이 우수하며 제품 선택의 폭이 넓고 취득하기 쉽다는 것이다. 하지만 결과에 대한 이유를 설명하지 못하고 변수 변환의 추가적이 노력이 필요하며 복잡한 학습과정이 필요하므로 모형을 구축하는 데 시간이 많이 걸리는 단점이 있다.

2.3 로지스틱 회귀분석

로지스틱 회귀분석(logistic regression)은 기존의 선형 회귀분석의 종속변수를 범주형으로 확장한 것이다[4]. 본 연구에서와 같이 종속변수가 일반적인 연속변수 혹은 측정치(나이, 체중, 온도 등)가 아닌 이항변수(binary variable, 남/여, 오류여부(Y/N) 등)인 경우 즉, 독립변수에 대한 그 분포를 수식하기 위해 일반적인 선형회귀나 다항회귀분석을 사용하기 어려운 경우에 사용된다.

로지스틱 회귀분석 과정은 확률의 분포를 다루는 것이며 종속변수가 되는 사건이 발생하게 되는 확률을 추정하는 것이다. 로지스틱 회귀분석은 2단계의 과정을 거친다. 첫 번째 단계는 각 집단에 속하는 확률의 추정치를 계산하는 것이다. 이진변수의 경우 집단 1에 속하는 확률로서 $P(Y = 1)$의 추정치를 얻는다(집단 0에 속하는 확률의 경우도 동일함).

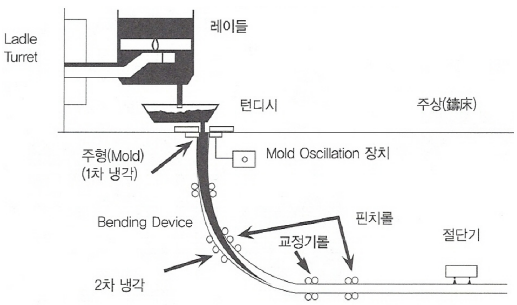
두 번째 단계는 각 관찰치를 어느 한 집단으로 분류하기 위해서 이러한 확률들에 분류 기준값(cut-off value)을 적용한다. 예를 들면, 이진변수의 분류 기준값이 0.7인 경우 $P(Y = 1) > 0.7$의 추정된 확률을 가지면 집단 1로 분류되고, 반면에 $P(Y = 1) \leq 0.7$의 추정된 확률을 가지면 집단 0으로 분류된다. 하지만 이러한 분류 기준값이 반드시 0.7일 필요는 없다. 분석대상사건의 발생확률이 낮을 경우에는 분류 기준값을 0.7이하로 정의할 수도 있지만, 0.7보다 높은 분류 기준값을 설정함으로써 해당 데이터를 충분히 집단 1에 속하는 것으로 분류할 수 있다.

2.4 연주공정

연속주조법은 직접적으로 용강을 주형에 주입함으로 위에서 주입되는 용강이 주형을 통하여 아래로 흐름에 따라 주형의 형상으로 냉각, 응고되면서 연속적으로 슬래브, 블룸, 빌릿 등을 제조하는 방법이다[2]. 연주공정 중 주형(mold)에서의 냉각을 1차 냉각, 주형(mold) 이후의 물과 공기에 의한 냉각을 2차 냉각이라 하며 연속주조에 대한 전체 공정은 [그림 1]과 같다.

연주공정은 설비비 삭감, 에너지 절감, 제품 실수율 향상, 제품생산 원가절감, 납기단축, 자동화,

기계화가 용이해서 작업환경을 개선시킬 수 있는 장점이 있는 반면에 연주공정 중 2차 냉각공정에 의해 대부분의 제품 균열 및 품질 결함이 발생하는 단점이 있다.



[그림 1] 연속주조공정

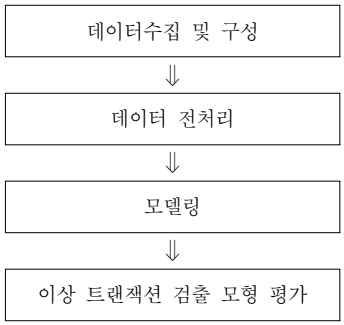
이와 같은 한계를 극복하기 위해서 연주공정 중 2차 냉각공정의 관리에 대한 다양한 연구가 진행되어 왔다. Cheung et al.[12], Hancock et al.[13], Ma et al.[14]은 냉각공정에서 발생하는 품질 결함을 줄이기 위해서 2차원 열 전달 모형을 개발하였으며, 이상민 외[7]는 최적의 2차 냉각패턴 설정을 위해 주편표면온도 측정에 의한 2차 냉각 존(zone) 내 주편표면온도 모델을 개발하였다. 그러나 기존의 연주공정에 대한 연구는 열 전달 모형 및 시간당 수익성 탐색에 대해서는 다양하게 이루어졌으나, 연주공정의 품질을 실시간으로 관리하기 위한 이상 트랜잭션 검출에 대한 연구는 미흡한 상태이다.

3. 연구방법 및 절차

3.1 분석 프로세스

본 연구에서는 국내 H철강 기업의 생산공정 데이터를 이용하여 연주공정상에서 발생하는 품질 결함을 파악하고 생산공정을 관리하기 위해 의사결정나무 분석을 이용하여 이상트랜잭션 검출모형을 제안하고 신경망 분석 및 로지스틱 회귀분석과 비교분석하여 제안한 방법의 유용성을 검증하고자 한다.

