

데이터마이닝을 이용한 스틸 파이프 생산 수율 측정에 관한 연구

김웅경 *, 김종완**, 남인길**

(Woong-Kyung Kim *, Jongwan Kim**, In-Gil Nam***)

요약 본 논문은 스틸 파이프 생산시 저불량율, 고수율을 가진 제품을 분류하고 모형화하기 위하여 과거 스틸 파이프 생산이력을 비교, 분석하여 주요 특성들이 불량율, 수율에 어떠한 영향을 미치는가를 파악함으로써 향후 스틸 파이프 생산 공정에서 저불량율, 고수율의 제품을 생산하는데 주요한 지표로 활용코자 하는데 그 목적이 있다. 과거 스틸 파이프 산업에 대한 주요 특성별 수율 측정에 대한 연구가 미흡하였으나, 본 논문에서 이를 구체화 하여 주요 특성별 불량률과 수율이 어떠한 형태를 나타내는지를 분류하고, 그 영향정도를 구분하고자 한다.

핵심주제어 : 데이터마이닝, 스틸 파이프, 의사결정트리, 생산 수율

Key Words : data mining, steel pipe, decision tree, production yield rate

I. 서론

현재 강관 즉 스틸 파이프는 굴뚝 산업 특성상 하향 국면으로 접어들고 있으며, 중국의 값싼 제품으로 인하여 가격 경쟁력에서 어려움을 겪고 있다. 또한 원자재 가격의 급등으로 전체적인 난항이 지속되고 있다. 이에 저부가가치 스틸 파이프 제품의 생산을 지양하고 고부가가치 및 고수익을 창출하는 제품으로의 전환이 필요한 시점이다. 세계 경제 위기와 불황속에서 스틸 파이프는 설비능력의 급증과 가동률 하락, 생산과 소비의 입지 불균형, 건설의존적인 수요구조, 중국으로부터 수입재 증가와 수출경쟁력 약화, 해외 Linepipe Projects 수주능력 취약 등의 문제점을 안고 있다.

이러한 문제점을 해결하기 위해서는 향후 저성장이 불가피한 건설용 강관, 특히 공급과잉이 심각한 중소구경 강관부문을 통합화하고 통합화 이후 잉여설비의 해외매각과 폐쇄 등을 추진, 공급과잉 완화 및 가동률 제고를 통해 국내 강관시장에서의

교섭력을 제고시키는 것이 필요하다. 또한 국내 강관 시장의 정체 극복을 위한 수출 확대를 위해서는 경쟁력을 갖춘 고부가가치 제품의 연구개발이 이루어져야 한다. 아울러 원소재인 열연강관의 공급부족을 해소하기 위해 상대적으로 국내 소재 수급여건 및 시황이 좋은 고부가가치 강관 생산 비중을 확대하는 Product Mix를 더욱 강화하고, 고부가가치 강관 개발은 소재 단계에서부터 결정된다는 점에서 강관사와 소재 공급사간의 협력을 강화해야 한다.

이러한 강관 산업의 문제점과 대응방안을 모색하기 위한 방안으로, 과거 강관 생산의 이력을 분석하여 주요 특성별 고부가가치 제품을 선별하고 각 공정별 최소 불량률을 가진 제품 모델을 구현하여, 고수익의 제품 생산을 지향하고 설비증설, 설비보수, 작업자 교육의 자료로 활용하고자 한다. 구체적 방법으로 주문을 통해 생산되는 각종 강관 제품들의 특성을 비교, 분석하여 어떠한 특성을 가지는 제품들이 가장 높은 수율을 창출해 내고, 적은 불량을 발생시키는지를 고찰한다. 이렇게 수집되고 비교·분석된 자료들을 데이터마이닝 기법으

* 대구대학교 대학원 컴퓨터정보공학과 석사과정

** 대구대학교 정보통신대학 컴퓨터IT공학부 교수

로 모델화함으로써 향후 생산관리를 위한 소중한 자료로 사용하는데 본 논문의 목적이 있다.

