데이터마이닝 기법의 생산공정데이터에의 적용

이형욱(한국생산기술연구원), 이근안(한국생산기술연구원), 최석우(한국생산기술연구원),
배기웅(한밭대학교 산업경영공학과), 배성민(한밭대학교 산업경영공학과)*

Analyzing Production Data using Data Mining Techniques

H. W. Lee (Digital Production Processing Team, KITECH), G. A. Lee (Digital Production Processing Team, KITECH), S. Choi (Digital Production Processing Team, KITECH),
K. W. Bae(Dept. of IME, HANBAT National Univ.), S. M. Bae (Dept. of IME, HANBAT National Univ.)

ABSTRACT

Many data mining techniques have been proved useful in revealing important patterns from large data sets. Especially, data mining techniques play an important role in a customer data analysis in a financial industry and an electronic commerce. Also, there are many data mining related research papers in a semiconductor industry and an automotive industry. In addition, data mining techniques are applied to the bioinformatics area. To satisfy customers' various requirements, each industry should develop new processes with more accurate production criteria. Also, they spend more money to guarantee their products' quality. In this manner, we apply data mining techniques to the production-related data such as a test data, a field claim data, and POP (point of production) data in the automotive parts industry. Data collection and transformation techniques should be applied to enhance the analysis results. Also, we classify various types of manufacturing processes and proposed an analysis scheme according to the type of manufacturing process. As a result, we could find inter- or intra-process relationships and critical features to monitor the current status of the each process. Finally, it helps an industry to raise their profit and reduce their failure cost.

Key Words : Data Mining (데이터마이닝), Type of Manufacturing Process(생산공정형태), Inter-process relationship(공정간 관계)

1. Introduction

1990 년대 후반부터 폭발적으로 증가하기 시작한 인터넷의 사용으로 인하여, 인터넷을 통해 생산되는 데이터의 양은 기하급수적으로 늘어나게 되었다. 또, 기업들은 인터넷의 폭발적인 사용이전부터 고객과의 긴밀한 관계 유지를 위해 고객데이터들을 수집하고, 분석해 왔기 때문에 기업들이 관리하고, 분석해야 하는 데이터의 양 또한 급격하게 증가하게 되었다.

이러한 현상은 생산현장에서도 마찬가지여서, 전세계의 기업들과 가격, 품질, 스피드 측면에서 경쟁을 하게 되었으며, 가격, 품질 측면에서의 경쟁력을 갖추기 위해 전사적 자원관리(enterprise resource planning, ERP), 생산설비 자동화(process automation) 등의 시스템과 설비를 갖추게 되었다.

이를 통해 생산현장에서도 많은 데이터들이 생성되게 되었으며, 이를 관리하기 위한 통계적 품질관리(statistical quality control, SQC), 식스 시그마 (six sigma) 등의 여러 관리 기법들이 적용되고 있는 것이 현실이다.

이러한 기법들로 인해 생산 현장에서 발생하고 있는 여러 상황에 대한 실시간 모니터링이 가능해진 것이 사실이나, 데이터들의 깊이 있는 분석을 통해 데이터들로부터 숨겨진 사실을 발견하기 위한 시도를 하는 수준까지는 이르지 못했던 것도 사실이다.

본 논문에서는 생산현장에서 생성되는 데이터를 데이터마이닝 기법을 이용해서 분석하는 방법에 대해서 설명하고자 한다. 데이터마이닝 기법은 대용량의 데이터로부터 의미 있는 정보를 추출하는데 유용하다고 알려져 있으며, 금융, 보험 등의 서비스업종에서 고객분석을 위해 많이 사용되어 왔으며, 생산현장에서 생성되는 공정 데이터 분석에도 많이 적용되어왔다. 또, 최근 들어 각광받고 있는 생물정보학(bioinformatics) 영역에서도 활발하게 이용되고 있는 기법이다. 본 논문에서는 많이 사용되고 있는 데이터마이닝 기법에 대해 간단히 설명하고, 이것이 라인생산방식을 따르는 생산현장에 적용되었을 때, 어떻게 이용이 되어야 할 것인지에 대해서 자세히 설명하도록 한다.

