데이터마이닝을 위한 공정데이터 품질개선

Enhancing Manufacturing Data Quality for Data Mining

*,#배성민[1], 이형욱[2], 이근안[2], 최석우[2], 박홍균[3]

*,#S. M. Bae(loveiris@hanbat.ac.kr)[1], H.W. Lee[2], G.A. Lee[2], S. Choi[2], H.K. Park[3]

[1] 한밭대학교 산업경영공학과, [2]한국생산기술연구원 디지털성형공정팀, [3](주)스페이스솔루션

Key words : Data Quality, Data Mining, Manufacturing Data

1. 서론

오늘날 기업이 더 많은 이윤을 창출하기 위하여 생산성을 높이고 제조원가를 낮추며 기업이 제공하는 제품과 서비스에 대한 고객의 경험과 만족도를 높이고자 하는 노력은 예전과는 달리 기업이 대상으로 하고 있는 고객의 범위가 전 세계적으로 넓어졌기 때문이다. 또한 기업의 경쟁 대상이 같은 국가 내의 기업뿐만 아니라 전 세계의 기업과의 경쟁으로 바뀌었고 이는 기업의 모든 업무 프로세스에서의 개선과 혁신을 추구하도록 바꾸었으며 기업에서 생성되는 모든 데이터들의 분석을 통해 현실을 직시하고 잠재적인 문제점을 미리 파악하고 이에 대비할 수 있는 체제를 구축하는데 큰 영향을 주었다.

이러한 관점에서 제조업에서는 예전에는 중요하게 여겨지지 않았거나 제한적인 목적으로 수집, 분석되던 데이터들이 최근 들어서는 데이터의 수집 범위가 넓어지고 대부분의 생산 공정에서 생성되는 많은 양의 데이터들이 다양한 시스템을 통해 수집되고 저장됨으로써 이를 분석하고자 하는 시도가 더 활발해지고 있다.

이렇게 수집된 데이터들은 기존에 통계적 품질관리(statistical quality control, SQC), 식스 시그마(six sigma), 전사적 품질관리(total quality management, TQM) 등의 여러 관리기법을 통해 체계적으로 관리되어 왔으며 이를 통해 관리상한(control limit) 등의 여러 지표들을 제시하여 생산현장에서의 관리 등에 이용되고 있다. 이러한 지표들은 생산 현장에서 발생하고 있는 여러 상황에 대해 실시간 모니터링이 가능하게 하는 장점을 가지고 있으나 데이터들의 깊이 있는 분석을 통해 지금까지 발견되지 않았던 유용한 지식을 도출해내 수준까지는 이르지 못했던 것이 사실이며 이러한 단점을 해결하기 위해 주로 서비스·금융업에서 이용되고 있던 데이터마이닝(data mining) 기법이 제조업에서도 이용되기 시작하였다.

데이터마이닝 기법은 대용량의 데이터로부터 의미 있는 정보를 추출하는데 유용하다고 알려져 있으며, 주로 금융·보험업 등의 서비스업종에서 고객 분석을 위해 많이 사용되어져 왔다. 그러나 이러한 데이터마이닝 기법으로부터 유용한 정보를 추출해 내기 위해서는 다양한 알고리즘(algorithm)의 적용에 선행하여 각각의 데이터마이닝 기법에 적절한 데이터를 확보하는 것이 더 중요한 문제이다.

본 논문에서는 생산 공정에서 적용될 수 있는 데이터마이닝 알고리즘에 대해 간단히 알아보고, 생산 공정에서 수집되는 데이터들의 일반적인 형태와 데이터들이 분석에 사용되기 위해서 갖추어야 할 전처리(pre-processing)과정에 대해서 알아보도록 한다. 생산 공정에서 생산되는 데이터들의 전처리를 통해 데이터의 품질(data quality)를 높이고 이를 분석에 사용함으로써 추후 도출될 지식들의 신뢰성을 높이는데 도움을 줄 수 있을 것이다.

2. 생산공정에 적용 가능한 데이터마이닝 기법

2.1 군집기법(clustering)의 적용

공정에서 측정해야 되는 데이터의 종류는 매우 많다. 예를 들어 반도체 공정의 경우, 한 공정에서 테스트를 위해 측정하는 변수의 개수는 600여개가 넘는 경우도 있다. 이런 경우 모든 변수에 대한 분석을 하기가 어렵기 때문에 비슷한 경향을 지닌 변수들을 그룹화하여 해당 그룹에서 가장 대표성을 가지는 상위 3~5개의 변수들을 도출하여 이에 대한 분석을 하는 것이 유리하며, 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.