II. 문헌연구

2-1. 스틸 파이프

스틸 파이프(강관)란 단면형상이 원형 또는 각형으로 내부가 비어있는 강재를 말하며, 강철로 만든 파이프이다. 유체 수송용, 기계의 구조물, 일반 구조물용 등 광범위하게 사용된다. 용접강관은 고압용으로, 리벳 이음매강관은 저압용으로 사용된다. 가스·증기·물 등을 보내는 데 광범위하게 사용되는 강관은 이음매 없는 강관 또는 단접강관인데, 이것은 보통 가스관이라고 한다. 강관은 제조원가가 싸고, 강한 강도를 유지하며 쉽게 변형되지 않는 장점이 있으며, 또한, 진동이나 충격에 강하고 절단이 용이하며 쉽게 용접할 수 있으며 고온에서 치수를 안정되게 유지하여 튼튼하고 오래 견디며 열전도율이 좋다[5]. 그림 1은 스틸 파이프 중 ERW(전기저항용접)강관의 제조공정을 보여준다[7].

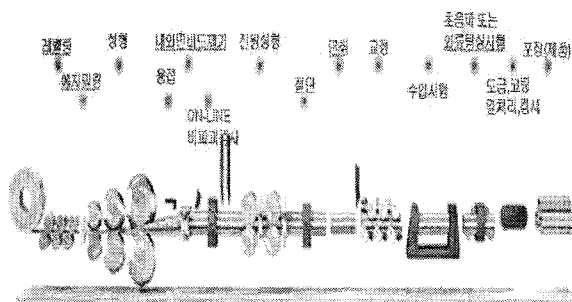


그림 1. 스틸 파이프의 제조공정(ERW)

2-1-1. 스틸 파이프 산업의 시장현황

국내 스틸 파이프 산업은 1990년대 이후 설비증가율이 소비증가율을 넘어서 구조적 공급과잉이 심화되었고, 2000년대 들어 중국산 저가재가 수입되어 국내 시장 잠식이 늘어나고 있다. 또한 타 철강업종 대비 시장구조가 파편화되어 불황시 과잉경쟁이 불가피한 상황이다. 이러한 스틸 파이프 산업의 경영실적 부진이 지속되는 가운데, 제품별/규모별 경기 양극화가 심화되고, 건설 및 유통거래 의존도가 높아 가격 변동성 확대가 불가피하다[4].

국내 스틸 파이프 산업의 문제점은 과열 경쟁으

로 인한 설비능력 급증과 가동율 하락, 대구경 설비는 경상권에 집중되고 국내 수요 60%는 경인권에 집중된 생산과 소비의 입지 불균형, 수요가보다 소재에 초점을 맞춘 공장건설 및 건설의존적인 수요구조 등이 지적되고 있다[5].

2-1-2. 스틸 파이프 산업의 향후전망

2008년 하반기부터 부각된 글로벌 금융위기 및 경기침체로 인해 국내 철강 수요는 급감하고 있으며 철강재 가격 역시 급락한 상황이다. 강관 부문 역시 국내 수요 위축으로 인해 실적이 둔화되고 있는 상황으로 경기불황의 골이 깊어질 것으로 보이는 2009년 1분기부터 실적 하락이 본격화될 것으로 전망된다. 특히, 공급과잉 및 낮은 가동률 등 국내 강관산업은 구조적으로 취약해 외부충격에 따른 타격이 상대적으로 클 것으로 보인다. Buy America로 글로벌 보호무역주의가 새로운 이슈로 떠오르고 있다. 각국의 자국 경기 부양을 위한 노력이 관세 인상 및 지원책이 이어지고 있는 상황으로 강관은 생산의 28%가량(2008년 11월 누계 기준) 수출되고 있으며 특히, 강관수출 중 대미 수출 비중은 69%에 달해 대미의존도가 가장 높은 것으로 나타났다. 따라서 경기침체로 인한 글로벌 강관 수요 감소 및 보호무역주의 강화로 국내 강관 수출의 부진은 더욱 커질 가능성이 높다. 본 연구는 스틸 파이프 생산 수율 측정을 데이터마이닝 기법으로 모형화하여, 향후 스틸 파이프 생산 공정에서 저불량률/고수율의 제품 생산을 목표로 한다[6].