본 연구의 분석 프로세스는 데이터수집 및 구성, 데이터 전처리, 모델링, 이상 트랜잭션 검출 모형 평가의 4단계로 구분 되며 [그림 2]와 같다.



[그림 2] 연구의 흐름도

데이터수집 및 구성단계에서는 국내 H철강 기업의 생산공정 데이터로부터 연주공정상의 이상 트랜잭션 검출 규칙을 생성하기 위해 필요한 데이터를 추출하였다. 데이터 전처리 단계에서는 데이터의 이해를 바탕으로 본격적인 분석에 앞서 데이터 전처리, 결측치 처리 등을 실행하였다. 모델링 단계에서는 클레멘타인 12.0으로 분석에 적합한 노드를 구성하여 모델을 수립하였다. 이상 트랜잭션 검출 모형 평가 단계에서는 의사결정나무를 이용한 이상트랜잭션 검출 모형과 신경망 및 로지스틱 회귀분석을 이용한 모형을 비교 평가하였다.

3.2 데이터수집 및 구성

데이터수집 및 구성은 국내 H철강 기업의 생산공정 데이터로부터 연주공정상의 이상 트랜잭션 검 출 규칙을 생성하기 위해 필요한 데이터를 추출하는 단계이다.

본 연구에서는 생산공정에서 수집된 대용량의 데이터 중에서 <표 1>과 같이 연주공정의 2차 냉각 존(zone)에서 생성되는 32개 변수의 데이터와 <표 2>와 같이 압연공정에서 생성되는 품질결함을 나타내는 13개 변수의 데이터를 대상으로 분석하였다.

〈표 1〉 연주공정 2차 냉각 존(zone) 변수

변수	변수설명	변수	변수설명
SCO_1BF_FLW	1B Fixed Spray Water 유량	SCO_1BF_MAF	1B Fixed Spray Air 유량
SCO_1BL_FLW	1B Loosed Spray Water 유량	SCO_1BL_MAF	1B Loosed Spray Air 유량
SCO_1CF_FLW	1C Fixed Spray Water 유량	SCO_1CF_MAF	1C Fixed Spray Air 유량
SCO_1CL_FLW	1C Loosed Spray Water 유량	SCO_1CL_MAF	1C Loosed Spray Air 유량
SCO_Z2F_FLW	Z2 Fixed Spray Water 유량	SCO_Z2F_MAF	Z2 Fixed Spray Air 유량
SCO_Z2L_FLW	Z2 Loosed Spray Water 유량	SCO_Z2L_MAF	Z2 Loosed Spray Air 유량
SCO_Z3F_FLW	Z3 Fixed Spray Water 유량	SCO_Z3F_MAF	Z3 Fixed Spray Air 유량
SCO_Z3L_FLW	Z3 Loosed Spray Water 유량	SCO_Z3L_MAF	Z3 Loosed Spray Air 유량
SCO_Z4F_FLW	Z4 Fixed Spray Water 유량	SCO_Z4F_MAF	Z4 Fixed Spray Air 유량
SCO_Z4L_FLW	Z4 Loosed Spray Water 유량	SCO_Z4L_MAF	Z4 Loosed Spray Air 유량
SCO_Z5F_FLW	Z5 Fixed Spray Water 유량	SCO_Z5F_MAF	Z5 Fixed Spray Air 유량
SCO_Z5L_FLW	Z5 Loosed Spray Water 유량	SCO_Z5L_MAF	Z5 Loosed Spray Air 유량
SCO_Z6F_FLW	Z6 Fixed Spray Water 유량	SCO_Z6F_MAF	Z6 Fixed Spray Air 유량
SCO_Z6L_FLW	Z6 Loosed Spray Water 유량	SCO_Z6L_MAF	Z6 Loosed Spray Air 유량
SCO_Z7F_FLW	Z7 Fixed Spray Water 유량	SCO_Z7F_MAF	Z7 Fixed Spray Air 유량
SCO_Z7L_FLW	Z7 Loosed Spray Water 유량	SCO_Z7L_MAF	Z7 Loosed Spray Air 유량

〈표 2〉 압연공정 품질결함 변수

변수	변수설명
Scab	Scab(표면결함-제강성결함)
기포흠	기포흠(표면결함-제강성결함)
선형흠	선형흠(표면결함-제강성결함)
연와흠	연와흠(표면결함-제강성결함)
이중판	이중판(표면결함-제강성결함)
C-L-Crack	C-L-Crack(표면결함-제강성결함)
Hole	Hole(표면결함-제강성결함)
유사비늘흠	유사비늘흠(표면결함-제강성결함)
E-Crack	E-Crack(표면결함-제강성결함)
성분	성분(기타 결함)
E- L-Crack	E- L-Crack(표면결함-제강성결함)
연주성스크래치	연주성 스크래치 (표면결함-제강성결함)

입력 변수는 연주공정의 2차 냉각 존(zone)에서 생성되는 32개 변수의 데이터를 선정하였고, 목표 변수는 압연공정에서 생성되는 품질 결함을 나타내는 변수 중에 표면등급이 2이상인 데이터를 선정하였으며 <표 3>과 같다.

