creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

공 학 석 사 학 위 논 문

공정개선을 위한 대용량 데이터 분석 및 데이터마이닝 적용사례 연구

*2007*년 *2*월

부 경 대 학 교 대 학 원

시스템경영공학과

권 유 진

공 학 석 사 학 위 논 문

공정개선을 위한 대용량 데이터 분석 및 데이터마이닝 적용사례 연구

지도교수 권혁무

이 논문을 공학석사 학위논문으로 제출함

*2007*년 *2*월

부 경 대 학 교 대 학 원

시스템경영공학과

권 유 진

권유진의 공학석사 학위논문을 인준함

*2006*년 *12*월 *22*일

주 심 공학박사 옥 영 석 인

위 원 공학박사 고 시 근 인

위 원 공학박사 김 영 진 인

목 차

제 1 장 서론 ····· 1
1.1 연구의 배경 및 목적 ····· 1
1.2 연구 내용 ····· 2

제 2 장 데이터마이닝 ····· 3
2.1 데이터마이닝의 배경 및 정의 ····· 3
2.2 데이터마이닝의 특징 ····· 4
2.3 데이터마이닝의 분석유형 및 기법 ····· 5
2.4 데이터마이닝 수행과정 ····· 8

제 3 장 데이터 분석사례 ····· 10
3.1 제품 소개 및 공정소개 ····· 10
3.2 Warp의 정의 및 불량원인 도출 ····· 12
3.3 통계 및 기술적 지식에 의한 분석 ····· 13
3.3.1 인자별 분석 결과 ····· 17
3.3.2 Slicing 공정의 작동 메커니즘 분석 ····· 19

제 4 장 데이터마이닝 기법의 적용사례 ····· 20
4.1 서론 ····· 20
4.2 군집분석 ····· 20
4.2.1 군집분석의 개념 ····· 20

4.2.2 K-mean 군집분석 ····· 22
4.2.3 군집분석의 장점 및 단점 ····· 24
4.2.4 군집분석의 적용 ····· 25
4.3 의사결정나무 분석 ····· 27
4.3.1 의사결정나무의 개념 ····· 27
4.3.2 의사결정나무의 형성과정 ····· 30
4.3.3 의사결정나무의 장점 및 단점 ····· 33
4.3.4 의사결정나무의 적용 ····· 34
4.4 결과 비교 ····· 36

제 5 장 결론 및 향후 연구과제 ····· 39

참고문헌 ····· 41
Abstract ····· 44

그림 목차

[그림 2.1] 데이터마이닝 기법 ····· 6
[그림 2.2] 데이터마이닝 수행과정 ····· 7
[그림 3.1] 실리콘 웨이퍼의 제조공정 ····· 11
[그림 3.2] Warp의 정의 및 형태 ····· 12
[그림 3.3] Multi wire saw 기계 ····· 13
[그림 3.4] 장비에 따른 Warp의 변화 ····· 15
[그림 3.5] Warp 데이터의 히스토그램 ····· 15
[그림 3.6] 공정 총 변동의 구성 ····· 16
[그림 3.7] Rod 위치별 Warp의 변화 ····· 18
[그림 3.8] 개선 후 Rod 위치별 Warp의 변화 ····· 19
[그림 4.1] K-mean 군집분석 ····· 23
[그림 4.2] 군집분석 결과····· 27
[그림 4.3] 의사결정나무의 예 ····· 29
[그림 4.4] 의사결정나무 형성과정 ····· 31
[그림 4.5] 의사결정나무분석의 Tree노드 결과 ····· 35

표 목차

[표 3.1] 장비별 Warp의 평균 및 표준편차 비교 ····· 14
[표 3.2] Rod의 패턴 ····· 18
[표 4.1] X요인의 패턴정리 ····· 27
[표 4.2] 의사결정나무 분석 결과 ····· 36
[표 4.3] 분석결과 비교 ····· 37

PUKYONG NATIONAL UNIVERSITY
부경대학교

제 *1* 장 서론

1.1 연구의 배경 및 목적

최근 컴퓨터 및 센서기술이 발달함에 따라 많은 데이터가 실시간으로 수집되는 공정이 늘어나고 있다. 공정을 개선하고자 할 경우, 이와 같은 공정에서 얻어진 대용량 데이터에서 필요한 정보를 얻을 수 있다면, 새로이 데이터를 수집하는 과정 없이도 데이터를 잘 활용하여 유용한 정보를 얻을 수 있을 것이다. 즉, DOE와 같은 실험을 통한 데이터 수집과정 없이 이미 수집되어 있는 데이터 속에서 유용한 정보를 도출함으로써 시간적인 감소뿐만이 아니라 대량의 데이터에 기초한 신뢰도 높은 결론을 도출할 수 있을 것이다.

