「정보시스템연구」 제20권 제3호
한국정보시스템학회
2011년 9월, pp. 1~24
http://dx.doi.org/10.5859/KAIS.2011.20.3.1

데이터 마이닝 기반의 6 시그마 방법론 : 철강산업 적용사례*

장 길 상**

<목 차>

Ⅰ. 서 론

최근 많은 기업 및 기관, 그리고 학자들에 의해서, 프로세스 개선 및 다양한 문제해결 방법으로써 6 시그마(Six Sigma) 및 데이터 마이닝(Data Mining) 기법들이 주목을 받고 있다(나수천, 2005; 다카나시 토모히로와 만넨 이사오, 2004; Fayyad와 Simoudis, 1995).

6 시그마 방법론은 기업에서 발생하는 결함을 지속적으로 감소시킴으로써 기업의 제품, 서비스, 그리고 프로세스를 향상시키기 위한 고객, 프로세스, 데이터, 그리고 재무성과 중심적인 경영혁신 접근법이라고 말할 수 있다. 6 시그마 방법론에서의 결함은 제품, 서비스, 그리고 프로세스에서 발생 가능한 모든 결함을 포함한다(서정훈, 2005; 배영일, 2002; Banuelas, 2002; Hahn, 1999; Hoerl & Snee, 2002; Montgomery, 2001; Harry, 1994; Pande et al., 2000; Pyzdek, 2003).

이러한 6 시그마 방법론에 관한 기존의 연구는 크게 다음과 같은 두 가지로 분류될 수 있다. 첫째, 6 시그마를 적용하기 위한 프로세스와 주요 성공요인에 관한 분야로서, Murugappan과 Keeni(2003)는 품질 개선을 위한 6 시그마 추진 프로세스는 Define, Measure, Analyze, Improve, Control의 5단계로 실시해야 효과가 있다고 주장하였다(Hammer, 2002). 또한 Kwak과 Anbari (2006)는 6 시그마 방법론의 전개, 이점, 그리고

* 이 논문은 2010년 울산대학교 연구비에 의하여 연구되었음.
** 울산대학교 경영정보학과 교수, gsjang@ulsan.ac.kr

과제를 검토하고 성공적인 6 시그마 프로젝트 구현에 영향을 미치는 핵심 요소들을 제시하면서, 6 시그마의 성공은 최고 경영층의 관여, 조직의 헌신, 문화적 변화, 그리고 효과적인 프로젝트 관리를 통해 이루어질 수 있다고 주장하였다. 두 번째는 산업별 6 시그마 적용 사례연구에 대한 분야이다. Tong et al.(2004)은 PCB(Printed Circuit Board) 제조 공정의 불량률 감소를 위해 6 시그마 방법론을 적용하였고, 박재홍 등(2001)은 철강산업에서 극저탄소강의 품질 저하의 원인을 파악하여 개선하기 위해 공정데이터 분석을 통한 최적 조업조건 도출을 위해 구간 세분화 방법을 이용한 6 시그마 방법론을 제시하였다. 또한, 김형욱과 김종안(2000)은 금융서비스 산업에서의 6 시그마 도입 가능성에 대한 연구를 수행하기 위해 증권부문의 창구영업의 대기시간 단축 프로젝트에 6 시그마 방법론을 적용하였으며, 서영주와 함효준(2001)은 조립라인의 6 시그마 추진 시에 통계적인 방법보다는 작업자에 대한 표준 작업의 유지를 위한 작업관리시스템의 중요성을 주장하면서, 제품 선정, 작업표준 설정, DPO(Defects Per Opportunity) 관리, 문제 선정, DMAIC에 의한 평가, 그리고 제도화로 구성되는 DPO에 의한 관리모형을 제시하였다.

한편, 최근 들어 기업의 정보화 및 자동화가 급속하게 진전되면서 경영관리시스템 뿐만 아니라 현장관리 및 제어시스템을 이용하여 대량의 데이터를 축적하는 기업이 증가하고 있다. 즉, 이러한 기업들에는 방대한 업무 및 공정 데이터가 실시간으로 축적되기 때문에, 축적된 데이터로부터 프로세스 개선을 위한 유용한 정보를 찾아내고 분석하는 것이 무엇보다도 중요하게 되었다. 지금까지 6 시그마 프로젝트에서는 대상 프로세스의 문제점과 그 근본 원인을 파악하고 개선하기 위하여 상관분석, 회귀분석, 실험계획법 등과 같은 고도의 통계적 기법이 이용되어 왔다. 이것은 사실과 데이터를 중심으로 문제를 해결하는 6 시그마 방법론의 특징 때문이다. 따라서 지금까지 데이터의 통계적 분석을 위해서는 Microsoft Excel이나 MINITAB과 같은 스프레드시트(Spreadsheet) 소프트웨어를 주로 이용하여 왔다. 그러나 이러한 스프레드시트 소프트웨어를 이용할 경우, 수만 건 이상 되는 대량의 데이터를 분석하기에는 한계가 있었다. 이러한 한계를 극복하기 위하여, 대량의 축적된 데이터를 분석하여 유용한 정보를 찾아주는 데이터 마이닝(Data Mining) 기법들을 이용할 수 있다. 데이터 마이닝이란 축적된 대량의 데이터로부터 데이터 간의 관계, 패턴, 규칙 등을 찾아내고, 이를 모형화해서 의사결정을 돕는 유용한 정보로 변환하는 일련의 과정이다(안진석 등, 1999; 홍태호와 김진완, 2006; Fayyad와 Simoudis, 1995; Su et al., 2003).

이러한 데이터 마이닝을 적용한 기존의 관련 연구들 살펴보면, 대부분 단일한 영역에서의 의사결정을 위한 의미 있는 정보를 도출하기 위한 것으로, 6 시그마 방법론과의 연계에 관한 연구는 거의 이루어지지 않는 실정이다. 다만 이와 유사한 연구로는, Hancock et al.(1996)은 회귀분석을 이용하여 자동차 벨브 주조 공정에 공정변수들과 품질 특성치 간의 관계를 규명하여 불량률 감소를 위한 공정조건을 도출하였다. 지원철과 김우주(1998)는 신경망을 이용하여 품질설계를 위한 시뮬레이션을 체계적으로 지원할 수 있는 시스템을 데이터 마이닝의 관점에서 설계 및 구현하였다. 또한 안진석 등(1999)은 제조

업에서 수집된 공정 데이터로부터 공정개선이나 품질개선에 유용한 정보를 추출하기 위하여 신경망, 의사결정나무, 연관규칙을 이용하여 연속형 변수의 범주화와 유사빈도의 군집화를 통한 최적공정조건 도출을 위한 방법을 제시하였다. 그리고 민광기 등(1998)은 철강산업에서 열풍로 공정의 에너지 사용량 절감을 위해 특성 추출과 신경망을 이용하여 열풍로 열효율을 예측하기 위한 모델을 제시하였다.

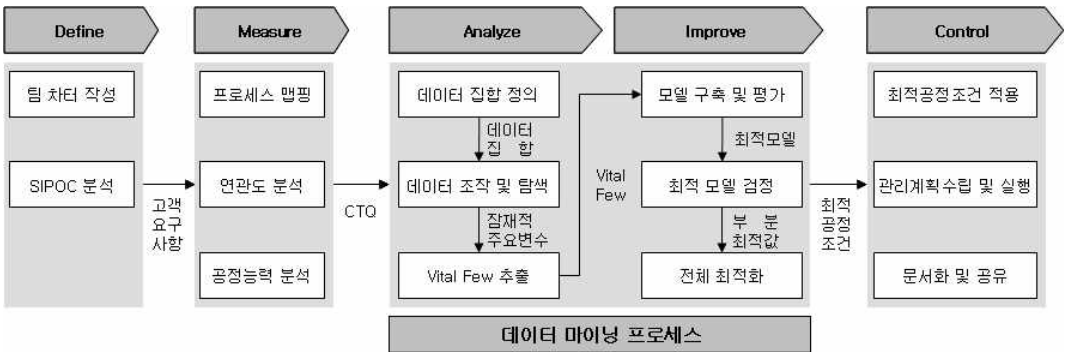
따라서 많은 조직에서 발생하는 문제해결을 위하여, 지금까지 각각 따로 적용되고 연구되어 오던 6 시그마 방법론과 데이터 마이닝 기법을 연계할 필요성이 제기되었다. 이러한 6 시그마 방법론과 데이터 마이닝 기법의 연계 방안에 관한 연구로는 Jang et al.(2005), Yachao(2008), 그리고 Jang & Jeon(2009)의 연구가 국제학술대회 Proceedings지에 발표되었다. 이들은 철강산업에서 데이터 마이닝 기법을 적용한 6 시그마 사례를 제시하였다.

이러한 관점에서, 본 논문에서는 최근 기업들에서 지속적으로 추진되고 있는 6 시그마 프로젝트에 대량의 데이터 분석을 효과적으로 수행하기 위한 데이터 마이닝 기반의 6 시그마 추진 방법론을 제시하고자 한다. 본 논문에서 제시한 방법론의 유용성을 입증하기 위하여, 대량의 데이터가 실시간으로 발생하는 "P" 철강회사의 열풍로 공정에서의 에너지 사용량 절감을 위한 6 시그마 프로젝트에 적용되어졌다.

Ⅱ. 데이터 마이닝 기반의 6 시그마 추진 방법론

2.1 본 방법론의 개요

일반적으로, 6 시그마 방법론에서는 상관분석, 회귀분석, 실험계획법 등과 같은 통계적 기법을 제공하는 MINITAB을 많이 이용한다. 그러나 많은 공정변수를 가지는 대량의 공정데이터가 실시간으로 축적되는 제조 산업에서는 MINITAB을 이용하여 축적된 데이터로부터 유용한 정보를 분석하고 찾아내는 데는 한계가 있다. 그러므로 본 논문에서는 기존의 6 시그마 방법론에 대량의 데이터를 분석할 수 있는 데이터 마이닝 기법을 적용하고자 한다. 특히, 본 방법론은 6 시그마 방법론의 다섯 단계 중에서 대량의 데이터를 분석해야 하는 Analyze 단계와



<그림 1> 데이터 마이닝 기반의 6 시그마 추진 방법론의 개요

Improve 단계에 초점을 맞추어 데이터 마이닝 기법을 적용한다. <그림 1>은 본 논문에서 제시하는 데이터 마이닝 기반의 6 시그마 추진 방법론을 보여주고 있다.