2. Literature Review

2.1 제조업에서의 생산형태

제조업은 여러 가지 생산형태를 가지게 된다. 특히, 대표적인 생산형태는 3 가지로 구분이 되는데, 흐름(flow)생산, 라인(line)생산, 잡샵(jobshop)생산방식이 그것이다.

흐름생산은 연속적인 공정을 가지는 장치산업으로 철강, 반도체, 화학분야가 해당된다. 특히, 반도체 분야의 공정은 수많은 공정과 테스트작업이 연속적으로 이루어지기 때문에 데이터마이닝을 적용한 사례가 많다. [1] [2]

라인생산방식은 가장 일반적인 제조업에서의 생산공정으로써 자동차부품, 기계부품 등의 가장 광범위한 영역에 적용된다.

잡샵형식의 생산은 단품이나 단량위주의 생산방식으로 금형, 중공업, 조선 등의 특화된 제품에 적용되는 방식이다.

2.2 데이터마이닝 기법

2.2.1 Classification

Classification 이란 클래스(class)를 구분해주는 규칙(rule)에 의해서 새로운 자료가 미리 정의되어 있는 클래스 중 어디에 속하는 것인지를 판단해주는 것을 의미한다. 규칙을 생성하기 위해서는 클래스가 정의되어 있는 데이터가 필요하며 규칙을 만들기 위한 알고리즘(algorithm), 그리고 도출된 규칙을 어떤 형태로 보여줄 것인지에 대한 고려가 필요하다.

가장 일반적인 Classification 도구로는 인공지능분야에서의 의사결정나무(decision tree)와 통계학분야의 k-NN(k-nearest neighborhood) 방법이 있는데, 이는 각 데이터의 속성(*attribute*)에 관련된 질문을 함으로써 이에 대한 결과에 따라 어떤 클래스에 속하는 지를 결정해 주는 것을 의미한다.

의사결정나무를 생성해 주는 알고리즘 가운데 가장 널리 쓰이고 있는 것은 Quinlan 에 의해 개발된 C4.5 라는 프로그램이다. [3] C4.5 에는 사용자들의 편의를 위해 생성된 의사결정나무를 규칙으로 바꾸어주는 기능이 포함되어 있으며, 효율적인 분류(classification)를 위해서 *gainratio* 라는 새로운 구분기준을 이용한다. 현재 C4.5 는 여러 기능들을 보강한 C5.0 까지 개발되어 있으며, C5.0 은 C4.5 에 비해 다양한 형태의 데이터를 처리할 수 있으며 더 나은 classification 결과를 보여준다. [4]

2.2.2 Neural Network

신경망(neural network)은 인간 대뇌의 기본 단위인 뉴런의 생리학적 모델을 본 떠 만든 것으로, 합(sum)이나 로그(log)함수 같은 간단한 함수를 계산하는 노드(node)가 있고, 노드와 노드사이의 연결에 가중치(weight)가 있어 하나의 노드의 값이 다른 노드로 연결 가중치와 곱해져서 전달된다.

일반적으로 신경망에서는 역전파 알고리즘(back propagation algorithm)이 이용되는데, 이는 LMS(least mean square) 알고리즘을 일반화 한 것이다. 즉, 초기 파라메터(parameter)값을 가중치로 사용하여 반복적으로 학습함으로써 가중치 값을 조정한다. 많은 학습자료를 이용하여 일정한 값으로 수렴하는 특징을 지닌 점은 HMM(hidden markov model)과 비슷한 점이나, 신경망은 자동적으로 분류를 계속하면서 스스로 학습되어 나가는 것이 HMM 과 다른 점이라고 할 수 있다. 특히, 신경 회로망 내의 연결가중치 조절방법은 신경 회로망의 학습방법을 결정한다고 볼 수 있으며, 그 예로는 Hebbian Learning Rule, Delta Rule, Generalized Delta Rule 및 Kohonen Learning Rule 등이 있다.

2.2.3 Clustering

군집화(clustering)은 classification 과 다르게 데이터에 미리 정의된 클래스가 존재하지 않으며, 정해진 알고리즘에 따라 비슷한 성격을 지니는 데이터들끼리 모아주는 역할을 하게 된다.