〈표 3〉 분석변수

구 분	변수명		비고
입력 변수	SCO_1BF_FLW	SCO_1BF_MAF	2차 냉각공정의 Water Flow, Air Flow 변수
	SCO_1BL_FLW	SCO_1BL_MAF	
	SCO_1CF_FLW	SCO_1CF_MAF	
	SCO_1CL_FLW	SCO_1CL_MAF	
	SCO_Z2F_FLW	SCO_Z2F_MAF	
	SCO_Z2L_FLW	SCO_Z2L_MAF	
	SCO_Z3F_FLW	SCO_Z3F_MAF	
	SCO_Z3L_FLW	SCO_Z3L_MAF	
	SCO_Z4F_FLW	SCO_Z4F_MAF	
	SCO_Z4L_FLW	SCO_Z4L_MAF	
	SCO_Z5F_FLW	SCO_Z5F_MAF	
	SCO_Z5L_FLW	SCO_Z5L_MAF	
	SCO_Z6F_FLW	SCO_Z6F_MAF	
	SCO_Z6L_FLW	SCO_Z6L_MAF	
	SCO_Z7F_FLW	SCO_Z7F_MAF	
	SCO_Z7L_FLW	SCO_Z7L_MAF	
출력 변수	surface_yn		압연공정의 이상 Transaction 중 표면등급 ≥ 2

또한 본 연구를 수행하기 위해 데이터마이닝 도구인 클레멘타인 12.0을 사용하였으며, 총 데이터 수는 10,999건으로 훈련 집합(training data set)과 테스트 집합(test data set)을 각각 7 : 3의 비율로 나누어서 수행하였다.

3.3 데이터 전처리

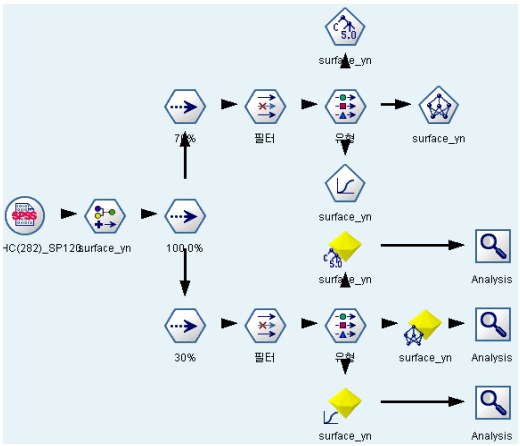
데이터 전처리는 데이터마이닝 분석 목적에 따라 데이터를 처리하는 일련의 과정을 말한다. 실제 데이터는 불완전(incomplete)하며, 잡음(noisy)이 있고 불일치(inconsistent)하기 때문에 데이터의 전처리가 필요하다. 일반적으로 데이터 전처리 단계에서는 데이터 중 결측치를 채워넣고, 잡음이 있는 데이터를 제거하며 이상치를 식별하고, 데이터 불일치를 교정한다[4].

본 연구에서는 실험 데이터 수가 많으므로 더 정확한 성과를 위해 결측치가 발생하는 레코드를 삭제하여 데이터를 전처리 하였다. 또한 연주공정에서 2초 간격으로 수집되는 전체 트랜잭션 중 같은 시간에 2차 냉각공정 중 다른 슬라브와 겹치는 트랜잭션은 제외하고 동일한 슬라브의 시간대 트랜잭션만 사용하였으며, 연주공정에서 품질결함을 판별하기 위한 데이터 셋을 만들기 위하여 연주공정의 데이터와 압연공정의 품질결함 데이터를 통합하였다.

3.4 모델링

본 논문에서는 모델링을 하기 위해 데이터마이닝 도구인 클레멘타인 12.0을 사용하였으며, 본 연구의 데이터 마이닝 수행노드는 [그림 3]과 같다.

본 연구에 사용된 모델링 및 노드에 대한 설명은 다음과 같다. 입력노드는 분석대상 파일을 불러오는 노드로 본 연구에서는 철강공정 중 대표 강종인 J강종을 대상으로 분석하였다. 재분류 노드는 “필드 유형”이 “이산형, 범주형, 이분형”인 경우 여러가지 범주(이산)값을 하나의 범주(이산)값으로 바꿀 때나 기존 값을 다른 값으로 바꿀 때 사용할 수 있는 노드로, 본 연구에서는 표면등급 1, 2, 3중 1을 nor로 2, 3을 out으로 지정하였다. 여기서 nor은 표면등급이 정상을 의미하며, out은 표면등급이 불량을 의미한다.