2.2. 본 방법론의 주요 프로세스

본 방법론에서 제시하는 각 단계별 주요 프로세스에 대한 설명은 다음과 같다.

(1) Define-Measure 단계

6 시그마를 적용할 프로젝트와 주요 고객을 정의하고, 고객의 요구사항을 분석한다. 분석된 고객요구사항을 이용하여 프로세스 맵핑, 연관도 분석 등을 통하여 핵심 개선 영역인 CTQ(Critical to Quality; 핵심개선영역)를 도출한다.

(2) Analyze 단계

대량의 데이터를 분석하여 CTQ에 영향을 미치는 근본원인인 Vital Few(핵심주요인자)를 도출한다. 이를 위해, 다음의 단계를 수행한다.

① 데이터 집합 정의 : 데이터 집합 정의 단계에서는 CTQ의 실제 공정으로부터 분석에 필요한 데이터 집합을 정의하고, 모델을 구축한 이후에 모델의 타당성 평가와 모델간의 비교를 위하여 데이터를 분할한다. 데이터 마이닝은 많은 변수를 가지고 있는 대량의 데이터를 대상으로 하여 다양한 방법론에 의한 분석을 수행하기 때문에 분석을 위한 데이터로써 모든 공정 데이터를 이용할 수는 없기 때문이다. 따라서 특정한 기간과 같은 조건을 이용하여 분석용 데이터를 추출하고, 추출된 데이터를 학습용(Training), 검증용(Validation), 평가용(Test)으로 각각 분리하여 분석용 데이터를 이용하여 모형을 구축하고, 평가용 혹은 검증용 데이터를 이용하여 모형의 비교나 최종적인 평가를 수행할 수 있도록 한다.

② 데이터 탐색 및 변형 : 데이터 탐색 및 변형 단계에서는 모델을 본격적으로 구축하기에 앞서, 분석하고자 하는 데이터 집합에 대한 탐색적 분석을 수행하여 데이터 집합 내의 많은 변수들 중에서 목표변수와 관련이 높은 잠재적인 주요 변수들을 선택하여 모델을 효과적으로 구축할 수 있도록 한다. 여기서 목표변수와 관련이 높은 변수들이란, 목표변수에 결정적인 영향을 미치는 주요변수가 아니라 잠재적으로 주요변수가 될 수 있는 잠재적인 주요변수의 집합을 의미한다. 또한, 결측치가 많은 데이터는 없는지, 이상치가 존재하는지의 여부를 검토하여 결측치가 지나치게 많은 데이터를 제거하고, 이상치를 제거하거나 대체하는 작업을 수행한다. 이때, 이상치는 항상 의미가 없거나 불필요한 것이 아니기 때문에 이상치를 충분히 검토한 후 필요한 경우 제거하거나 변환해야 한다.

③ Vital Few 추출 : Vital Few 추출 단계에서는 데이터 탐색 및 변형 단계에서 도출된 잠재적인 Vital Few를 이용하여 모델 구축에 필요한 Vital Few를 선택한다. 여기서 Vital Few란, 목표변수에 결정적인 영향을 미치는 변수의 집합을 의미한다. 이러한 작업을 수행하기 위하여, 목표변수에 대한 잠재적 Vital Few의 영향도 분석과 다변량 상관분석, 그리고 결정계수(R-Square) 혹은 카이제곱(Chi-Square)을 이용한 변수 선택법을 통해 모델 구축을 위한 주요변수인 Vital Few를 추출한다.

(3) Improve 단계

추출된 Vital Few의 최적화를 통하여 CTQ를 최적화하기 위한 모델을 구축하고 평가하여 부분 최적값과 최적 모델을 선정한다. 그리고 전체 최적화를 통하여 실제 공정에 이용할 수 있는 최적 공정 조건을 도출한다. 이를 위해, 다음의 단계를 수행한다.

① 모델 구축 및 평가 : 모델 구축 및 평가 단계에서는 Vital Few 추출 단계에서 선택된 Vital Few에 해당하는 데이터 집합을 입력받아 다양한 형태의 모델을 구축하고, 이들을 비교·평가한다. 모델을 구축하기 위하여 회귀분석(Regression), 의사결정나무(Decision Tree), 인공신경망(Neural Network) 등의 대표적인 데이터 마이닝 기법을 이용하고, 평가를 통해 가장 예측력이 뛰어난 하나의 모델을 최적 모델로 선정한다.

② 최적 모델 검정 : 최적 모델 검정 단계에서는 모델 구축 및 평가 단계에서 선정된 최적 모델을 이용하여 Vital Few별 조건에 따른 시뮬레이션을 실시한다. 이를 통해 최적 모델의 적합성을 평가하고, Vital Few에 대한 최적값 존재 가능 영역을 분석하여 최적 조건을 도출한다. 그리고 구간화 알고리즘을 이용하여 Vital Few에 대한 최적 구간을 도출한다.

③ 전체 최적화 : 전체 최적화 단계는 최적 모델 검정 단계에서 도출된 Vital Few에 대한 최적 구간을 실제 공정에 적용하기에 앞서 재검증 하는 단계이다. 이것은 데이터 마이닝을 통하여 제시되는 구간이 범위가 넓기 때문에 실제 공정에 적용하는데 필요한 최적의 구간을 찾기가 쉽지 않기 때문이다. 따라서 이를 해소하기 위하여, 모델 구축 및 평가 단계에서 선정된 최적 모델과 최적 모델 검정 단계에서 도출된 Vital Few에 대한 최적 구간을 이용하여 부가적인 시뮬레이션과 실험계획법(Design of Experiments)을 실시한다. 이를 통하여, 실제 공정에 적용 가능한 Vital Few에 대한 최적공정조건을 도출한다.

(4) Control 단계

핵심 개선영역에 도출된 최적공정조건을 실제 공정에 적용한다. 그리고 이를 지속적으로 유지하고 필요할 경우 개선하기 위한 관리계획을 수립하고 실행한다. 또한, 프로젝트를 통해 발생한 모든 산출물을 체계화하고 이를 공유한다.

<표 1>은 6 시그마의 Analyze 및 Improve 단계에서 데이터 마이닝 프로세스와 그 추진내용, 그리고 적용 가능한 데이터 마이닝 기법을 요약하고 있다.

본 논문에서는 데이터 마이닝을 위하여 강력한 데이터 마이닝 도구 중의 하나인 SAS Enterprise Miner를 이용하였다. SAS Enterprise Miner는 본 방법론이 제시하는 데이터 마이닝 프로세스를 수행하기에 적합한 다양한 데이터 마이닝 기법들을 제공한다. 지금부터는 <표 1>에서 6 시그마의 Analyze 및 Improve 단계의 데이터 마이닝 프로세스에서 적용하는 데이터 마이닝 기법에 대하여 간략하게 기술하고자 한다.

① Input Data Source: 분석에 이용할 데이터 집합을 선택하여 데이터의 역할(Training(학습용)-초기 모델 개발, Validation(평가용)-모형 평가용으로 초기 모형을 미세 튜닝, Test(검증용)-모형 검증 및 평가, Score(예측용)-새로운 데이터 집합을 입력하여 생성된 모델을 활용한 예측용, Document-텍스트 마이닝 적용용)을 지정하고, 데이터 마이닝 과정에서 변수가 각각 어떠한 역

<표 1> Analyze-Improve 단계별 항목과 데이터 마이닝 기법

단계	항목	기능	데이터 마이닝 기법
Analyze	데이터집합 정의	실제 공정으로부터 분석에 필요한 데이터 집합 정의	- Input Data Source - Data Partition
	데이터 탐색 및 변형	데이터집합의 이상치 제거 및 변환	- Multiplot - Distribution Explorer - Filter Outlier - Replacement
		데이터집합의 탐색적 분석을 통한 잠재적 Vital Few 도출	- Insight - Regression
	Vital Few 추출	CTQ에 영향을 미치는 핵심 주요인자인 Vital Few 추출	- Regression - Decision Tree - Insight - Variable Selection
Improve	모델 구축 및 평가	분석 모델의 구축	- Regression - Decision Tree - Neural Network
		최적 모델의 선정	- Assessment - Scoring
	최적 모델 검정	Vital Few의 최적 구간 도출	- Simulation - Insight - Interactive Grouping
	전체 최적화	최적 공정 조건 도출	- Simulation (Crystal Ball) - DOE (MINITAB)

할을 수행하는지 변수의 레벨(Interval, Binary, Ordinal, Nominal)을 지정한다. 따라서 데이터 마이닝 프로세스의 가장 처음에 위치한다.

② Data Partition: Input Data Source로부터 입력된 데이터 집합을 학습용, 평가용, 그리고 검증용 데이터 집합으로 분할한다. 학습용 데이터 집합은 모델의 사전 조정을 위해 이용되고, 검증용 데이터 집합은 평가하는 동안에 모델의 중요도를 관찰하고 조정하며, 모델을 평가하는데 이용된다. 평가용 데이터 집합은 모델을 평가하기 위하여 이용할 수 있도록 하는 부가적인 데이터 집합이다. 분할방법에는 단순(Simple Random), 층화(Stratified), 사용자 정의(User Defined) 방법이 있다.

③ Multiplot: 분석에 이용할 데이터 집합에 존재하는 대량의 데이터를 그래프를 이용하여 시각적 탐색을 통하여 데이터의 분포, 패턴, 추세를 이해할 수 있도록 하며 변수간의 관계정보도 제공한다. 제공되는 그래프는 막대그래프와 산점도이다.

④ Distribution Explorer: 분석에 이용할 데이터 집합에 존재하는 대량의 데이터를 다차원 히스토그램(Multidimensional Histogram)으로 탐색할 수 있도록 한다. 이를 통하여, 한 시점에 3개의 변수에 대한 분포를 살펴볼 수 있다.

⑤ Insight: 데이터의 탐색과 분석을 위한 대화식 도구로서, 여러 가지 형태의 그래프와 통계적 분석 결과를 통하여 데이터의 특성을 탐색할 수

있도록 한다. 또한 데이터와 그래프, 통계분석 결과가 상호 연동되는 기능을 제공함으로써 데이터에 대한 탐색 능력을 향상시켜 준다. 일반화된 선형 모델을 통해 일변량 및 다변량 분포의 탐색은 물론이고 회귀 적합, 상관분석, 주성분분석 등 다변량 분석도 가능하다.