2-2. 데이터마이닝

2-2-1. 데이터마이닝의 개념

데이터마이닝(Data mining)이란 수많은 데이터들로부터 단순한 검색 기법으로는 얻기 힘든 유용한 정보를 얻어내는 과정이라고 할 수 있다[1]. 이러한 데이터마이닝은 기업의 관점에서 보았을 때, 기업이 보유하고 있는 자료와 기타 외부자료를 포함하여 사용가능한 데이터를 기반으로 숨겨진 지식, 기대하지 못했던 패턴, 새로운 법칙과 관계를 발견하고 이를 실제 경영의 의사결정 등을 위한 정보로 활용하고자 하는 것이라 정의한다.

데이터마이닝은 대용량의 관측된 자료를 다루며, 이론보다는 실무위주의 컴퓨터 중심적인 방법이다.

또한 경험적 방법(adhockery method)에 근거하며, 일반화된 결과를 토출하는데 초점을 두고 있다. 데이터마이닝 기법들은 통계학(statistics), 컴퓨터과학(computer science), 경영학(management), 인공지능(artificial intelligence), 공학(engineering)과 같은 분야에서 개발되고 사용되어진다[3].

2-2-2. 연구에 사용된 데이터마이닝 기법

데이터마이닝에 대한 기법들은 여러 가지가 있으나, 본 논문에서 활용한 2가지 기법만 간단히 설명한다.

첫째, 회귀분석(Regression)은 변수들 사이의 관련성을 규명하기 위하여 어떤 수학적 모델을 측정하고 추정된 모형을 사용하여 필요한 예측을 하거나 관심 있는 통계적 추론을 하는 기법이다. 회귀 분석 기법을 분류하면, 독립변수가 한 개이고 종속변수가 한 개인 단순선형회귀(simple linear regression)와 독립변수가 두 개 이상이고, 종속변수가 한 개인 다중선형회귀(multiple linear regression), 독립변수와 종속변수의 관계를 곡선(비선형)으로 가정하여 분석하는 비선형회귀(non-linear regression)가 있다.

둘째, 의사결정트리(decision tree)는 데이터마이닝의 분류(classification) 작업에 주로 사용되는 기법으로 과거에 수집된 데이터의 레코드들을 분석하여 이들 사이에 존재하는 패턴, 즉 분류별 특성을 속성의 조합으로 나타내는 분류모형을 트리 형태로 만드는 것이다. 분류나 예측의 근거를 알려주기 때문에 이해하기 쉽고, 어떠한 속성들이 각각의 분류값에 결정적인 영향을 주는가를 쉽게 파악 가능하며, 모형구축에 소요시간이 단축된다는 장점이 있는 반면, 연속형 데이터를 처리하는 능력이 신경망(neural networks)이나 통계기법에 비해 떨어져 예측력이 감소하고, 모형을 구축하는데 사용되는 표본의 크기에 지나치게 민감하다는 단점이 있다.

2-2-3. 데이터마이닝의 과정

데이터마이닝 과정은 아래 그림 2와 같이 5단계로 구분한다. 첫째, 방대한 양의 데이터(모집단)에서부터 특성을 유지한 작은 양의 데이터(표본)를 추출하는 샘플링(Sampling), 이미 알고 있는 사실들을 확인하여 수치화하는 작업을 시작으로 하여 보유하고 있는 수많은 변수들의 관계를 살펴보는 탐색(Exploration), 기존의 변수를 이용하여 새로이 생성해야 할 변수를 고려하고 결정하는 변형 및

조정(Modification), 선정된 주요한 변수를 사용하여 다양한 모형을 적합해 보는 모형화(Modeling), 모형의 효과를 비교하여 가장 좋은 모형을 선택하고, 그 효과를 재평가하는 평가(Assesment) 단계로 구성된다.