[그림 3] 데이터마이닝 수행노드

표본 노드는 표본을 추출하는 노드로 대용량 데이터에서 효과적인 작업 수행을 위하여 사용되는 노드이다. 일반적으로 전반적인 추세를 살펴보거나 예비 모델링을 할 경우 장시간에 과도한 작업시간을 피하기 위해 표본을 추출한다. 본 연구에서는 표본 노드를 통해 훈련 집합(training data set)과 테스트 집합(test data set)의 비율을 7 : 3으로 나누어서 실험하였다. 필터 노드는 데이터 필드명을 변경하거나 필드를 이후 스트림에서 사용할 것인지를 지정하는 노드로 클레멘타인에서 자주 사용되며, 유형 노드와 마찬가지로 소스 노드에 필터 노드의 기능이 포함되어 있어서 필터 노드를 추가하지 않고도 데이터 처리를 할 수도 있다. 본 연구에서는 전체 데이터 필드 중에서 입력필드와 출력필드만 필터하였다. 유형노드는 데이터 유형을 지정하거나 변경할 수 있게 해주는 노드로 소스 노드에서도 이 기능을 지원하지만, 대용량 데이터의 유형을 지정하거나 변경할 때 모델링 노드 앞에서 이 노드를 사용하는 것이 더 유용하다. 본 연구에서는 필드의 타입을 입력필드는

레인지(range)로 출력필드는 셋(set)으로 설정하였으며, 입력필드의 방향은 인(in)으로 출력필드의 방향은 아웃(out)으로 설정하였다. C5.0 노드는 의사결정나무 또는 규칙 집합을 생성하기 위하여 C5.0 알고리즘을 이용하는 노드로 정보 획득값이 최대값을 가지는 입력필드를 분류필드로하여 나무 모델을 만든다. Net 노드는 신경망 학습을 통해 생성된 속성을 내재하고 있는 모델노드로, 새로운 데이터가 이 노드를 통과하게 되면 내재된 정보에 따라 예측값을 생성하게 된다. 로지스틱 노드는 2개 이상의 이산형 값을 갖는 목표 필드와 설명 필드들 간의 인과 관계를 로지스틱 함수를 이용하여 추정하는 모델링 노드이다. 마지막으로 분석노드는 모델이 가진 예측력을 평가하는데 사용되며, 모델 노드에서 하나 또는 그 이상 생성된 서로 다른 모델들 간의 예측 값이 실제 목표 값을 얼마나 정확하게 예측하는지를 비교하는데 사용된다. 본 연구에서는 분류행렬표 및 결과값을 통해 모델이 가진 예측력을 평가하는데 사용하였다.

3.5 이상 트랜잭션 모형 평가

본 연구에서는 제시한 모형의 적절성을 확인하기 위해 분류행렬표를 사용하였다. 분류행렬표란, 예측 모형이 특정 데이터 집합에 대해 수행한 정분류와 오분류의 요약정보를 보여준다. 정오분류의 행과 열은 각각 실제 집단과 예측 집단에 대응되며 본 논문에서는 C_0을 정상 트랜잭션으로, C_1을 이상 트랜잭션으로 표시하였다.

분류행렬표의 대표적인 측정치는 다음과 같다[4].

(1) 민감도(sensitivity)는 주요 집단의 소속 레코드를 정확하게 판별하는 능력을 말한다. 이 지표는 실제 C_0 집단을 C_0 집단으로 정확하게 분류할 확률로써 $n_{0,0}/(n_{0,0}+n_{0,1})$로 측정된다.

(2) 특이도(specificity)는 C_1 집단의 소속 레코드를 정확하게 판별하는 능력을 말한다. 이 지표는 실제 C_1 집단을 C_1 집단으로 정확하게 분류할 확률로써 $n_{1,1}/(n_{1,0}+n_{1,1})$로 측정된다.

(3) 위양성률(false positive rate)은 C_0 집단으로 분류된 레코드 중에서 실제 C_1 집단을 C_0 집단으로 잘못 분류한 레코드의 비율을 의미하며, $n_{1,0}/(n_{0,0}+n_{1,0})$로 측정된다.

(4) 위음성률(false negative rate)은 C_1 집단으로 분류된 레코드 중에서 실제 C_0 집단을 C_1 집단으로 잘못 분류한 레코드의 비율을 의미하며, $n_{0,1}/(n_{0,1}+n_{1,1})$로 측정된다.

(5) 정확도(accuracy)는 전체집단에서 각각의 집단을 정확하게 분류하는 정도를 나타내는 것으로 $(n_{0,0}+n_{1,1})/n$으로 측정된다. 여기서 n은$(n_{0,0}+n_{0,1}+n_{1,0}+n_{1,1})$을 의미한다.

이 지표들 간의 균형을 맞추는 분류기준값을 찾기 위해 1차원 테이블을 사용하여 이 지표들과 분류기준값을 도표로 작성하는 것이 도움이 된다. 아래 <표 4>는 분류행렬표를 나타낸 것으로 본 연구에서는 정확도, 민감도, 특이도를 계산하여 예측모형의 유용성을 검증하였다.

〈표 4〉 분류행렬표

		예측집합	
		C_0(정상)	C_1(이상)
실제 집합	C_0 (정상)	$n_{0,0}$ = 정확하게 분류된 C_0의 개수	$n_{0,1}$ = C_1으로 잘못 분류된 C0의 개수
	C_1 (이상)	$n_{1,0}$ = C_0으로 잘못 분류된 C_1의 개수	$n_{1,1}$ = 정확하게 분류된 C_1의 개수