⑥ Variable Selection: 목표변수를 분류하거나 예측하는데 이용되는 입력변수의 중요도를 R-square 혹은 트리 기반의 Chi-square 기준으로 평가하여 우선순위를 나타내 주며, 목표변수와 입력변수 간에 비선형적 관계가 있는 경우에는 각 변수를 자동으로 그룹핑하여 새로운 변수를 생성하고 제시해 준다. 입력 변수가 많은 경우 Variable Selection을 이용하여 유의한 변수를 사전에 줄여주고 이후 모델링 기법(회귀분석, 의사결정나무, 인공신경망 등)을 적용함으로써 분석과정의 효율성을 향상시킬 수 있다.

⑦ Data Set Attributes: 분석과정에서 Input Data Source에서 지정한 데이터의 속성을 변경할 필요가 있는 경우 데이터 집합의 속성을 변경할 수 있도록 한다. 즉 변수에 대한 사용여부, 역할, 레벨(Interval, Binary, Ordinal, Nominal), 정렬 순서 등을 변경할 수 있다.

⑧ Filter Outlier: 데이터 집합으로부터 이상치를 확인하고 제거할 수 있도록 한다. 제거방법은 변수의 형태에 따라 명목형 변수는 빈도 기반으로, 연속형 변수는 범위 기반으로 지정할 수 있으며, 빈도수가 적은 희박한 값 또는 극단적인 이상치를 제거할 수 있다. 이상치는 모델로부터 도출되는 결과에 크게 영향을 미칠 수 있으므로, 이를 사전에 확인하여 제거함으로써 분류 혹은 예측을 위한 모델의 정확도를 높일 수 있다.

⑨ Replacement: 결측값이나 특정한 값을 가지는 관찰치의 값을 다른 값으로 대체할 수 있도록 한다. 연속형 변수의 경우는 평균, 중위수, 범위의 중간값, 트리 기반 혹은 분산, 통계량 기반 등의 방법으로 결측치를 대체하고, 명목형 변수의 경우는 최빈값, 상수, 트리 기반 혹은 분산 기반의 대체를 통해 결측치를 대체한다. 또한 특정한 값 또는 범위의 값을 다른 값으로 대체할 수 있다. 주의할 사항은 결측값에 대한 처리와 특정한 값 대체 중에서 실행하는 순서에 따라 대체되는 값이 다를 수 있다.

⑩ Interactive Grouping: 각각의 입력변수가 목표변수를 최적으로 분류하는 기준을 근거로 변수를 자동으로 그룹핑해 주며, 분석자의 의도에 따라서 다양한 클래스를 형성하고, 대화식으로 입력되는 변수 값들을 원하는 클래스로 그룹화 할 수 있도록 한다. 이때 이용되는 목표변수는 반드시 이진형 변수여야 한다.

⑪ Regression: 선형 및 로지스틱 회귀분석 모델을 제공하는 분석 모델로서, 입력 변수로 연속형 및 명목형 변수 모두 이용할 수 있고 목표변수의 형태에 따라 예측값(연속형) 또는 분류(이진형 또는 순서형)를 위한 확률값을 제공한다. 로지스틱 회귀분석의 경우 연결함수로 Logit, Probit, CLogLog(곰페르츠 모형)을 제공한다. 변수 선택법으로는 단계적(Stepwise), 전진(Forward), 그리고 후진(Backward) 선택법을 제공하며 선택기준으로 AIC, SBC, 오분류율, 이익/손실함수 등을 활용할 수 있다. 또한 입력변수의 숫자를 고려한 최적의 모형 개발을 위한 최적화 방법으로 Trust-Region(40개 이하), Quasi-Newton(40~400개), Conjugate Gradient (400개 이상) 등을 제공한다.

⑫ Decision Tree: 목표변수의 형태에 따라

CHAID, CART, 그리고 C.45 등 다양한 알고리즘을 적용하여 최적으로 분류하는 입력변수와 중요도, 변수별 임계값을 제시하는 분류 및 예측 기법이다. 분석 결과에 대한 해석이 용이하고, 비선형 관계 및 입력변수간의 교호작용도 자동으로 고려하며 선형성, 등분산성 등의 제약을 받지 않는 비모수적 기법이다. 분리규칙으로는 명목형 목표변수-카이제곱, 지니 지수, 엔트로피 지수, 연속형 목표변수-분산분석의 F 통계량, 분산의 감소량이 있다. 이 기법은 시스템에 의한 자동적인 학습과 사용자 시나리오 기반의 대화식 학습 모두를 지원하고 있어, 사용자가 여러 가지 제약사항을 반영한 모델을 개발할 수 있다.

⑬ Neural Network: 복잡한 구조의 데이터에 대한 정확한 예측을 위한 반복적인 학습과정을 통하여 내재된 패턴을 발견하고 예측하는 비선형 모형이다. 예측모형 생성을 위한 아키텍처에는 일반 선형 모형(Generalized Liner Model), 다층 인식자(Multilayer Perceotron), RBF (Radial Basis Functions), 그리고 광범위한 조합, 활성화, 오차 함수 등이 포함된다. Neural Network의 최적화를 위하여 설명변수 개수, 은닉층(Hidden Layer) 개수, 각 층(Layer)의 개수, Neuron의 수, Activation/Combination 기능, Training 방법 등을 설정하여야 한다.

2.3. 본 방법론의 이점

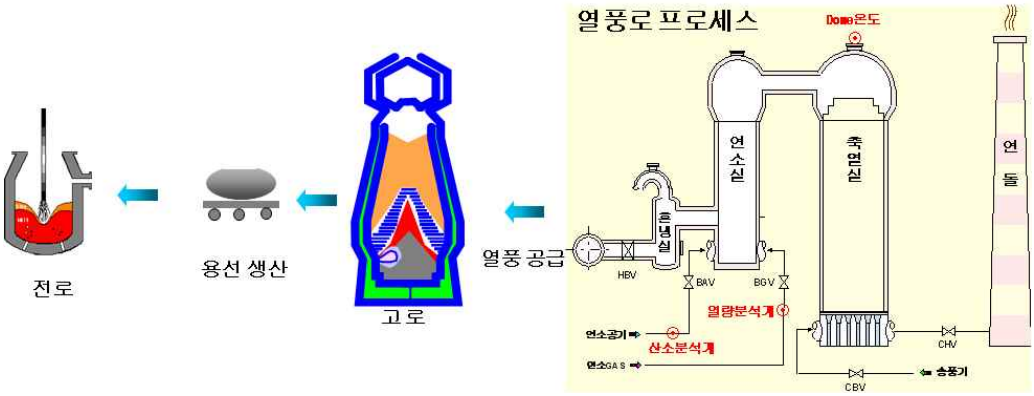
본 논문에서 제시하는 방법론은 MINITAB 등과 같은 스프레드시트 형태의 분석 도구를 이용한 기존의 6 시그마 방법론과 비교하여 다음과 같은 이점을 가질 수 있다: ① 효율적인 데이터 조작이 가능하다. ② 대량의 데이터 분석이 가능하다. ③ 비선형 예측을 위한 모델의 개발이 용이하다. ④ 통계적 가설을 검정하고 해석하기 용이하다. ⑤ 정확하고 일반화된 모델 개발이 용이하다. ⑥ 다양한 모델의 평가와 측정, 그리고 시각적 분석이 가능하다. ⑦ 데이터 마이닝 알고리즘에 의한 최적 통제범위 설정이 가능하다. ⑧ 하나의 분석 다이어그램을 통해 데이터 조작과 분석을 통합할 수 있다.

Ⅲ. 사례연구 : 철강산업 적용사례

본 절에서는, 제시한 데이터 마이닝 기반의 6 시그마 방법론을 “P” 철강회사의 열풍로 공정(Hot Stove Process)의 에너지 사용량 절감을 위한 6 시그마 프로젝트에 적용하고자 한다. 철강산업은 에너지 소비량이 많은 자본 집약적 산업으로 가동률, 생산성, 품질, 그리고 낮은 원가가 주요 경쟁 요소이다. 철강 산업의 제조 공정은 하부 공정이 상위 공정의 산출물과 상태에 매우 의존적이고 대량의 데이터가 발생한다는 특징을 가지고 있다. 따라서 전체적인 최적화가 다른 산업들 보다 더욱 중요하기 때문에 다변량 분석 및 다차원 분석이 필수이다. 본 절에서는 이러한 특징을 가지는 철강산업의 전체 공정 중 용선(Hot Metal)을 생산할 목적으로 고로(Blast Furnace)에 열풍(Hot Air)을 공급하는 열풍로 공정에 중심을 둔다.

3.1 Define 단계

열풍로 공정은 용선을 생산하기 위해 필요한 1,100~1,300℃의 열풍을 고로에 안정적으로 공

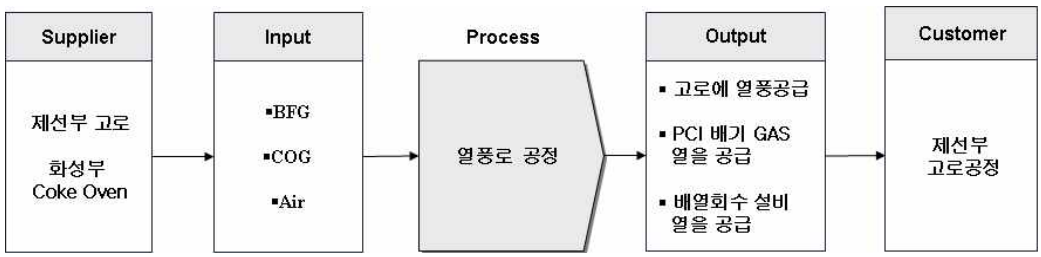


<그림 2> 열풍로 공정의 구성

급하는 공정으로, 통상 고로 1기에 4기의 열풍로가 설치된다. 열풍로 공정에서는 '연소'와 '송풍'이라는 작업이 병렬로 수행된다. 즉, 2기의 열풍로가 송풍을 하는 동안, 2기의 열풍로는 연소 과정에 돌입한다. 이러한 열풍로는 일반적으로 연소실과 축열실로 구성되어 있다. 열풍로는 "연소"와 "송풍"을 교호로 조반하여 조업하고 있고, "연소"라 칭하는 시간은 혼합 가스(m-gas)를 연소실에서 연소시켜 축열실 연와를 가열하고 연돌로 배기하는 약 60분의 시간이다. <그림 2>는 열풍로 공정의 구성을 보여준다.