2.2 분류기법(classification)의 적용

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

즉, C4.5의 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 3개~5개 정도의 변수들이 집중적으로 관리해야 할 필요가 있는 중요 변수들이 될 수 있다.

또한, 의사결정나무를 만들기 위한 데이터들을 성능테스트에 관련된 것으로 제한하면 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.

2.3 신경망(neural network)의 적용

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.

3. 데이터마이닝을 위한 공정데이터 품질요소

앞 절에서 소개한 다양한 데이터마이닝 기법을 적용하여 유용한 지식을 도출해내기 위해서는 분석에 사용되는 데이터들에 대한 품질(quality)에 대한 고려가 선행되어야 한다. 하지만 데이터 품질을 개선하기 위해서는 기업에서 운영되고 있는 공정과 각 공정에서 수집되는 데이터들의 성격(nature)에 대한 깊이 있는 이해가 선행되어야 하며 데이터들을 수집하기 위한 비용(cost)과 기업 내에서의 활용성에 대한 고려가 필수적인 요소이다. 아무리 좋은 데이터라고 할지라도 이를 수집하는데 많은 비용이 든다거나 분석에는 유용하나 기업에서 활용이 불가능한 데이터를 수집하는 것은 큰 의미가 없기 때문이다.

이러한 관점에서 데이터 환경(data environment)은 데이터를 수집(collecting), 저장(storing), 사용(using)하는데 관련된 모든 사항을 의미한다. 또, POP(point of production), 전사적 자원관리(enterprise resource management, ERP), 각종 데이터베이스로부터

수집되는 데이터들뿐만 아니라 이에 관련된 프로세스정보, 생산 규칙, 생산 방법 등에 대한 모든 사항을 포함하게 된다.

본 절에서는 생산현장에서의 데이터품질을 향상시키기 위하여 고려해야 할 7가지 요인에 대해 언급하기로 한다.

- 다양한 데이터 소스(multiple data source)

생산 현장에서 사용되는 시스템들은 서로 다른 목적으로 개발되어 서로 다른 사용자들이 사용하게 되는 것이 일반적이다. 이때, 수집되는 데이터들은 여러 시스템에서 다양한 이름(name)으로 저장되는데, 같은 정보를 가지고 있는 데이터들은 항상 같은 이름으로 저장되고 사용될 수 있도록 수정되는 것이 필요하다. 이를 해결함으로써 여러 부서의 사용자들은 분석 결과에 대해 토론을 하고 이를 현장에 적용가능하게 된다.

- 수기데이터 수집(manual-data collection)

모든 공정에서 자동화된 데이터수집이 가능한 것은 아니며 실제 현장에서는 작업자들이 직업 데이터를 취득, 입력하게 되는 과정을 거쳐 수집되는 데이터들이 상당 부분을 차지하고 있다. 이러한 과정에서는 데이터를 취득하고 입력하는 부분에 대한 규정(rule)를 정의해 줌으로써 작업자가 달라지는 경우에도 같은 품질의 데이터를 취득할 수 있게 되는 장점이 생긴다.

- 수집·저장되는 데이터의 양 (volume of data)

생산현장에서 수집되는 데이터들의 양은 제품 또는 공정에 따라 엄청난 양의 데이터를 생성시킨다. 이러한 데이터들이 매시간 생성되고 저장되면 분석에 걸리는 시간을 매우 증가시킬 뿐만 아니라 기업에서 가용한 컴퓨팅 자원을 낭비할 수도 있다. 그러므로 분석에 필요한 데이터의 양을 추정하여 이에 대한 데이터들의 수집을 계획하고, 저장하는 것이 필요하다.

- 효율적 메타데이터의 사용(efficient metadata encryption)

저장된 데이터들에 대한 부가적인 정보를 제공하는 메타데이터는 잘 사용된다면 데이터들에 대한 많은 정보를 추가적으로 사용자들에게 제공할 수 있다. 하지만, 이러한 메타데이터에 대한 정의 및 규칙이 전사적인 관점에서 결정되어야 한다. 특히 분석에 있어서 데이터 가공(data manipulation)에서 메타데이터 정보를 활용하는 경우가 많기 때문에 이러한 규칙들이 서로 혼동되어 진다면 잘못된 분석결과를 유발할 수 있게 된다.