4. 실험 및 평가

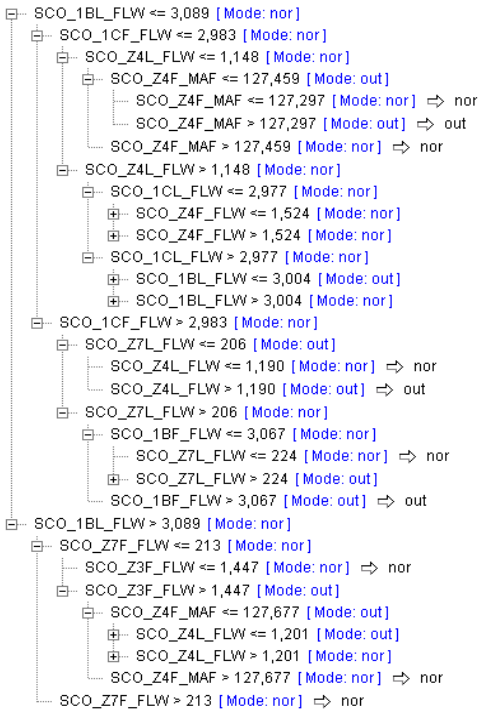
4.1 이상 트랜잭션 분석결과

본 연구에서는 연주공정에서 생성되는 트랜잭션을 통해 사전적으로 품질결함 유·무를 판정하기 위해 이상 트랜잭션 검출 모형을 생성하였다. 이상 트랜잭션을 도출하기 위해 연주공정의 2차 냉각 존(zone) 변수를 입력변수로 지정하고, 압연공

정의 품질결함 변수를 출력변수로 지정하였다.

J강종에 대한 연주 공정의 품질 불량의 패턴을 찾기 위하여 의사결정나무 분석을 한 결과는 [그림 4]와 같으며, 정상 트랜잭션은 nor로 이상 트랜잭션은 out로 표시된다.

예를 들면 J강종은 첫 번째 기준인 SCO_1BL_FLW가 3089이하이고, SCO_1CF_FLW가 2983이하이고, SCO_Z4L_FLW가 1148이하이고, SCO_Z4F_MAF가 127,459이하일 때 이상트랜잭션이 발생함을 알 수 있다. 이와 같은 결과로 J강종의 이상트랜잭션 발생에 영향을 미치는 변수가 SCO_1BL_FLW ,SCO_1CF_FLW, SCO_Z7F_FLW, SCO_Z4F_MAF, SCO_Z4L_FLW 라는 것을 알 수 있다.

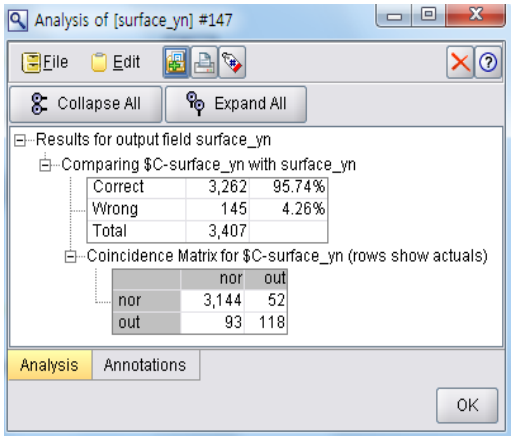


[그림 4] 연주공정 이상 트랜잭션 검출 규칙

J강종의 품질과 생산성 향상을 위해서는 연주공정 이상 트랜잭션 검출규칙을 통해 도출된 민감변수 중점 관리함으로써 품질과 생산성을 높일 수 있을 것이다.

4.2 의사결정나무 분석 결과

본 연구를 분석하기 위해 데이터마이닝 도구인 클레멘타인 12.0을 사용하였으며, 의사결정나무 분석을 실행한 출력 결과는 [그림 5]와 같다. 이전에 모델링에서 설명하였듯이 클레멘타인의 분석 노드는 생성된 모형의 예측값이 실제값에 얼마나 잘 맞는지를 보여주는 노드로 서로 다른 모형의 예측값을 비교하는 데도 사용되며, 각 모형의 예측값이 얼마나 일치하는지도 보여주는 '일치성'을 확인하는데도 사용된다.



[그림 5] 의사결정나무 분석 결과

[그림 5]의 결과를 해석하면 다음과 같다. 첫 번째 지표는 surface_yn의 실제값과 C5.0 rule 모형의 예측값인 $C-surface_yn를 비교한 것이다. C5.0 모형은 예측값이 95.74%의 예측율을 보이고 있으며, 세 가지 기법 중에 가장 높은 예측율을 나타내었다.

두 번째 지표는 정오분류표인 matrix를 통해 모형의 평가값을 나타내고 있으며, 평가값의 해석은 다음에 나오는 정오분류표를 통해 자세히 설명하고자 한다.

본 연구에서 제안한 의사결정분석 모형을 통해 J강종을 평가한 결과는 <표 5>와 같으며. 에러율이 4.26%, 정확도가 95.74%, 민감도가 98.37%, 특이도가 55.92%이다.

〈표 5〉 의사결정나무 분석 분류행렬표

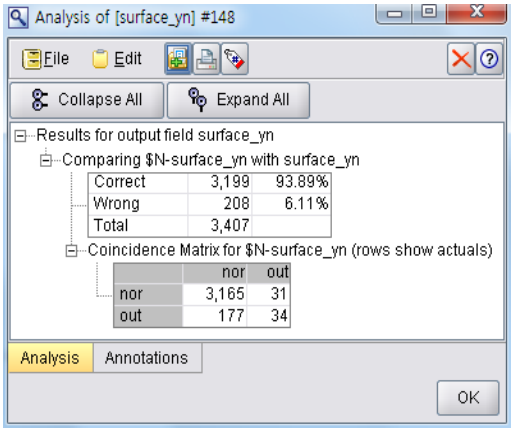
		예측집합		계
		nor	out	
실제 집합	nor	3,144	52	3,196
	out	93	118	211
계		3,237	170	3,407

error rate = 4.26%, accuracy = 95.74%
sensitivity = 98.37%, specificity = 55.92%

의사결정나무 분석 평가결과 정확도와 민감도는 우수하지만 이상 트랜잭션을 이상 트랜잭션으로 분류할 가능성인 특이도가 다른 결과값에 비해 낮음을 알 수 있다.