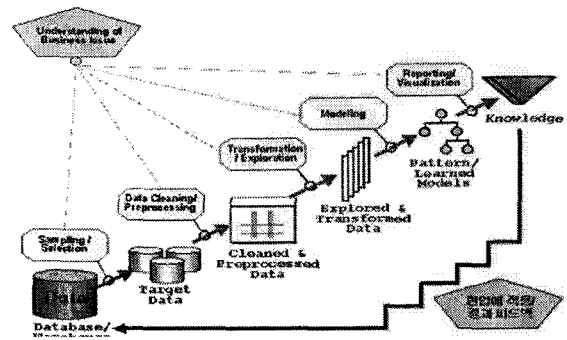


그림 2. 데이터마이닝의 처리단계

III. 스틸 파이프 생산 수율과 불량률 모델링

3-1. 연구모델 설정

이 장에서는 스틸 파이프 생산과 관련한 특성별 수율과 불량률을 모델링하여 연구 목적을 달성하기 위한 기초를 마련하고자 한다. 본 연구는 다음과 같은 연구 질문을 가지고 연구를 시작하였다.

첫째, 기존 생산된 스틸 파이프의 주요 특성별 수율과 불량률은 어떠한 형태로 모형화되며, 이로 인해 예측할 수 있는 결과는 무엇일까?

둘째, 수율과 불량률에 각 특성의 상관관계 및 연관성은 어떠하며 가장 큰 영향을 주는 특성(인자)은 어떤 것인가?

두 가지의 연구 질문을 해결하기 위하여 과거 스틸 파이프의 주요 특성별 생산자료와 수율, 불량률의 자료를 수집하여 분석하였으며, 스틸 파이프 생산에 영향을 끼치는 주요 특성은 표 1과 같이 파악되었다.

구분	변수명	설명
독립변수	물종	제품의 종류
	기준	제품의 제원 혹은 설계명세서
	관단	제품의 단면처리방법
	외경	제품의 바깥 지름
	두께	제품의 두꺼운 정도
	길이	제품의 길이
	패킹	제품의 포장방법
	코팅	제품의 표면처리방법
종속변수	수율	투입량 대비 제품량(정품)
	불량률	작업수량 대비 불량수량

표 1. 스틸 파이프의 주요 특성

3-2. 연구모형별 측정결과

첫 번째 연구 질문을 측정하기 위해서 2004년 4월부터 2009년 1월까지 OO제강에서 생산된 스틸 파이프의 생산이력을 추출하여, 불필요 및 오류 데이터를 제거한 후, Oracle 9i를 이용하여 데이터베이스(database)화 한 후, PowerBuilder 10.5를 이용하여 모형 및 시스템을 구현하였다. 이를 각 특성별 모형으로 나타내면 그림 3, 그림 4와 같다.