열풍로 공정은 중요한 에너지 소비자로써, 고로에서 소비되는 전체 에너지의 10~15%를 소비한다. 따라서 열풍을 가열하기 위한 에너지 소비량의 절감은 용선의 생산원가에 중요한 영향을 미친다. 따라서 에너지 소비량에 직접적으로 영향을 미치는 요인들을 수치적으로 표현할 수 있는 열풍로 공정을 주요 개선 영역인 CTQ로 선정하였다. 그리고, SIPOC 분석을 실시하여 열풍로 공정에 대한 내·외부 고객과 투입물과 산출물을 도출하였다. <그림 3>은 열풍로 공정에 대한 SIPOC 분석의 결과를 보여준다.

SIPOC 분석의 결과를 토대로, 열풍로의 에너지 소비량 절감을 위한 고객 요구사항을 분석하기 위해 VOC(Voice of Customer) 및 VOB(Voice of Business)를 조사하였다. 그리고 그 결과를 이용하여 CCR(Critical Customer Requirement) 및 CBR(Critical Business Requirement)을 도출하였다. <표 2>는 Define 단계에서 도출된 결과물을 보여준다.



<그림 3> 열풍로 공정에 대한 SIPOC 분석 결과

<표 2> Define 단계의 결과물

구분	내용
VOC	- 설비 정상화 및 개선 필요 - 설비 감시화 필요
VOB	- Combustion control model 정상화 - Hot stove 일상관리 강화 - Hot stove 설비관리 강화
CCR	- 배가스 온도 250℃ 및 O_2 농도 2.2% - 풍온 설정에 맞는 Dome 온도 관리 - Gas 원 단위 절감
CBR	- 열풍로 돔 온도 도달 시간 10분 이내 - 열풍로 Dome온도 1290℃로 관리 - Checker 수금물 온도를 350℃ 이내로 관리

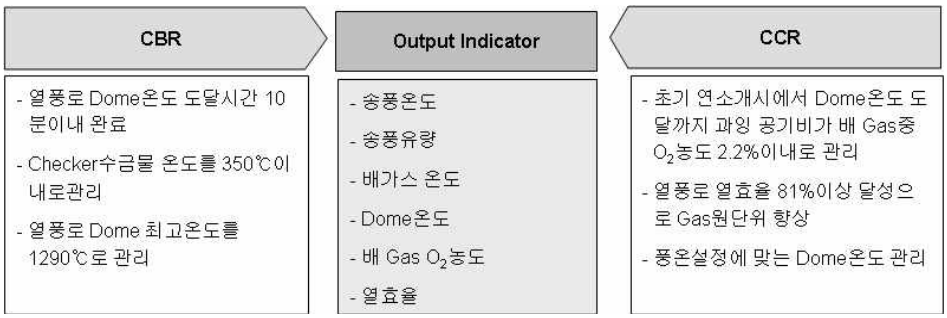
3.2 Measure 단계

Measure 단계에서는 Define 단계의 결과물을 기반으로 하여 고객의 요구사항을 정의하고, 개선을 위해 필요한 정보 및 데이터를 수집한다.

그리고 개선활동에 대한 기준과 목표를 구체화하여 CTQ를 선정한다. 먼저, Define 단계의 산출물인 CCR과 CBR을 기반으로 한 프로세스 매핑 작업을 통해 열풍로 공정에 대한 Output Indicator를 도출한다. <그림 4>는 Output Indicator의 도출 결과를 보여준다.

다음으로, 핵심 개선 영역인 CTQ를 도출하기 위해 CCR 및 CBR과 Output Indicator와의 연관도 분석을 실시한다. 연관도 분석에서는, Output Indicator를 기준으로 인자들 간의 관련성에 따라 점수를 부여하고, 점수의 합계가 높은 순서로 순위를 부여한다. 그리고 순위가 가장 높은 Output Indicator를 CTQ로 선정한다. 연관도 분석의 결과, 열효율(48점), 돔 온도(42점), 배가스 온도 및 O2 농도(36점), 송풍온도 및 유량(30점)의 순으로 순위가 결정되었으므로, 순위가 가장 높은 열효율이 CTQ로 선정되었다. <표 3>은 연관도 분석의 결과를 보여주고 있다.

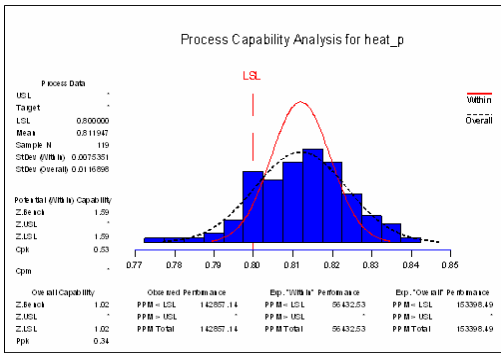
마지막으로, CTQ인 열효율에 대한 공정능력을 분석하고 현재 시그마 수준을 산정한다. <그림 5>와 같이, 열효율에 대한 공정능력을 분석한 결과 시그마 수준을 나타내는 Z_bench 값이 1.02σ로 나타났다. 이를 현재의 시그마 수준을 나타내는 Z_st로 변환한 결과, 열효율 공정의 현재의 시그마 수준은 2.52σ (Z_st = Z-bench + 1.5 = 1.02σ + 1.5) 수준으로 다소 낮게 나타났다. <그림 5>는 열효율에 대한 공정능력 분석의 결과를 보여준다.



<그림 4> Output Indicator 도출 결과

<표 3> 연관도 분석 결과

Output Indicator	CBR			CCR			합계	순위
	열풍로 돔 온도 도달 시간 10분 이내	Checker 수금물 온도를 350℃ 이내로 관리	열풍로 Dome온도 1290℃로 관리	배가스 온도 250℃ 및 O_2 농도 2.2%	풍온 설정에 맞는 Dome 온도 관리	Gas 원 단위 절감		
돔 온도	9	3	9	3	9	9	42	2
배가스 온도 및 O_2 농도	3	9	3	9	9	3	36	3
열효율	9	3	9	9	9	9	48	1
송풍온도 및 유량	3	3	9	3	3	9	30	4



<그림 5> 열효율에 대한 공정능력 분석

3.3 Analyze 단계

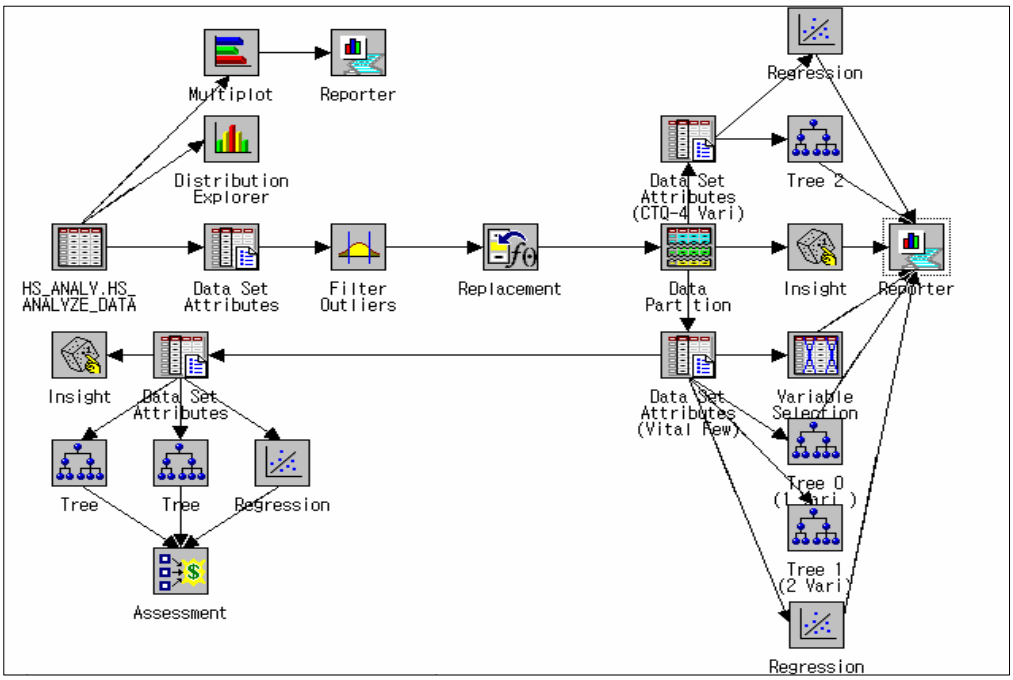
Analyze 단계에서는 개선의 대상이 되는 열풍로 공정의 실제 공정 데이터를 바탕으로 정성적인 인과관계를 이해하고 통계적인 상관관계를 분석한다. 이를 기반으로, CTQ인 열풍로 열효율에 직접적으로 영향을 미치는 핵심 원인 인자인 Vital Few를 도출한다. 이러한 활동을 수행하기 위해 데이터 집합 정의, 데이터 조작 및 탐색, Vital Few 추출의 절차에 따라 데이터 마이닝 기법을 적용한다. 앞서 기술하였듯이, 본 논문에서는 데이터 마이닝을 위하여 SAS Enterprise Miner를 이용한다. <그림 6>은 Analyze 단계에서 Vital Few를 도출하기 위해 이용한 Mining Flow Diagram을 보여주고 있다.

(1) 데이터 집합 정의

데이터 집합 정의 단계에서는, 실제 공정으로부터 분석에 필요한 데이터 집합을 정의한다. 먼저 4M1E(Machine, Man, Method, Material, Environment)에 따라서 프로젝트 팀원의 분임토의와 브레인 스토밍(Brainstorming)을 실시하여 열풍로 열효율에 영향을 미치는 인자들을 도출하였다. 도출된 인자들은 KJ(Kawakita Jiro) 법을 적용하여 분류하였고, AHP(Analytic Hierarchy Process) 분석을 실시하여 다음과 같은 4개의 잠재적 원인을 도출하였다.

- 최적의 과잉 공기비 설정이 필요하다.
- 연소말기 Mix gas 사용량이 과다하게 사용된다.
- 설비노후화로 방산열이 많다.
- Blast furnace gas, Coke oven gas 발열량이 부정확하다.