4.3 신경망 분석 결과

본 연구에서 클레멘타인을 이용하여 신경망 분석 결과를 실행한 출력 결과는 [그림 6]과 같다.



[그림 6] 신경망 분석 결과

[그림 6]의 결과를 해석하면 다음과 같다. 첫 번째 지표는 surface_yn의 실제값과 신경망 모형의 예측값인 $N-surface_yn를 비교한 것이다. 신경망 모형은 예측값이 93.89% 이상의 예측율을 보이고 있으며, 세 가지 기법 중 두 번째로 높은 예측율을 나타내었다. 두 번째 지표는 정오분류표인 matrix를 통해 모형의 평가값을 나타내고 있으며, 평가값의 해석은 다음에 나오는 정오분류표를 통해 자세히 설명하고자 한다.

〈표 6〉 신경망 분석 분류행렬표

		예측집합		계
		nor	out	
실제 집합	nor	3,165	31	3,196
	out	177	34	211
계		3,342	65	3,407

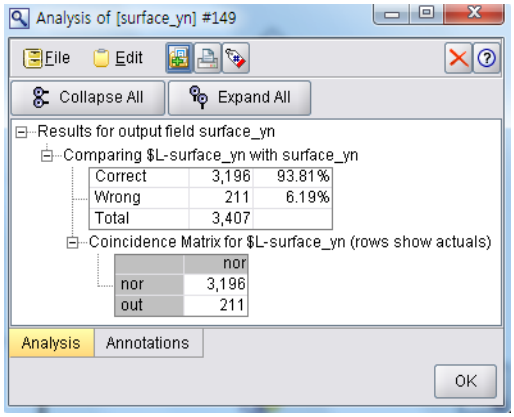
error rate = 6.11%, accuracy = 93.89%
sensitivity = 99.03%, specificity = 16.11%

본 연구에서 제안한 신경망분석 모형을 통해 J 강종을 평가한 결과는 <표 6>과 같으며. 에러율이 6.11%, 정확도가 93.89%, 민감도가 99.03%, 특이도가 16.11%이다.

신경망 분석 평가결과 정확도, 민감도는 우수하나, 이상 트랜잭션을 정상 트랜잭션으로 분류할 가능성인 특이도가 다른 결과값에 비해 낮음을 알 수 있다.

4.4 로지스틱 회귀분석 결과

본 연구에서 클레멘타인을 이용하여 로지스틱 회귀분석 결과를 실행한 출력 결과는 [그림 7]과 같다.



[그림 7] 로지스틱 회귀분석 결과

[그림 7]의 결과를 해석하면 다음과 같다. 첫 번째 지표는 surface_yn의 실제값과 로지스틱 회귀모형의 예측값인 $L-surface_yn를 비교한 것이다. 로지스틱 회귀분석 모형은 예측값이 93.81%의 예측율을 보이고 있으며, 근소한 차이지만 세 가지 기법중에선 가장 낮은 예측율을 보여주고 있다. 두 번째 지표는 정오분류표인 matrix를 통해 모형의 평가값을 나타내고 있으며, 평가값의 해석은 다음에 나오는 정오분류표를 통해 자세히 설명하고자 한다.

본 연구에서 제안한 로지스틱 회귀분석 모형을 통해 J강종을 평가한 결과는 <표 7>과 같으며, 에러율이 6.19%, 정확도가 93.81%, 민감도가 100%이나 특이도의 값이 0임을 알 수 있다.

〈표 7〉 로지스틱 회귀분석 분류행렬표

		예측집합		계
		nor	out	
실제 집합	nor	3,196	0	3,196
	out	211	0	211
계		3,342	65	3,407

error rate = 6.19%, accuracy = 93.81%
sensitivity = 100%, specificity = 0%

로지스틱 회귀분석 평가결과 정확도와 민감도는 우수하나, 이상 트랜잭션을 정상 트랜잭션으로 가능성인 특이도가 다른 결과값에 비해 월등하게 낮음을 알 수 있다.

4.5 분석 기법별 결과 비교

본 연구에서 제시한 세 가지 분석 기법인 의사결정나무 분석, 신경망 분석, 로지스틱 회귀분석의 결과에 따른 오차율(error rate), 정확도(accuracy), 민감도(sensitivity), 특이도(specificity)의 값은 <표 8>과 같다.

각 분석기법 별 정확도는 의사결정나무 분석 > 신경망 분석 > 로지스틱 회귀분석 순이며, 민감도는 로지스틱 회귀분석 > 신경망 분석 > 의사결정나무 분석 순이고, 특이도는 의사결정나무 분석 > 신경망 분석 > 로지스틱 회귀분석 순임을 알 수 있다.