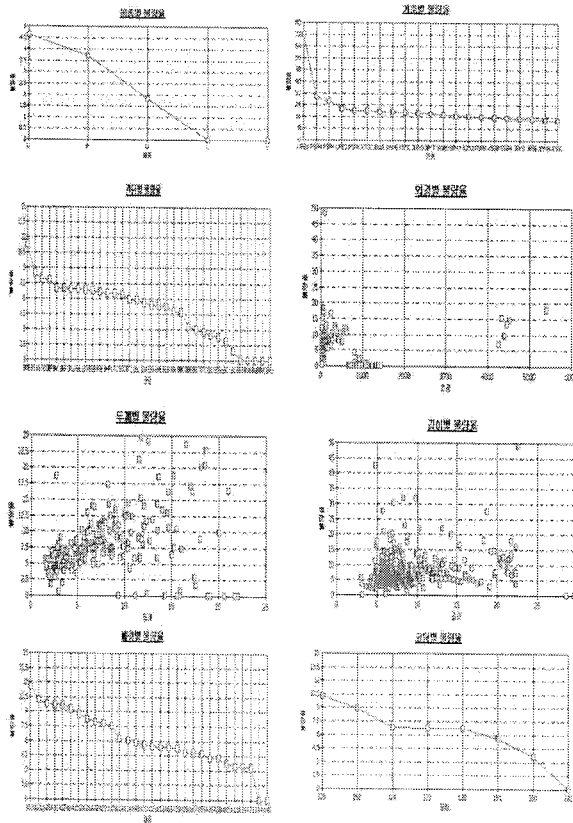


그림 3. 주요 특성별 불량률

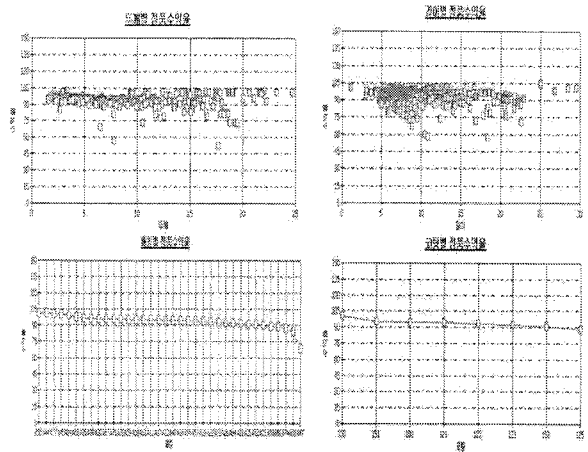
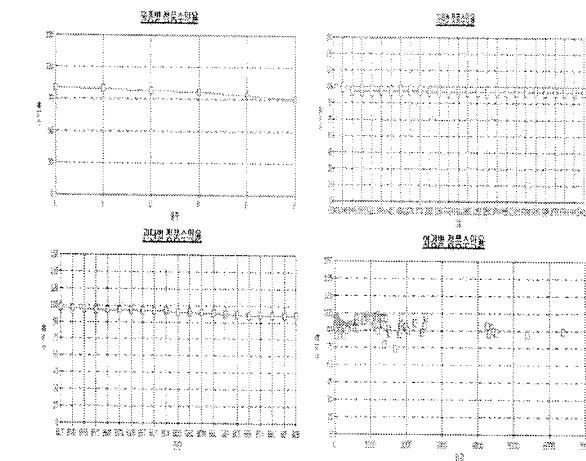


그림 4. 주요 특성별 수율

이러한 모형의 결과로 볼 때, 불량률의 경우, 외경은 중소구경일 때 집중 발생하며, 두께는 두꺼워질수록 불량률이 높아지고, 길이별로는 약 5~10m 구간에서 높은 발생률을 나타내는 것을 알 수 있다. 수율의 경우, 전반적으로 고른 분포를 나타내는데, 대체로 90~105% 사이의 분포를 보이며, 일부 두께가 두꺼운 제품과 길이가 약 5~10m 구간에서 낮은 수율을 나타내고 있다.

두 번째 연구 질문에 대한 측정은 Witten와 Frank[1]가 보급한 잘 알려진 데이터마이닝 개발 도구인 WEKA 틀을 이용하여 회귀분석과 의사결정트리의 두 가지 데이터마이닝 기법을 적용하였다. 회귀분석은 스틸 파이프의 주요 특성 가운데 외경, 두께, 길이를 독립변수로 하고, 불량률과 수율을 종속변수로 하여 선형 회귀식을 구한 후, 데이터베이스화 한 첫 번째 모형과 비교분석하였다.