다음으로, 도출된 4개의 잠재적 원인을 토대



<그림 6> Analyze 단계에서의 Mining Flow Diagram

로 하여, 열풍로 연소관리 시스템으로부터 분석용 데이터 집합을 정의하였다. 초기 데이터 집합은 데이터 수집기간 5개월 동안 하루 1회의 측정주기로 모두 44개의 변수로 구성되었다. 이 중에서 날짜, 시간 등과 같은 분석 목표와 무관한 변수 6개를 제외하고, 최종적으로 38개의 변수로 데이터 집합을 구성하였다. <표 4>는 이렇게 정의된 데이터 집합의 구성을 보여주고 있다.

(2) 데이터 탐색 및 변형

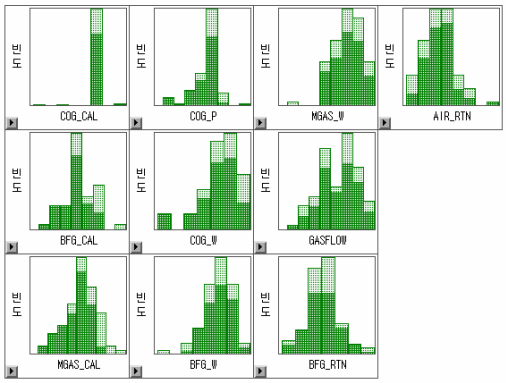
데이터 탐색 및 변형 단계에서는, 먼저, 본격

<표 4> 데이터 집합의 구성

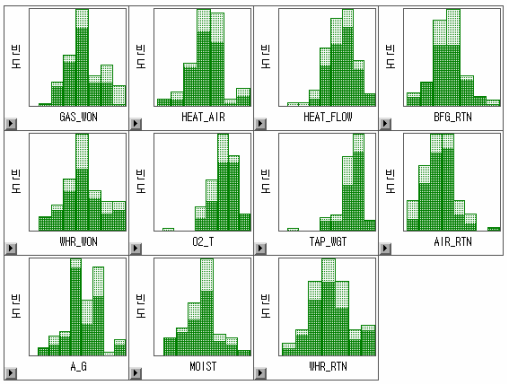
CTQ	분류	변수	합계
열풍로 열효율 (heat_p)	가스 성분	cog_cal, cog_p, mgas_w, bfg_cal, cog_w, gasflow, mgas_cal, bfg_w, bfg_rtn, Air_rtn	10
	설비조건	gas_won, heat_air, heat_flow, moist, whr_won, whr_rtn, o2_t, tab_wgt, air_rtn, a_g, bfg_rtn	11
	온도	fly_temp, mgas_tmp, out_tmp_max, heat_tmp, si_tmp_avg, bfgout_tmp, cog_tmp, bfgin_tmp, out_tmp_set, dom_tmp_avg, out_tmp_o2, si_tmp_min, dom_tmp_max, air_tmp, cool_tmp, out_tmp_avg, si_tmp_max	17
	제외 변수	ban_per, ban_wgt, mcv, date_t, trt_day, trt_pwr	6

적인 데이터 마이닝을 수행하기 이전에 데이터 집합에 대한 탐색적 분석을 통해 잠재적 주요변수를 선택하고, 데이터 집합 내에 존재하는 이상치를 제거 및 변환하는 작업을 수행한다. 이를 위해, 데이터 집합을 Insight로 분석하여 이상치의 존재여부를 확인하였다. 그 결과, cog_p, gas_won, heat_p, mgas_w 등의 변수에서 이상치가 존재하였으므로, SAS Enterprise Miner의 Filter Outlier와 Replacement 기능을 이용하여 이상치를 제거하고 변환하였다.

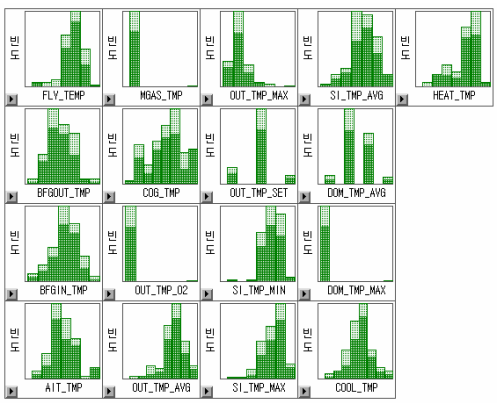
다음으로, 데이터 집합에 대한 탐색적 분석을 통해 잠재적 Vital Few를 선택한다. 이를 위해, 데이터 집합의 분류별로 목표변수인 열풍로 열효율에 대한 분포를 조사한다. 본 논문에서는 Insight를 이용하여 열효율이 80% 이상인 경우에 대하여, 가스성분, 설비조건, 그리고 온도에 해당하는 각각의 변수별로 분포 조사를 실시한다. 목표변수에 대한 개별 변수들의 분포를 직관적으로 살펴보기 위하여 조건을 'heat_p > 0.8'으로 설정하고, 분류별 변수들에 대한 히스토그램을 살펴보면, 가스성분 변수들의 경우에는 <그림 7>과 같이 ‘cog_p', 'mgas_w', 'cog_w', 'gasflow', 'mgas_cal', 'bgf_w' 등의 변수들이 상대적으로 열효율에 대한 영향도가 있음을 알 수 있다, 설비조건 변수들의 경우는 <그림 8>과 같이 'gas_won', 'heat_air', 'a_g', 'moist' 등의 변수들이 상대적으로 열효율에 대한 영향도가 있음을 알 수 있고, 온도영향 변수들의 경우에는 <그림 9>와 같이 ’air_tmp', 'cool_tmp' 등의 변수들이 상대적으로 열효율에 대한 영향도가 있음을 확인할 수 있었다.



<그림 7> 가스성분 변수의 히스토그램



<그림 8> 설비조건 변수의 히스토그램



<그림 9> 온도영향 변수의 히스토그램

그리고 Insight에서 제공하는 일반화된 선형회귀모형을 이용하여 목표변수에 대한 특성요인

영향분석을 실시한다. <그림 10>은 회귀분석을 통해 도출된 결과를 보여준다.

The DMREG Procedure

Analysis of Parameter Estimates

Parameter		DF	Estimate	Standard Error	t Value	Pr>|t|
COG_CAL		1	-0.00005	9.086E-7	-58.79	<.0001
COG_P		1	-0.0179	0.00243	-7.34	<.0001
COG_TMP		1	3.717E-6	8.563E-6	0.43	0.6657
COG_W		1	-2.87E-7	6.001E-8	-4.79	<.0001
COOL_TMP		1	-0.00078	0.000014	-57.45	<.0001
DOM_TMP_AVG	1259	1	0.000123	0.000399	0.31	0.7592
DOM_TMP_AVG	1285	1	5.643E-6	0.000102	0.06	0.9560
DOM_TMP_AVG	1286	1	-0.00003	0.000101	-0.33	0.7417
DOM_TMP_AVG	1287	1	0.000070	0.000108	0.65	0.5202
DOM_TMP_MAX		1	5.937E-6	3.956E-6	1.50	0.1385
FLV_TEMP		1	-0.00025	0.000089	-2.82	0.0064
GASFLOW		1	-8.5E-6	3.914E-8	-217.22	<.0001
GAS_WON		1	9.126E-9	1.238E-8	0.74	0.4639
HEAT_AIR		0	0	.	.	.
HEAT_FLOW		1	0.000263	8.627E-7	304.88	<.0001
HEAT_TMP		1	0.000857	0.000020	43.83	<.0001
MGAS_CAL		0	0	.	.	.
MGAS_TMP		1	-0.00002	0.000012	-1.49	0.1415
MGAS_W		0	0	.	.	.
MOIST		1	-3.32E-6	7.511E-6	-0.44	0.6599
O2_T		1	2.202E-6	2.802E-8	78.60	<.0001
OUT_TMP_AVG		1	-3.94E-6	0.000020	-0.20	0.8458
OUT_TMP_MAX		1	1.447E-6	7.579E-6	0.19	0.8492
OUT_TMP_O2		1	-7.61E-8	8.239E-8	-0.92	0.3595
OUT_TMP_SET	272	1	-0.00018	0.000332	-0.54	0.5936
OUT_TMP_SET	280	1	0.000159	0.000123	1.29	0.2008
OUT_TMP_SET	282	1	0.000100	0.000119	0.84	0.4059
SI_TMP_AVG		1	7.95E-6	6.873E-6	1.16	0.2518
SI_TMP_MAX		1	-2.69E-6	0.000011	-0.25	0.8012
SI_TMP_MIN		1	-0.00002	7.836E-6	-2.28	0.0263
TAP_WGT		1	4.558E-7	6.544E-7	0.70	0.4887
WHR_RTN		1	4.69E-12	1.96E-11	0.24	0.8120
WHR_WON		0	0	.	.	.

<그림 10> 선형 회귀모형을 이용한 특성요인 영향 분석 결과

분석 결과, 데이터 집합에 존재하는 38개의 변수 중에서 T-Value가 크고, P-Value(Pr>|t|)가 0.05보다 작은 17개의 변수(21개의 변수 제외)가 잠재적 Vital Few로 선정되었다. <표 5>는 데이터 탐색 단계를 통해 선정된 잠재적 Vital Few를 보여준다.

마지막으로, 본격적인 분석 모형에 이용하기 위하여 SAS Enterprise Miner의 Data Partition이라는 도구를 이용하여 데이터 집합을 분석용(Training), 검증용(Test), 그리고 평가용(Validation)으로 분할한다. 본 사례에서는, 데이터 분할 방법으로 단순 임의추출법을 이용하였고, 분할 비율은 분석용: 90%, 평가용: 0%, 검증용: 10%로 수행하였다.