〈표 8〉 각 기법의 분류행렬표 결과 비교

	decision tree	neural network	logistic regression
error rate	4.26%	6.11%	6.19%
accuracy	95.74%	93.89%	93.81%
sensitivity	98.37%	99.03%	100%
specificity	55.92%	16.11%	0%

세 가지 기법 모두 93% 이상의 높은 정확도와 민감도를 나타내었지만, 특이도 결과값의 차이로 인해 세 가지 기법에 대한 차이가 나타났다. 즉, 모형을 평가할 때 정확도와 오류율만을 가지고 평가하면 적절한 평가가 이루어 질 수 없으며, 민감도와 특이도를 고려한 평가가 이루어져야 함을 보여주고 있다. 따라서 정확도, 오류율, 민감도, 특이도를 모두 고려한 결과, 의사결정나무 분석의 특이도와 다른 기법의 특이도보다 높으므로 의사결정나무 분석이 가장 우수한 성능을 나타낸다는 것을 알 수 있었다.

5. 결론 및 제언

최근 경영환경의 급격한 변화로 많은 기업들은 경쟁우위를 유지하기 위해 품질에 영향을 미치는 요인과 사전에 품질을 관리할 수 있는 방법을 모색하고 있으며, 이를 위해 대용량의 데이터를 수집 및 활용하여 품질관리를 하고 있다. 그러나 철강기업들은 생산 공정상에서 생성되는 대용량의 데이터를 수집 및 저장하고 있지만 제품의 품질관리를 위해 활용하지 못하고 있는 실정이다. 특히 철강공정 중 대부분의 품질결함이 발생하는 연주공정을 실시간으로 관리하지 못함으로써 생산성이 감소되는 문제점이 있다.

따라서 본 연구에서는 국내 H철강 기업의 생산 공정 데이터를 이용하여 연주공정상에서 발생하는 품질결함을 파악하고, 실시간으로 생산공정을 관리하기 위해 데이터마이닝 기법인 의사결정나무분석을 이용한 연주공정상의 이상 트랜잭션 검출 모형을 제안하였으며, 신경망 분석 및 로지스틱 회귀분석을 이용한 모형과 비교분석 하였다.

분석결과 세 가지 기법 모두 93% 이상의 높은 정확도와 민감도를 나타내었지만, 특이도 결과값의 차이로 인해 세 가지 기법에 대한 차이가 나타났다. 즉, 정확도, 오류율, 민감도, 특이도를 모두 고려한 결과 의사결정나무 분석의 특이도와 다른 기법의 특이도의 차이로 인해 의사결정나무 분석이 가장 우수한 성능을 나타낸다는 것을 알 수 있었다

본 연구 결과를 통해 제안한 모형은 철강기업의 연주공정에서의 실시간 품질관리 및 사전 품질관리에 유용할 것으로 기대되며, 더 나아가서는 공정 중 이상 원인을 해결하는 과정에서 소요되는 시간과 비용의 감소를 통해 공정의 생산성을 향상시킬 수 있을 것으로 기대된다.

그러나 본 연구는 다음과 같은 한계점을 안고 있고 또 이에 따른 후속 연구를 필요로 하고 있다.

첫째, 데이터마이닝 모형에는 본 연구에서 사용된 모형 외에 여러 다른 모형이 존재하고 개발되고 있으나, 본 연구에서는 데이터마이닝 모형 중 의사결정나무(C5.0), 신경망(Quick), 로지스틱 회귀분석만을 사용했다는 것이다. 둘째, 실험 데이터의 여러 철강종류 중에서 대표 강종인 J강종만을 사용함으로써 철강강종 전체에 일반화하기에는 한계가 있을 수 있다.

추후 연구에는 본 연구를 기반으로 이러한 한계점을 보완하여 좀 더 세밀하고 다양하게 분석 할 수 있기를 기대한다.

참 고 문 헌

[1] 배성민, 이형욱, 이근안, 최석우, 박홍균, “데이터마이닝을 위한 공정 데이터 품질개선”, 「한국정밀공학회 2007년 춘계학술대회 논문집」, (2007), pp.795-796.

[2] 배정운, 「기초철강지식」, 한국철강신문, 2008.

[3] 변성규, 강창욱, 심성보, “데이터마이닝 기법을 이용한 제조 공정내의 불량항목별 예측방법”, 「한국산업경영시스템학회지」, 제27권, 제2호(2004), pp.10-16.

[4] 신택수, 홍태호, 「비즈니스 인텔리전스를 위한 데이터마이닝」, 사이텍미디어, 2009.

[5] 유성진, 강부식, 홍한국, “휴대용 카메라 모듈(CCM) 제조 라인에 대한 데이터마이닝 기반 품질관리 시스템 구축”, 「지능정보연구」 제14권, 제4호(2008), pp.89-101.

[6] 이극노, 이홍철, “이동통신 고객 분류를 위한 의사결정나무(C4.5)와 신경망 결합 알고리즘에 관한 연구”, 「한국지능정보시스템학회지」, 제9권, 제1호(2003), pp.139-155.

[7] 이상민, 이인렬, 신영길, “연속주조 공정에서 최적의 2차 냉각 Pattern 설정에 관한 연구”, 「춘계학술강연 및 발표개요집」, (1994), pp.205-205.