그러나 대용량 데이터로부터 의미 있는 정보를 도출하는 것은 쉬운 일이 아니다. 일반적으로 많은 양의 데이터를 확보하고 있을 경우 검출력의 향상으로 공정의 미세한 변화에 보다 민감하게 탐지할 수 있을 것으로 생각이 된다. 하지만, 현실적으로 이러한 대용량의 데이터를 이용하여 의미 있는 정보를 얻고자 할 경우 불필요한 정보 혹은 많은 잡음 요인들의 개입으로 인하여 의미 있는 정보의 도출이 더 어려울 수도 있다. 따라서 기존의 방대한 데이터를 효과적으로 분석하기위해서는 단순히 통계적인 분석에 의한 방법뿐만이 아니라 공정에 대한 경험과 기술적인 지식을 활용하여 의미 있는 정보를 도출하는 방법이 제시될 수 있다. 즉, 공정에 대한 경험과 기술적인 지식을 활용하여 분석의 기본 방향을 설정하고 많은 양의 데이터를 체계적으로 분석한 후 분석 결과를 실질적인 측면에서 재검토하여 의미 있는 결과를 도출할 수 있다.

본 연구에서는 대용량 데이터에 대한 통계적인 분석과 함께 공정에

대한 기술적인 지식을 활용하여 공정에 대한 의미 있는 정보를 도출하는 과정을 소개하며, 또한 데이터마이닝 기법 중 군집분석과 의사결정나무를 이용하여 그 결과를 비교하고자 한다.

1.2 연구 내용

본 논문은 공정개선을 위한 대용량 데이터 분석사례로 통계적인 분석뿐만이 아니라 공정에 대한 기술적인 지식을 활용하여 분석한 사례와 데이터마이닝을 이용한 분석을 비교하고자 한다.

본 논문은 다음과 같이 구성되었다.

2장에서는 대용량 데이터를 분석하는 하나의 기법인 데이터마이닝에 대한 개념을 소개한다.

3장과 4장에서는 실제 대용량 데이터를 분석한 사례를 소개한다. 3장에서는 대용량 데이터에 대한 통계적인 분석과 함께 공정에 대한 기술적인 지식을 활용하여 공정에 대한 의미 있는 정보를 도출하는 과정을 소개하였으며, 4장에서는 데이터마이닝 기법 중 군집분석과 의사결정나무에 대한 소개와 함께 실제 적용한 사례를 소개한다.

마지막으로 5장에서는 결론 및 향후 연구과제에 대해 논의한다.

제 2 장 데이터마이닝

2.1 데이터마이닝의 배경 및 정의

컴퓨터 성능의 향상, 데이터 수집·저장기술의 발달, 데이터베이스 관리시스템과 데이터 웨어하우스 기술의 광범위한 사용은 기업 내부에 대량의 데이터를 축적하는 것을 가능하게 하였다. 그러나 데이터를 수집하고 저장하는 능력에 비해 그것을 분석하고 정보와 지식을 획득하는 능력이 미흡하여 데이터 과잉 및 유용한 정보의 빈곤문제가 발생하였다. 따라서 기업들은 방대한 양의 데이터에 내재된 정보와 지식을 발견 또는 채굴하는 능력의 제고에 관심을 가지게 되었으며, 이러한 관심에 따라 발생한 것이 데이터마이닝이다. [4]

데이터마이닝의 개념이 등장한 이후, 여러 분야에서 연구되고 발전되면서 데이터마이닝에 대한 다양한 정의가 내려지고 있다.

컴퓨터 과학(Computer Science)분야에서는 데이터마이닝을 패턴인식기술뿐 아니라 통계적, 수학적 분석방법을 이용하여 저장된 거대한 자료로부터 우리에게 유익하고 흥미 있는 새로운 관계, 성향, 패턴 등 다양하고 가치 있는 정보를 찾아내는 일련의 과정으로 정의하고 있으며, MIS 분야에서는 거대한 데이터베이스 혹은 자료에서 유용한 정보를 유출하는 일련의 과정뿐 아니라 값진 정보를 사용자가 전문적 지식 없이 사용할 수 있는 의사결정지원시스템의 개발과정을 통틀어 데이터마이닝이라고 한다. 또한 통계학에서는 데이터마이닝을 올바른 의사결정을 지원하기 위한 자료 분석(Data Analysis) 및 모델선택(Model Selection)이라고 정의한다. 이와 같은 정의를 바탕으로 정리를 하면, 의사결정을 위해 대용량의 데이터로부터 이전에 알려지지 않은 관계, 패턴, 규칙 등을

탐색하고 찾아내어 모형화 함으로써 유용한 지식을 추출하는 반복적인 프로세스라고 볼 수 있다. [14]

데이터마이닝으로 정보를 찾아내는 방법은 어떤 특정 기법과 그 기술자체만을 의미하는 것은 아니고, 업무상의 문제를 이해하고 이러한 문제를 해결하기 위하여 정보기술을 적용하는 포괄적인 과정을 의미한다. 즉, 유용한 정보의 추출을 위한 방법론이라고 할 수 있다. 따라서 데이터마이닝을 효율적으로 수행하기위하여 각종의 통계적인 기법과 데이터베이스기술뿐만이 아니라 산업공학, 신경망, 인공지능, 전문가시스템, 퍼지논리, 패턴인식, 기계학습, 불확실성 추론, 정보검색에 이르기까지 각종 정보기술과 기법들을 사용하게 된다. 이러한 데이터마이닝을 통하여 거대한 데이터 속에 있어있는 전략적인 정보를 발견할 수 있으며 또한 이러한 정보를 이용하여 주요 비즈니스 프로세스 개선에 사용될 수도 있다. [2]