(3) Vital Few 추출

이 단계에서는, 데이터 탐색을 통해 도출된 잠재적 Vital Few로부터 열풍로 열효율에 영향을 미치는 핵심인자인 Vital Few를 추출한다. 이를 위해, Insight, Variable Selection, Regression, 그리고 Decision Tree와 같은 네 가지 기법을 이용하여 잠재적 Vital Few로부터 Vital Few를 추출하기 위한 분석을 수행하였다. 먼저, Insight는 데이터의 탐색과 분석을 위한 대화식 도구로, 입력변수와 목표변수에 대한 히스토그램과 산점도를 생성하여 변수들 간의 관계를 직관적으로 분석할 수 있게 한다. 따라서 히스토그램과 산점도를 이용하여 열효율에 대한 변수별 영향도를 분석한 결과, bfg_cal, mgas_cal, cog_p, 그리고 mgas_w가 열효율과 관련이 있을 가능성이 있음을 확인할 수 있었다. 그리고 변수간의 관계를 분석하기 위해 다변량 상관분석을 실시하였다. 분석 결과, 열효율에 대한 상관계수가 높고, P-Value가 0.05보다 작은 bfg_cal, mgas_cal, cog_p, mgas_w가 역시 열효율과 관련이 있을 가능성이 있음을 확인할 수 있었다.

두 번째로, Variable Selection은 목표변수를

<표 5> 잠재적 Vital Few의 구성

CTQ	분류	변수	합계
열풍로 열효율 (heat_p)	가스 성분	cog_cal, cog_p, mgas_w, bfg_cal, gasflow, mgas_cal, bfg_w	7
	설비조건	gas_won, heat_air, heat_flow, moist, a_g	5
	온도	mgas_tmp, heat_tmp, out_tmp_o2, cool_tmp, out_tmp_avg	5

예측하거나 분류할 때 입력변수의 중요도 평가를 통해 목표변수와 무관한 변수는 제거하고, 관련성이 높은 변수를 선택하여 제시하는 도구이다. 이를 통한 분석 결과, 데이터 집합에 존재하는 17개의 변수 중 mgas_w, bfg_w, cog_p, mgas_cal 등을 포함하는 14개의 변수가 선택되었다. 그러나 선택된 14개의 변수들의 결정계수가 매우 낮으므로, 선택된 입력변수에 대한 결정계수를 누적하여 표현한 Effect 도표를 이용하여 목표변수에 대한 영향도가 상대적으로 높은 mgas_cal, gas_won, a_g, heat_air를 핵심인자로 선정하였다.

세 번째로, Regression을 이용하여 핵심인자를 선택하기 위한 분석을 수행하였다. 변수 선택을 위한 방법으로는 Best Subset, Backward, Forward, 그리고 Stepwise의 4가지 방법을 이용하였다. 핵심인자를 선택하는 기준은 다음과 같다. ① T-Value가 크고 P-Value가 0.05보다 작은 변수들에 우선순위를 두고, ② 변수별 중요도를 의미하는 Effect T-scores와 변수별 추정치를 의미하는 Parameter Estimates의 절대 값을 비교하여 이 값이 높은 4개의 변수를 최종적으로 선택한다. Best Subset은 모든 입력변수를 사용하여 모형 적합을 수행하는 방법이다. 이를 이용하여 분석한 결과, 조건 ①에 따르면, a_g, bfg_cal, bfg_w, cog_p, gas_won, 그리고 heat_air가 핵심인자로 고려될 수 있다. 핵심인자로 고려되는 6개의 변수에 대하여 조건 ②를 적용하여 보면, a_g, bfg_cal, bfg_w, 그리고 cog_p를 핵심인자로 선택할 수 있다. Backward 방법을 적용하여 분석한 경우, 조건 ①과 조건 ②를 적용해 보면, a_g, bfg_cal, bfg_w, 그리고 cog_p를 핵심인자로 선택할 수 있다. Forward 방법을 적용하여 분석한 경우, 조건 ①과 조건 ②를 적용하여보면, a_g, bfg_w, 그리고 gas_won을 핵심인자로 선택할 수 있다. Stepwise 방법을 적용하여 분석한 경우에도 조건 ①과 조건 ②를 적용하여 보면, a_g, bfg_w, 그리고 gas_won을 핵심인자로 선택할 수 있다. 이렇게 Regression을 이용하여 최종적으로 핵심인자를 도출하기 위해, 앞에서 수행한 4가지 방법을 통해 도출된 결과를 비교 및 평가하여, 최종적인 핵심인자 도출을 위해, 각 방법별로 선정된 변수들의 순위에 따라 평가 순위를 산정하였다. 그 결과 a_g, bfg_w, bfg_cal, 그리고 cog_p가 최종적으로 핵심인자로 선정되었다.

네 번째로, Decision Tree를 이용하여 핵심인자를 선택하기 위한 분석을 수행하였다. 여기서, 목표변수인 열효율(heat_p)은 연속형 변수이기 때문에, 분리 기준(Splitting criterion)을 F test로 설정하고, 유의수준은 0.2로 설정하였다. 분리를 위한 관측치의 수는 2로 설정하고, 이외의 나머지 조건들은 기본 설정을 유지하였다. 이러한 조건에 따라 상위의 4개의 변수를 핵심인자로 선정하기 위해 4단계로 분석을 수행하였다. 1단계에서는 17개의 변수를 입력변수로 하여 분석을 수행하였다. 그 결과 mgas_cal이 최상위 변수로 선정되었으므로 핵심인자로 선택하였다. 2단계에서는 1단계에서 선정된 mgas_cal을 제외한 16개의 변수를 입력변수로 하여 분석을 수행한 결과 cog_p가 최상위 변수로 선정되었으므로 핵심인자로 선택하였다. 이와 같은 과정으로, 3단계와 4단계 분석을 수행하여 각각 bfg_cal과 mgas_w를 핵심인자로 선택하였다.

마지막으로, 지금까지 분석에 이용한 4가지 도구에서 도출된 결과들을 평가하여 최종적인

핵심인자를 선택한다. 평가를 위한 방법은 다음과 같다. ① 모든 입력변수들에 대해서 4가지 방법에 의해 도출된 중요도에 따른 순위를 나열한다. ② 변수별 순위와 빈도를 비교하여 높은 순위가 많은 순서로 평가 순위를 결정하고, 상위의 변수 4개를 핵심인자로 선정한다. 이러한 조건에 따라 평가를 수행한 결과, mgas_cal, cog_p, a_g, 그리고 mgas_w가 핵심인자로 결정되었다. <그림 11>은 핵심인자 선택을 위한 평가 결과를 요약하고 있다.

Name	Insight	Variable Selection	Regression	Decision Tree	Assesment
mgas_w	7	5		4	4
bfg_w	4		2		
cog_w	5				
cog_p	2		4	2	2
mgas_cal	1	1		1	1
bfg_cal	3		3		
cog_cal					
heat_p					
heat_tmp					
moist		4			
dom_tmp_avg					
out_tmp_o2					
a_g		3	1	3	3
mgas_tmp					
ait_tmp					
Gas_won	6	2			

<그림 11> Vital Few 선정결과 요약

아래 <표 6>은 이렇게 최종 선정된 열풍로 열효율에 대한 Vital Few를 정리하고 있다.

<표 6> 열풍로 열효율에 대한 Vital Few

Vital Few	설명
mgas_cal	Mix gas calorie
cog_p	Coke-oven gas percentage
a_g	Air gas percentage
mgas_w	Mix gas volume

3.4 Improve 단계

Improve 단계에서는 Analyze 단계에서 도출된 Vital Few의 최적화를 통하여 CTQ를 최적화하기 위한 모델을 구축하고 평가한다. 그리고 최적의 모델을 선정하여 실제 공정에 이용할 수 있는 최적 공정 조건을 도출한다. 이러한 활동을 수행하기 위하여, 모델 구축 및 적용, 모델 평가 및 최적화, 그리고 전체 최적화의 단계로 데이터 마이닝 기법을 적용한다.

(1) 모델 구축 및 평가

첫 번째 단계에서는, Vital Few를 이용하여 열

<표 7> 분석을 위한 모델의 종류

모델	기법	설명
Regression	Default	교호작용을 고려하지 않은 다중선형회귀분석
	Interaction	교호작용을 고려한 다중선형회귀분석
	Polynomial	다항회귀분석
Decision Tree	F_test	F 통계량에 따른 분리
	Variance Reduction	분산 감소량에 따른 분리
Neural Network	I-H(3)-O	I: 입력층, H: 은닉층, O: 출력층 (괄호 안의 숫자는 노드 수)
	I-H(5)-H(3)-O	
	I-H(7)-H(3)-O	

<표 8> 모델 평가 결과

모델	기법	Root ASE	Test Root ASE	기술적 판단
Regression	Default	0.00415	0.93811	-
	Polynomial	0.00406	1.07394	-
	Interaction	0.00415	0.93811	-
Decision Tree	F_test	0.00396	0.00732	3
	Variance Reduction	0.00216	0.00953	-
Neural Network	I-H(3)-O	0.00203	0.01511	-
	I-H(5)-H(3)-O	0.00419	0.00730	2
	I-H(7)-H(3)-O	0.00371	0.00737	1

효율에 대한 다양한 모델을 구축하고 실제 공정 데이터를 이용한 분석을 실시한다. 그리고 구축된 모델들을 평가하여 최적 모델을 선정한다. 본 논문에서는 분석을 위해 Regression, Decision Tree, Neural Network을 이용하여 모두 8개의 모델을 구축하였다. 각각의 모델로 입력되는 데이터는 Analyze 단계에서 도출된 Vital Few에 대한 데이터 집합이다. <표 7>은 분석을 위하여 구축한 모델의 종류를 보여주고 있다.

분석을 위해 적용된 8가지 모델을 Assessment를 이용하여 Root ASE(Average Squared Errors)를 기준으로 평가한 결과, Root ASE가 작고 Root ASE와 Test Root ASE 사이의 차이가 적은 세 번째 Neural Network [I-H(7)-H(3)-O]이 최적의 모델로 선정되었다. <표 8>은 모델 평가의 결과를 보여준다.

(2) 최적 모델 검정

두 번째 단계에서는, 선정된 최적 모델을 기반으로 Vital Few별 존재 가능 영역을 탐색하기 위하여, SAS code 노드를 이용하여 새로운 데이터를 생성하고, 이를 적용하여 시뮬레이션을 실시하였다. 이 시뮬레이션 결과를 산포도, 등고선도를 이용하여 Vital Few별 최적 값의 존재가능 영역 및 분포상태를 확인하여 각 인자별 탐색범위로 선정하고, 이것을 시뮬레이션 조건으로 설정하였다. 그 결과는 <표 9>와 같다.