가장 일반적으로 쓰이는 군집화 알고리즘에는 Self-Organizing Map 또는 코호넨맵으로 알려진 SOM 이라는 것이 있다.

SOM 에서는 뉴런이 n 차원으로 구성된 격자의 노드에 위치하게 되며, 각각의 뉴런은 경쟁학습관계에 있는 입력 패턴이나 입력 패턴의 클래스에 따라 선택적으로 구성된다. 즉, 격자에 있는 뉴런의

공간의 위치가 입력패턴의 고유 특성과 일치하도록 입력 패턴의 자형적인 지도를 형성하게 된다.

3. 데이터마이닝의 생산공정에의 적용

3.1 생산 공정 정형화

공정데이터를 분석하기 위해서는 분석하고자 하는 공정(process)을 정확히 파악하고, 각 공정에서 생성되는 데이터와 그에 대한 판정기준(criteria), 그리고 공정간의 상관관계에 대한 정보를 수집하는 것이 가장 첫 번째 단계이다.

생산공정에 대한 상세한 이해는 데이터 분석 및 해석에 많은 도움을 줄 수 있다. 예를 들어, 현재 수집되고 있는 데이터들 가운데 어떤 데이터들을 추출(extraction)해야 할 것이며, 변환(transformation)과정에서 정규화(normalization)는 어떻게 해야 할 것인지, 또 분석 결과들에 대한 해석(interpretation)에 도움을 줄 수 있다.

이러한 과정은 실제 현장에서 측정하고 있는 데이터들이 분석에 사용될 수 있는지를 파악하는데 도움을 줄 수 있으며, 만일 필요한 데이터들이 수집되고 있지 않다면 추후 개선을 통해 데이터들의 수집을 제안할 수 있다. 또, 데이터의 수집주기, 데이터의 품질에 대한 여러 가지 개선사항들을 도출하는데 도움을 줄 수 있으며, 분석 후 유용한 데이터로 검증이 된 데이터들은 추후 공정분석을 위한 데이터마트(data mart)를 구축하는데 포함시킬 수도 있다.

이를 위해서는 ISO 9000 또는 14000 등의 품질 인증을 받은 업체의 경우에는 인증에 필요한 문서들과 공정 관리자, 시스템 관리자들과의 인터뷰를 통해 필요한 사항들을 파악하는 것이 필요하다.

또, 모든 공정을 대상으로 하여 분석하기는 매우 시간이 많이 걸리기 때문에, 드릴다운(drill down) 방식으로 분석하기 위해서 가장 문제가 많이 발생하고 있는 공정과 그와 관련된 공정을 파악하는 것이 매우 중요하다.

3.2 Clustering 을 이용한 분석방법

공정에서 측정해야 되는 데이터의 종류는 매우 많다. 예를 들어 반도체 공정의 경우, 한 공정에서 테스트를 위해 측정하는 변수의 개수는 600 여개가 넘는 경우도 있다.

이런 경우 모든 변수에 대한 분석을 하기가 어렵기 때문에 비슷한 경향을 지닌 변수들을 그룹화하여 해당 그룹에서 가장 대표성을 가지는 상위 3~5 개의 변수들을 도출하여 이에 대한 분석을 하는 것이 유리하다. 이를 Feature construction 이라고 하는데. 이 때 유용하게 쓰일 수 있는 방법이 군집(clustering)분석이다.

군집분석을 통해 각 공정에서 분석의 대상이 될 수 있는 변수의 숫자를 획기적으로 줄일 수 있으며 이를 통해 좀 더 세밀한 분석을 수행할 수 있게 된다.

3.3 C4.5 를 이용한 분석방법

앞서 소개한 데이터마이닝 기법 가운데, 공정분석에서 가장 유용하게 쓰일 수 있는 기법 가운데 하나가 의사결정나무를 이용한 방법이다.