- 입력 데이터의 형식 (input data form)

현장에서 수집되는 데이터들은 개발에의 편의성 또는 빠른 데이터의 저장을 위해 분석에 바로 사용되기에는 어려운 형태로 저장되어 있는 경우가 많다. 이러한 경우, 데이터마이닝 기법을 적용하기 위해서 데이터의 전처리가 매우 까다로운 경우가 대부분이다. 이러한 문제를 해결하기 위해서는 데이터 분석에 필요한 데이터 형태(form)를 미리 정의하고 이를 고려한 데이터스키마(data schema)를 설계 또는 변경하는 것이 필요하다. 특히 실시간 분석이 필요한 경우 데이터의 전처리과정은 분석과정 전체에서 큰 부분을 차지하는 경우가 많기 때문이다.

- 데이터 수요에 대한 변화 (changing data needs)

데이터마이닝의 결과로 도출된 정보로 인하여 또는 관리상의 목적이 변경됨으로써 기존에 관리하던 변수들이 바뀌게 되는 경우가 생기게 된다. 또는 새로운 관리변수에 대한 등장으로 인해 기존에 수집되던 데이터들의 형태가 변경되는 경우가 발생하게 된다. 이러한 경우 시스템에서 새로운 데이터의 수집 규칙을 빠르게 대응해 줌으로써 새롭게 변경된 데이터 수요에 빠르게 대처할 수 있게 된다.

- 분산 데이터 시스템 (distributed data system)

지역적으로 분산된 시스템에서 수집된 데이터들의 분석은 서로 다른 시스템에 저장되어 있는 데이터들의 통합과정을 거쳐야 한다. 이러한 경우 가장 빈번하게 발생하는 문제점은 서로 일치하지 않는 (inconsistent) 데이터에 대한 통합 문제이다. 서로 다른 시스템간의 데이터 통합은 시스템자체의 문제일 뿐만 아니라 조직 내에서 사용되는 데이터들의 전반적인 구조에 영향을 미칠 수 있다. 이를 해결하기 위한 방법으로는 전사적 차원의 데이터웨어하우스(data warehouse) 또는 SAN(storage area network)의 구축을 통해 전사적인 관점에서 통합된 데이터의 저장을 추진할 수 있다.

4. 결론

제조업에서 생산성 및 품질의 향상, 제조원가의 절감 등을 통한 이윤향상은 글로벌 시대에서의 경쟁에서 생존하기 위한 최소조건이 되었다. 이러한 조건을 만족시키기 위해서는 기존에 사용되던 관리기법뿐만 아니라 데이터마이닝 기법을 활용하여 기업 내에서 생성되고 저장되는 데이터들을 심도 있게 분석함으로써 이를 활용하고자 하는 노력이 필요하다.

이러한 관점에서 생산 공정에서의 데이터품질에 대한 고려는 분석결과의 신뢰성을 높이고 추후 지속적인 데이터분석에 활용될 수 있다는 점에서 큰 의미를 가진다. 다양한 데이터 소스에 대한 고려, 데이터의 양에 대한 고려, 메타데이터의 사용, 수요 데이터의 변화의 측면은 제조업뿐만 아니라 서비스업에도 적용될 수 있는 부분이며, 특히 수기데이터의 수집부분에 대한 요인은 현재 우리나라의 생산현장에서 적용될 수 있는 부분이기도 하다. 이러한 생산현장에서의 데이터품질을 향상시킴으로써 더 많은 데이터들이 효율적으로 수집, 저장되고 이를 통해 의미 있는 데이터 분석결과를 도출해 내는데 도움을 줄 수 있을 것이다.

후기

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

참고문헌

1. J. Ross Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, 1993
2. J.H. Lee, S.J. Yu, and S.C. Park, "Design of Intelligent Data Sampling Methodology based on Data Mining Technology," IEEE Trans. on Robotics and Automation, Vol. 17, No. 5, 637-649, 2001.
3. Carlo Batini and Monica Scannapieco, Data Quality: Concepts, Methodologies and Techniques, Springer, 2006.
4. Yang W. Lee, Leo L. Pipino, James D. Funk, and Richard Y. Wang, Journey to Data Quality, MIT press, 2006.