다음으로, 위의 시뮬레이션 조건을 기반으로 Vital Few의 최적 구간을 도출하기 위하여, IGN(Interactive Grouping Node) 알고리즘을 이용하여 Vital Few에 대한 구간화를 실시하였다. 본 연구에서는 83% 이상의 열효율에 대한 Vital Few의 최적 구간이 도출되었고, 그 결과는 아래 <표 10>과 같이 도출되었다.

<표 9> 시뮬레이션을 위한 조건

Vital Few	범위	간격
mgas_cal	930 ~ 1,000Kcal/N㎥	10Kcal/N㎥
cog_p	3.0 ~ 7.0%	0.1%
a_g	1.0 ~ 1.2%	0.05%
mgas_w	3,300,000 ~ 3,800,000N㎥/D	1,000N㎥/D

<표 10> IGN에 의한 Vital Few의 최적 구간

구분	데이터 마이닝 결과	실제값 ('08.05.15)
mgas_cal (Kcal/Nm³)	976 이하	930
cog_p(%)	6.0 이하	5.4
a_g(%)	1.0	1.08
mgas_w (Nm³/D)	3,668,825 이하	3,634,000

(3) 전체 최적화

세 번째 단계에서는, 이전의 단계에서 도출된 최적 모델과 Vital Few의 최적 구간을 이용한 전체 최적화를 수행하여 최적 공정조건을 도출한다. 이전 단계에서 수행한 데이터 마이닝 작업으로부터 도출된 최적 구간은 <표 10>과 같이 범위가 넓다. 그래서 실제 열풍로 공정에 즉시 적용하는데 필요한 최적 구간을 찾기 어렵다는 문제가 있기 때문이다. 이러한 활동을 수행하기 위하여, 본 단계에서는 Microsoft Excel 기반의 시뮬레이션 도구인 Crystal Ball을 이용한 시뮬레이션을 수행한다. 그리고 시뮬레이션의 결과를 토대로 MINITAB을 이용한 RSM(Response Surface Methodology) 분석을 실시하여 최적 공정조건을 도출하고자 한다.

시뮬레이션을 위하여, 우선 Crystal Ball을 이용한 시뮬레이터를 구축한다. 시뮬레이터를 구축하는 목적은 열풍로의 열효율을 예측하기 위함이다. 시뮬레이션 기법으로는 불확실한 상황 하에서의 의사 결정을 목적으로 하는 Monte Carlo 시뮬레이션 기법을 적용하였다. 본 시뮬레이션을 위하여 다음의 다변량 회귀모형이 설정되었다..

열효율(Heat_p) = 1.2364 +0.000000021*mgas_w -0.0006*mgas_cal - 0.01298*cog_p+0.1857*a_g

위 식은 Vital Few들이 열효율을 가장 잘 설명하는 회귀식으로, 변수들 간의 다중공선성 분석 결과 분산팽창계수(VIF) < 10 이하로 다중공선성에는 문제가 없다고 판단되었다.

여기서, 시뮬레이션을 위한 데이터는 데이터 마이닝에 이용한 데이터를 이용하였다.

다음으로, MINITAB을 이용하여 RSM 분석을 위한 실험 테이블을 생성한다. 실험에 사용될 인자는 Vital Few이고, 실험에서 이용되는 각 인자의 수준은 <표 11>에 주어진 것과 같다.

<표 11> 실험에서 이용되는 각 인자의 수준

인자	최저	최대
mgas_cal(Kcal/Nm³)	930	1,000
cog_p(%)	3.0	7.0
a_g(%)	1.0	1.2
mgas_w(Nm³/D)	3,300,000	3,800,000

<표 11>을 기초로 생성된 실험 테이블은 모두 32개의 실험 조건을 가진다. 각각의 실험 조건들은 Monte Carlo 시뮬레이터로 Random Number를 생성하기 위한 Seed 값으로 이용된다. <그림 12>은 RSM 분석을 위한 실험 테이블을 보여준다.

Worksheet 3 ***

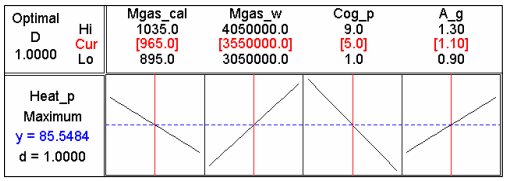
	C1	C2	C3	C4	C5	C6	C7
	StdOrder	RunOrder	Blocks	Mgas_cal	Mgas_w	Cog_p	A_g
1	1	1	1	930	3300000	3	1.0
2	6	2	1	1000	3300000	7	1.0
3	17	3	1	895	3550000	5	1.1
4	21	4	1	965	3550000	1	1.1
5	22	5	1	965	3550000	9	1.1
6	18	6	1	1035	3550000	5	1.1
7	29	7	1	965	3550000	5	1.1
8	25	8	1	965	3550000	5	1.1
9	11	9	1	930	3800000	3	1.2
10	12	10	1	1000	3800000	3	1.2
11	2	11	1	1000	3300000	3	1.0
12	15	12	1	930	3800000	7	1.2
13	20	13	1	965	4050000	5	1.1
14	7	14	1	930	3800000	7	1.0
15	28	15	1	965	3550000	5	1.1
16	8	16	1	1000	3800000	7	1.0
17	13	17	1	930	3300000	7	1.2
18	26	18	1	965	3550000	5	1.1
19	4	19	1	1000	3800000	3	1.0
20	9	20	1	930	3300000	3	1.2
21	5	21	1	930	3300000	7	1.0
22	31	22	1	965	3550000	5	1.1
23	19	23	1	965	3050000	5	1.1
24	16	24	1	1000	3800000	7	1.2
25	30	25	1	965	3550000	5	1.1
26	10	26	1	1000	3300000	3	1.2

<그림 12> RSM 분석을 위한 실험 테이블

세 번째로, 구축된 시뮬레이터를 이용한 시뮬레이션과 RSM 분석을 수행한다. 시뮬레이션은 MINITAB으로 생성한 실험 테이블에 있는 실험 조건들을 이용하여 수행되고, 각각의 실험 조건당 시뮬레이션 횟수는 10,000번이다. 시뮬레이션을 통하여 예측된 열효율은 RSM 분석을 위한 실험 테이블에 입력한다. RSM 분석을 실시한 결과, Vital Few와 열효율에 대한 최적 공정조건이 아래와 같이 도출되었다.

- mgas_cal: 965Kcal/N㎥
- cog_p: 5.0%
- a_g: 1.1%
- mgas_w: 3,550,000N㎥/D
- 열효율: 85.55%

<그림 13>는 RSM 분석의 결과를 보여준다.

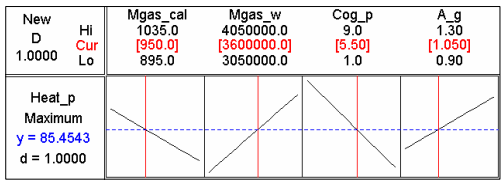


<그림 13> RSM 분석의 결과

열풍로의 현재 조업 조건을 고려하는 경우에는 Vital Few와 열효율에 대한 최적 공정 조건이 아래와 같이 도출되었다.

- mgas_cal: 950Kcal/N㎥
- cog_p: 5.5%
- a_g: 1.1%
- mgas_w: 3,600,000N㎥/D
- 열효율: 85.45%

<그림 14>은 열풍로의 현재 조업 조건을 고려한 RSM 분석의 결과를 보여준다.

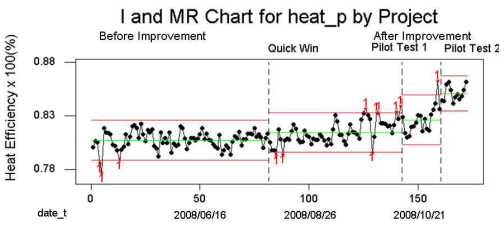


<그림 14> 현재 조업 조건을 고려한 RSM 분석의 결과

마지막으로, 시뮬레이션과 RSM 분석으로부터 도출된 최적 공정조건의 신뢰성을 확인하기 위하여, 83%의 열효율을 목표로 2차례의 Pilot test를 수행하였다. Pilot Test 결과, 4개의 Vital Few의 최적 조건이 아래와 같이 도출되었다.

- mgas_cal: 970Kcal/N㎥
- cog_p: 5.0~5.6%
- a_g: 1.07%
- mgas_w: 3,600,000N㎥/D

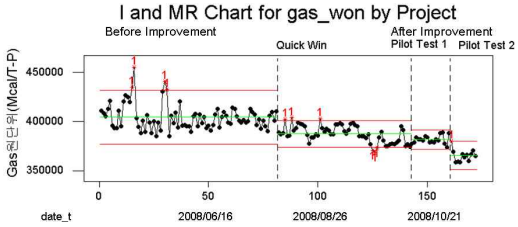
두 차례의 Pilot Test의 결과로부터, 열효율의 향상뿐만 아니라 에너지 소비량의 감소도 확인할 수 있었다. <그림 15>를 살펴보면, 개선 전의 열효율은 80.7%였지만, 첫 번째 Pilot Test에서는 82.6%로 향상되었고, 두 번째 Pilot Test에서는 85.1%로 향상되었음을 알 수 있다.



<그림 15> 열효율에 대한 IMR 차트

<그림 16>에서는, 개선 전의 에너지 소비량은 404.3Mcal/T-P였으나 첫 번째 Pilot Test에서는 381.4Mcal/T-P로 감소하였고, 두 번째 Pilot Test에서는 365.1Mcal/T-P로 감소하였음을 알

수 있다. Ton-Pig 당 Mix Gas 사용량은 용선 1톤을 생산하기 위해 요구되는 에너지를 말한다.



<그림 16> 에너지 소비량에 대한 IMR 차트

3.5 Control 단계

개선 결과가 지속되도록 이를 표준화, 시스템화 하고, 과제 수행에서 배운 내용을 공유한다.

3.6 기대효과

본 논문에서 제시한 데이터 마이닝 기반의 6 시그마 방법론을 본 프로젝트에 적용한 결과, 열풍로의 열효율이 기존의 80.2%에서 83.9%로 3.7%가 증가되었고, 열풍로 공정의 시그마 수준 역시 기존의 2.53σ에서 4.89σ로 2.37σ 수준이 증가하였음을 확인할 수 있었다. 이를 재무적 성과로 환산하여 보면, 연간 1,396,000,000원이 절감됨을 확인할 수 있었다. <표 12>은 본 방법론을 적용한 6 시그마 프로젝트의 개선 결과를 보여주고 있고, <표 13>는 재무적 성과를 보여주고 있다.