일반적으로 각 공정에서 측정되는 데이터의 항목은 그 수가 매우 많으며, 측정결과 또한 '양호' 또는 '불량' 의 명확히 구분된 2 개의 클래스를 결과값으로 가지게 된다. 이러한 데이터의 특성상 의사결정나무, 특히 C4.5 를 이용한 공정 데이터의 분석은 각 공정에서 양호와 불량을 구분하는 가장 중요한 특성이 무엇인지 도출해 낼 수 있다.

즉, C4.5 의 수행결과로 생성된 의사결정나무에서 가장 상위 루트 노드(root node)에 나타난 변수가 양호와 불량을 구분하는데 가장 중요한 역할을 하는 변수가 된다.

이러한 관점에서 의사결정나무에서 루트 노드를 포함한 상위 3 개~5 개 정도의 변수들이 집중적으로 관리해야 할 필요가 있는 중요 변수들이 될 수 있다.

또한, 의사결정나무를 만들기 위한 데이터들을 성능테스트에 관련된 것으로 제한 하면 수많은 성능테스트 항목 가운데 꼭 해야만 하는 항목과 하지 않아도 되는 항목을 구분할 수 있다. 이러한 분석을 통해서 어떠한 테스트가 최종 성능에 영향을 미치는지를 파악함으로써 테스트에 걸리는 시간과 비용을 줄이는데 도움을 줄 수 있다.

3.4 Neural Network 을 이용한 분석방법

신경망은 주로 에러율(error rate)에 대한 예측(forecasting)에 주로 사용된다. 각 공정에서 중요한 변수들을 도출하고 이러한 변수들이 최종 수율(yield) 에 어떠한 영향을 미치는지를 파악할 수 있으며 이에 대한 학습을 통해 새로운 환경으로 바뀌었을 때 또는 새로운 변수가 추가되었을 때 최종 수율이 어떻게 바뀔 것인지에 대해 예측을 하는데 도움을 줄 수 있다.

신경망을 이용한 예측에서는 어떤 변수를 사용하여 예측하는지 얼마나 신뢰성이 있는 데이터들이 사용되는지에 대한 사항들이 결과에 큰 영향을 미치기 앞부분에서 적절한 변수를 추출하는 것이 매우 중요하다.

4. 결론 및 추후 연구과제

생산공정에서 생성되는 데이터의 양은 매우 방대하다. 그러나 지금까지 이러한 데이터들을 단순히 경향을 파악하거나 또는 현황을 파악하는데 주로 사용되어 왔고, 데이터 속에 숨어있는 사실들을 파악하기 위한 용도로는 거의 사용되지 않았던 것이 사실이다.

이를 해결하기 위한 방법 가운데 하나인 데이터마이닝 기법은 금융, 보험, 통신영역에서 유용하다는 것이 검증되어 왔음에도 불구하고, 생산분야에서는 주로 반도체산업에서만 제한적으로 사용되어 왔던 것이 사실이다.

본 논문에서는 여러 가지 데이터마이닝 기법들을 소개하고 이들이 실제 생산공정의 분석에 사용되기 위해서는 어떻게 해야 할 것인지에 대해 논의하였다.

앞으로의 연구과제는 이러한 분석방법론을 우리나라의 대표산업 가운데 하나인 자동차 부품업계에 적용시켜 봄으로써 품질 및 성능개선에 유용한 정보들을 도출하고 이를 현장에 적용함으로써 도출된 정보에 대한 평가를 해보고자 한다.

후 기

본 연구는 산업자원부의 중기거점 개발사업인 "웹기반 SMART 제조시스템 개발" 과제의 지원으로 수행되었으며, 이에 도움을 주신 관계자 여러분들께 감사 드립니다.

참고문헌

1. 백동현, 한창희, "데이터마이닝을 이용한 반도체 FAB 공정의 수율 개선 및 예측," 한국지능정보시스템학회지, 제 9 권, 제 1 호, pp. 157-177, 2003
2. J. H. Lee, S. J. Yu, and S. C. Park, "Design of Intelligent Data Sampling Methodology based on Data Mining Technology," IEEE Transaction on Robotics and Automation, Vol. 17, No. 5, pp. 637-649, 2001
3. J. Ross Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publisher, 1993
4. http://www.rulequest.com