구분	산출공식	상관계수
불량률	$= (-0.0113 * \text{mm_od}) + (0.7232 * \text{mm_thick}) + 6.5197$	0.18
수율	$= (0.0173 * \text{mm_od}) + (-0.7885 * \text{mm_thick}) + 93.8538$	0.15

표 2. WEKA 틀을 이용하여 얻어진 회귀식

그러나 표 2에서 보듯이 회귀분석으로 구한 회귀식은 상관계수(correlation coefficient)가 불량률은 0.18, 수율은 0.15밖에 안되므로 만점인 1.0에 비하면 너무 낮아서 상관관계가 없는 것으로 판단되었다. 본 연구의 모델링 기법으로 회귀분석 기법은 적합하지 않은 것으로 결정하고, 의사결정

트리를 이용하여 모델링을 수행하였다. 의사결정트리의 경우 종속변수의 값이 숫자가 아닌 문자가 되어야 하므로 종속변수인 불량률과 수율을 문자로 변환하기 위하여 불량률과 수율을 아래 표 3과 같이 구간 값으로 정의하였다.

수율구간	불량률구간	종속변수
> 100	= 0	A
100 <= AND > 90	1 >= AND <= 3	B
90 <= AND > 80	3 > AND <= 6	C
80 <= AND > 70	6 > AND <= 10	D
70 <= AND > 60	10 > AND <= 20	E
60 <= AND > 50	20 > AND <= 30	F
50 <= AND > 40	30 > AND <= 40	G
40 <= AND > 30	40 > AND <= 50	H
30 <= AND > 20	50 > AND <= 60	I
20 <= AND > 10	60 > AND <= 70	J
10 <= AND > 6	70 > AND <= 80	K
6 <= AND >= 1	80 > AND <= 90	L
= 0	90 > AND <= 100	M

표 3. 종속변수 구간별 변환표

의사결정트리기법을 통해 나타난 분석결과는 불량률은 약 61%, 수율은 약 69%의 분류 정확도를 나타내는데, 트리모형은 그림 5, 그림 6과 같다.

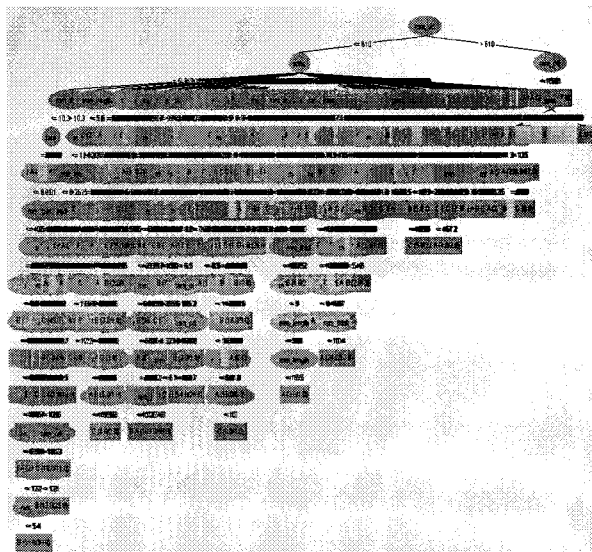


그림 5. 의사결정트리를 이용한 불량률 트리

이러한 의사결정트리 모형의 결과는 불량률과 수율 모두 독립변수 중 외경이라는 특성이 트리의 루트노드가 되며, 외경이 610mm 보다 큰지 아닌

지에 따라 1차 분류트리가 나뉘지고, 610mm보다 같거나 작을 경우에는 기호에 따라서, 클 경우에는 다시 불량률은 외경이 1500mm 보다 큰지 작은지에 따라서, 수율은 외경이 1220mm 보다 큰지 작은지에 따라서 분류되고 있음을 보여준다.