<표 12> 6 시그마 프로젝트의 개선 결과

항목	기존(A)	목표	실적(B)	증감(B-A)
열풍로 열효율	80.2	81.4	83.9	3.7
시그마 수준	2.52	3.40	4.89	2.37

<표 13> 6 시그마 프로젝트의 재무적 성과

항목	내용
재무효과	1,396,000,000원/년
산출근거	열풍로 열효율 향상으로 인한 에너지 사용량 절감
산출식	{(개선전Gas원단위-개선후Gas원단위) ×용선생산량}×Mixed Gas 단가 = {(430.5-416.5Mcal/T-P)}×8,587Ton/D ×32.46원/Mcal×365일×0.98 = 1,396백만원/년

Ⅳ. 결론 및 향후 연구과제

본 논문에서는 데이터 마이닝 기반의 6 시그마 추진 방법론을 제시하였고, 제시된 방법론은 제철공정의 열풍로 시스템의 열효율을 향상시키기 위한 프로젝트 수행에 적용되었다. 그 결과, 6 시그마 프로젝트에 데이터 마이닝을 적용함으로써 얻을 수 있는 이점으로서는 첫째, 데이터 마이닝이 6 시그마 프로젝트 수행 시에 하나의 도구 또는 지원 시스템으로써 프로젝트 성과를 향상시키며, 둘째, 6 시그마 프로젝트 수행 시 대규모 데이터를 분석해야 할 경우에 정확한 결과들을 도출함으로써 프로젝트 수행 기간을 단축시키며, 셋째, 데이터 마이닝이 6 시그마 프로젝트의 Analyze 단계와 Improve 단계에 유용하게 적용될 수 있음을 확인할 수 있었다. 특히, 방대한 데이터 및 공정들 간에 강한 상호작용 및 비선형 관계와 같은 특성을 갖는 연속 제조공정들에서, 다양한 비선형 기법들을 지원하는 데이터 마이닝은 Vital Few 및 적용 가능한 최적 운용조건들을 정확하게 도출하는데 적합함을 확인하였다. 또한 본 프로젝트에서는 열풍로 시스템의 최적 운용조건을 도출하고 구현하기 위하여,

인공신경망 기반의 feed-forward guidance simulator 및 데이터 마이닝을 이용한 multi-stage combustion pattern control system을 개발하였다. 이러한 시스템들의 도입으로 인해서 연구대상 제철회사에서는 최소의 운용비용을 달성하였고, 15%의 열풍로 에너지 비용을 절감할 수 있었다.

향후 연구방향으로는 6 시그마 프로젝트가 기업의 경영전략과 일치하는 방향으로 추진되어야 하기 때문에, 이를 위한 균형성과지표(BSC: Balanced Scorecard)를 6 시그마 프로젝트에 도입할 필요가 있고, 그리고 6 시그마 프로젝트 수행 결과가 기업의 업무 프로세스에 적용되어 수행되고, 또한 6 시그마 자체가 Define - Measure - Analyze - Improve - Control 프로세스를 따라 추진되기 때문에, 이를 지원하고 운영하기 위한 비즈니스 프로세스 관리(BPM: Business Process Management) 시스템의 도입이 필요할 것으로 사료된다(신정범 등, 2009; 김재전 등, 2008).

참고문헌

김재전, 노희옥, 박재성, 김상민, 유일, “BSC 방법론을 이용한 광가입자망사업 성과분석 연구,” 정보시스템연구, 제17권, 제4호, 2008, pp.79-98.

김형욱, 김종안, “식스시그마 추진기법 활용사례 연구,” 품질혁신, 제1권, 제2호, 2000, pp. 80-91.

나수천, 6시그마 國富論, 길벗, 2005.

다카나시 토모히로, 만넨 이사오, 프로세스 매니지먼트, 길벗, 2004.

민광기, 최대화, 한종훈, 장근수, “특성 추출과 신경회로망을 이용한 열풍로 열효율에 대한 모델링,” 한국가스학회지, 제2권, 제4호, 1998, pp.60-66.

박재홍, 변재현, 김창현, 정창원, 최영대, “구간세분화 방법을 이용한 철강산업체의 6시그마 프로젝트 추진사례,” 품질혁신, 제2권, 제1호, 2001, pp.57-65.

배영일, “6 시그마 경영의 이해와 실천,” CEO Information, 제349호, 2002, pp. 1-25.

서영주, 함효준, “조립라인에서의 6시그마 구축에 관한 연구,” 대한설비관리학회지, 제6권, 제1호, 2001, pp.91-100.

서정훈, 경영전략부문의 6 시그마 사용 사례 연구, 울산대학교 경영대학원 석사학위논문, 2005.

신정범, 김재균, 장길상, “SOA 및 BPM 기반의 정보시스템 구축 방법론: 고객지향 수주생산 환경에서의 제품 BOM 관리 적용사례,” 정보시스템연구, 제18권, 제1호, 2009, pp.77-95.

안진석, 고용민, 장중순, “데이터 마이닝을 이용한 최적 공정조건 탐색,” 대한설비관리학회지, 제4권, 제2호, 1999, pp.129-144.

지원철, 김우주, “품질설계시뮬레이션 지원시스템의 설계 및 구현,” 한국경영학회 추계학술대회 논문집, 1998, pp.385-388.

홍태호, 김진완, “데이터 마이닝의 비대칭 오류비용을 이용한 지능형 침입탐지시스템 개발,” 정보시스템연구, 제15권, 제4호, 2006, pp.211-224.

Antony, J. and Banuelas, R., "Key Ingredients for

the Effective Implementation of Six Sigma Program," *Measuring Business Excellence*, Vol.6, Issue 4, 2002, pp.20-27.

Fayyad, U. M. and Simoudis, E., "Knowledge Discovery and Data Mining: Tutorial," *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal, Canada, 1995.

Hahn, G. H., Hill, W., Hoerl, R. W., and Zinkgraf, S. A., "The Impact of Six Sigma Improvement: A Glimpse into the Future of Statistic," *The American Statistician,* Vol.53, No.3, 1999, pp. 208-215.

Hammer, M., "Process Management and the Future of Six Sigma," *MIT Sloan Management Review , Winter*, 2002, pp. 26-32.

Hancock, W., Yoon, J., and Plont, R., "Use of Ridge Regression in the Improved Control of Casting Process," *Quality Engineering*, Vol.8, No.3, 1996, pp.395-403.

Harry, M. J., The Vision of Six Sigma: *A Roadmap for Breakthrough*, Sigma Publishing Company, 1994.

Hoerl, R.W. and Snee, R.D., *Statistical Thinking: Improving Business Performance*. Duxbury Press/Thompson Learning, San Jose, 2002.

Jang, G. S., Jeon, J. H., and Lee, D. H., "Development of Multi-Stage Combustion Pattern Control System in Hot Stove Using Data Mining," *Proceedings of the 9th World Multi-Conference on Systemics,* Cybernetics and Informatics (WMSCI 2005), Orlando, Florida, USA, July 10 - 13, 2005, pp.136-140.

Jang, G. S. and Jeon, J. H., "A Six Sigma Methodology Using Data Mining: A Case Study on Six Sigma Project for Improving Heat Efficiency of a Hot Stove System in a Korean Steel Manufacturing Company," *Communications in Computer and Information Science 35*, 2009, pp.72-80.

Kwak, Y. H. and Anbari, F., "Benefits, Obstacles, and Future of Six Sigma Approach," *Technovation,* Vol.26, No.5-6, 2006, pp.708-715.

Montgomery, D.C., *Introduction to Statistical Quality Control*, 4th Edition, Wiley, New York, NY, 2001.

Pande, P., Neuman R., and Cavanaugh, R., *The Six Sigma Way*, McGraw-Hill, 2000.

Pyzdek, P., *The Six Sigma Handbook*, McGraw-Hill, 2003.

Su, D., Chen, W., Feng, J., Wang, H., and Su, Y., "Data Mining in Metallurgical Industry Process," *Computer Aided Chemical Engineering*, Vol.15, 2003, pp.1364-1369.

Tong, J., Tsung, F., and Yen, B., "A DMAIC Approach to Printed Circuit Board Quality Improvement," *International Journal of Advanced Manufacturing Technology*, 2004, pp.523-531.

Yachao W., "Data Mining from Simulation of Six Sigma in Manufacturing Company," *2008 International Conference on Computer Science and Software Engineering*, 2008, pp.423-426.

장길상(Jang, Gil-Sang)

저자는 울산대학교 산업공학과를 졸업하고, 한국과학기술원(KAIST)에서 산업공학 석사와 경영정보공학 박사를 취득하였다. 또한 한국국방연구원(KIDA) 선임연구원, 한국오라클 기술지원팀장, 동국대학교 경주캠퍼스 정보경영학전공 조교수를 거쳐, 현재 울산대학교 경영정보학과 교수로 재직중이다. 주요 관심분야로 생산정보시스템, 사례기반추론 시스템, DB 응용, ERP, e-Business 시스템, 객체지향 개발 방법론, 6 시그마 등이다.

<Abstract>

A Six Sigma Methodology Using Data Mining : A Case Study of "P" Steel Manufacturing Company

Jang, Gil-Sang

Recently, six sigma has been widely adopted in a variety of industries as a disciplined, data-driven problem solving approach or methodology supported by a handful of powerful statistical tools in order to reduce variation through continuous process improvement. Also, data mining has been widely used to discover unknown knowledge from a large volume of data using various modeling techniques such as neural network, decision tree, regression analysis, etc. This paper proposes a six sigma methodology based on data mining for effectively and efficiently processing massive data in driving six sigma projects. The proposed methodology is applied in the hot stove system which is a major energy-consuming process in a "P" steel company for improvement of heat efficiency through reduction of energy consumption. The results show optimal operation conditions and reduction of the hot stove energy cost by 15%.

Keywords : Six Sigma, Data Mining, Steel Manufacturing Company, Process Improvement, Hot Stove.

* 이 논문은 2011년 1월 4일 접수되어 1차수정(2011년 2월 28일)과 2차수정(2011년 6월 23일)을 거쳐 2011년 6월 28일 게재 확정되었습니다.