[8] 임세헌, 허연, “의사결정나무를 이용한 온라인 자동차 보험 고객 이탈 예측과 전략적 시사점”, 「Information Systems Review」, 제8권, 제3호(2006), pp.125-134.

[9] 장남식, “사전 세분화를 통한 고객 분류모형의 효과성 제고에 관한 연구”, 「Information Systems Review」, 제7권, 제2호(2005), pp.23-40.

[10] 장남식, 홍성완, 장재호, 「데이터마이닝」, 대청, 1999.

[11] Berry, M. J. and G. Linoff, “Data Mining Techniques : For Marketing, Sales, and Customer Support”, New York : John Wiley and Sons, 1997.

[12] Cheung, N., C. A. Santos, J. A. Spim, and A. Garcia, “Application of a Heuristic Search Technique for the Improvement of Spray Zones Cooling Conditions in Continuously

Cast Steel Billets," *Applied Mathematical Modelling*, Vol.30, No.4(2006), pp.104-115.

[13] Hancock, W. M., Yoon, J. W., and Plot, R., "Use of Ridge Regression in the Improved Control of Casting Process", *Quality Engineering*, Vol.8, No.3(1998), pp.395-403.

[14] Ma, J. C., Z. Xie, Y. Ci, and G. L. Jia, "Simulation and Application of Dynamic Heat Transfer Model for Improvement of Continuous Casting Process", *Materials Science and Technology*, Vol.25, No.5(2009), pp.636-639.

[15] Ture, M., F. Tokatli, and I. K. Omurlu, "The Comparisons of Prognostic Indexes Using Data Mining Techniques and Cox Regression Analysis in the Breast Cancer Data", *Expert Systems with Applications,* Vol.36, No.4(2009), pp.8347-8254.

◈ 저 자 소 개 ◈

김 재 경 (jaek@khu.ac.kr)

서울대학교에서 산업공학 학사, 한국과학기술원에서 경영정보시스템 전공으로 석사 및 박사학위를 취득하였으며 현재 경희대학교 경영대학 교수로 재직하고 있다. 미국 미네소타 주립대학교, 그리고 텍사스 주립대학교(달라스)에서 교환교수를 역임하였다. 주요 관심분야로는 비즈니스 인텔리전스, 추천시스템, 유비쿼터스 서비스, 사회 네트워크 분석 등이다. 저탄소 녹색성장국민포럼 그린IT분과 위원, 경희대학교 경영대학 BK21 사업단장, Information Technology and Management(SSCI)저널의 AE(Associate Editor)를 역임 중이다

권 택 성 (tskwon@khu.ac.kr)

삼육대학교에서 경영정보학 학사, 경희대학교 일반대학원 경영컨설팅학과에서 서비스경영 전공으로 석사학위를 취득하였으며, 현재 조선호텔 정보혁신 부서에 재직하고 있다. 주요 관심분야로는 비즈니스 인텔리전스, 데이터마이닝, CRM, 추천시스템, 사회 네트워크 분석, 그린 비즈니스/IT 등이며, 2010년 한국지능정보시스템학회 춘계학술대회에서 논문을 발표하였다.

최 일 영 (choice102@khu.ac.kr)

경희대학교에서 경제학 학사, 동 대학원에서 경영정보시스템 전공으로 경영학 MIS 전공으로 석사학위를 취득하였다. ㈜캐논코리아비즈니스 솔루션에서 대리로 근무 후 경희대학교 박사과정에 BK21사업 전일제 장학생으로 진학하여 2011년 박사학위를 취득하였으며 현재 경희대학교 경영대학 학술연구교수로 재직하고 있다. 주요 관심분야로는 CRM, 데이터마이닝, 그린 비즈니스/IT, 사회네트워크분석 등이며 경영과학회지, 경영과학, 정보관리학회지, 지능정보연구 등에 논문을 게재하였다.

김 혜 경 (kimhk@khu.ac.kr)

현재 경희대학교 경영대학에서 연구교수로 재직하고 있다. 경희대학교 물리학과에서 학사, 일반대학원 경영학과에서 MIS 전공으로 석사학위와 박사학위를 취득하였다. 주요 관심분야는 고객관계관리, 상품 추천 시스템, 사회 네트워크 분석, 복잡계 시스템 등이며, IEEE Transactions on Systems, Man, and Cybernetics(Part A : Systems and Humans), IEEE Transactions on Services Computing, International Journal of Information Management, Expert Systems, Expert Systems With Applications, 등 다수의 국제학술지에 관련논문을 게재하였다. Workshop on Information Technologies and Systems, Workshop on eBusiness 등 다수의 국제학술대회에서 논문을 발표하였다.

김 민 용 (andy@khu.ac.kr)

서울대학교 경영학과를 졸업하고, KAIST 경영과학과에서 MIS전공 공학석사와 박사학위를 취득하였으며, 현재 경희대학교 경영대학에서 교수로 재직하고 있다. 미국 카네기멜론 대학 SDS(Social and Decision Sciences) 학과의 방문교수로서 지식경영과 유비쿼터스 컴퓨팅을 연구하였다. 주요 연구분야는 유비쿼터스 컴퓨팅 응용, 지식경영, 비즈니스 인텔리전스 등이다. Behavior and Information Technology, Decision Support Systems, Journal of Knowledge Management, Expert Systems with Applications 등 다수의 외국학술지에 논문을 게재하고 있다.