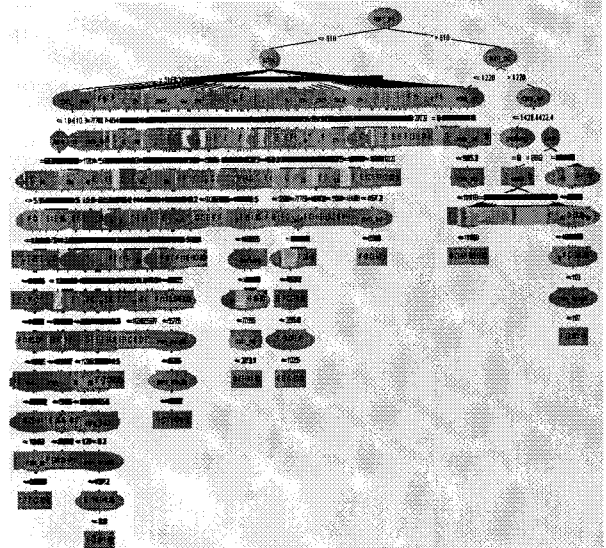


그림 6. 의사결정트리를 이용한 수율 트리

일반적으로 강관은 통상 외경 610mm를 기준으로 그 이상이면 대구경, 미만이면 중소구경 파이프로 구분된다. ERW 강관은 외경의 규격 범위가 최대 609.6mm으로 중소구경에 해당되는데, 분류트리의 결과로 볼 때, ERW 강관은 기호에 의해서, 그 외 주로 대구경 범위인 플랜더, 스파이럴 강관의 경우에는 외경에 의해서 불량률, 수율이 영향을 많이 받는다는 사실을 확인한 것이 본 연구의 의미있는 발견으로 판단된다.

IV. 결론 및 향후과제

본 연구는 스틸 파이프 생산에 영향을 미치는 주요 특성에 대한 연구모형을 제시하였으며, 주요 특성으로는 표 1에 나타난 것처럼 물종, 기호, 판단, 외경, 두께, 길이, 패킹, 코팅을 독립변수로 하고 불량률과 수율을 종속변수로 하여 독립변수에 대한 종속변수를 분류하여 독립변수가 종속변수에 어떠한 영향을 미치는가를 파악함으로써 향후 스틸 파이프 생산시 저불량율, 고수율의 제품을 생산하는데 주요한 지표로 활용코자 하는데 그 목적이 있다.

현재 스틸 파이프 산업은 여러 가지 불황과 위

기로 부가가치나 수익을 고려한 생산보다는 일단 고객의 수요에 대응하기 위해서 先생산 後수익평가의 형태를 보이고 있다. 그리하여, 여러 스틸 파이프 공급사간에 과도한 경쟁 및 설비 가동을 대비 저마진의 결과를 초래한다. 과거 스틸 파이프 산업에 대한 주요 특성별 수출 측정에 대한 연구가 미흡하였으나, 본 논문에서 이를 구체화 하여 주요 특성별 불량률과 수출이 어떠한 형태를 나타내는지를 분류하고, 그 영향정도를 구분하려 한데에 그 의의가 있다고 하겠다.

향후에는 본 논문을 기초로 하여 통계학에서 주로 사용하는 주요 특성간의 인자분석(factor analysis), 다변량 분석(multivariate analysis) 등을 통하여 독립변수의 제거 및 조정을 통한 분류정확도를 높이는 연구를 지속적으로 수행하여 분류결과 및 모형에 대한 신뢰도를 높이는 과정이 필요하다.

참고문헌

- [1] I. Witten and E. Frank, Data Mining: practical machine learning tools and techniques 2nd ed. Morgan Kaufmann, 2005.
- [2] 최종후외, “데이터마이닝 예측 및 활용”, SPPP아카데미, 2002년.
- [3] 백덕현, “근대 한국 철강산업 성장사”, 한국철강신문, 2007년.
- [4] 임종원, “철강산업의 경쟁과 협력”, 서울대학교, 2006년.
- [5] <http://www.kosa.or.kr>-한국철강협회.
- [6] <http://www.kmj.co.kr>-한국철강신문.
- [7] <http://www.hysco.com/>-현대하이스코.