데이터마이닝의 활용분야는 매우 다양하다. 판매에서는 고객의 구매패턴을 분석할 수 있어, 일부 고객에게만 집중적으로 판촉을 하는 Target Marketing을 할 수 있다. 제조업에서는 불량품 진단, 공장 자동화 등을, 통신에서는 고객이탈 방지, 의료에서는 질병진단, 보험에서는 Claim분석 등을 할 수 있다. 또한, 금융 분야에서는 신용평가, 신용카드 사기 감지, 증권 가격예측 등을 할 수 있다. [5]

2.2 데이터마이닝의 특징

데이터마이닝은 몇 가지 관점에서 특징을 가지고 있는데, 이를 간단히 요약하면 다음과 같다.

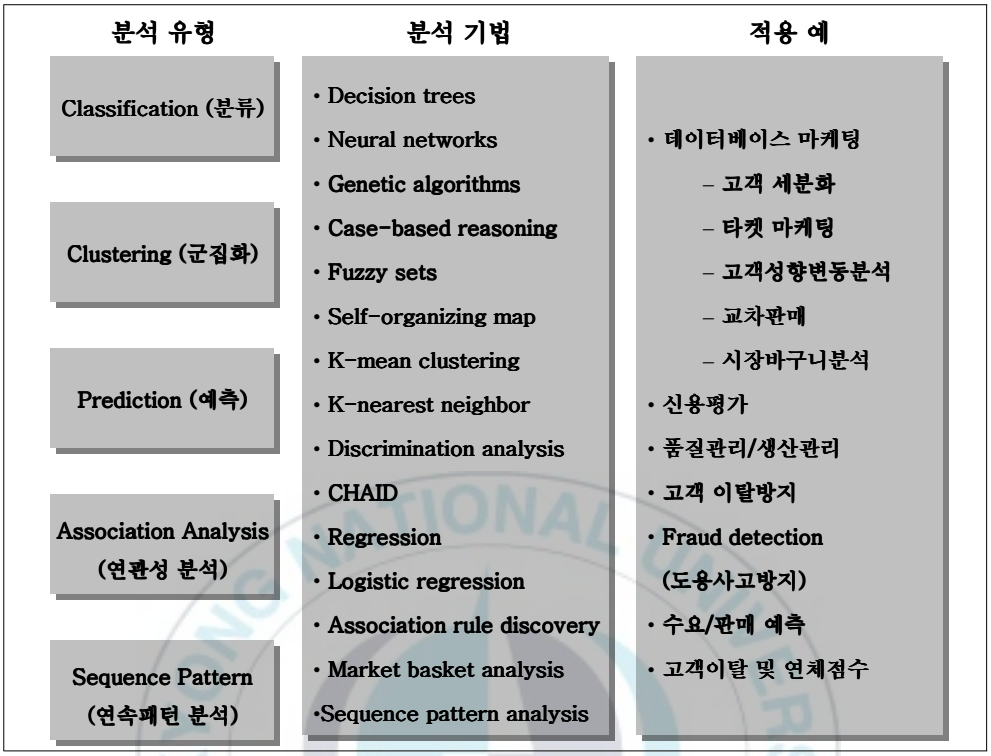
1 대용량의 관측 가능한 자료처리 : 데이터마이닝은 대용량의 관측 가능한 자료를 다룬다. 실험 자료는 가설검정 등의 구체적인 문제에 답

하기 위해 여러 요인들이 통제되고 조작된 가운데 만들어진다. 그러나 관측 자료는 시간의 흐름에 다라서 비계획적으로 축적되며, 자료 분석을 염두에 두고 수집되지 않는 것이 일반적이다.

- 컴퓨터 중심적 기법 : 데이터마이닝은 컴퓨터 중심적 기법이다. 현대의 컴퓨터 중심적 기법들은 기존의 기법들로서는 해결하기 곤란한 경우에 있어, 이를 해결하기 위하여 컴퓨터의 강력한 처리속도와 능력을 활용할 수 있도록 해주고 있다.
- 경험적 방법에 근거 : 대다수의 데이터마이닝 기법들은 수학적으로 증명되고 발전된 것이 아니라 경험에 기초하여 개발되었다. 이러한 기법들은 그 특성이 수리적으로 밝혀지지 않은 것들이 많다.
- 일반화에 초점 : 데이터마이닝의 주요 관심은 통계적 추론과 검정보다는 예측모형의 일반화에 있다.
- 의사결정 지원 : 데이터마이닝은 기업의 다양한 경영상황 하에서 경쟁력 확보를 위한 의사결정을 지원하기 위해서 활용될 수 있다.
- 활용분야 : 데이터마이닝 기법들은 통계학(statistics), 컴퓨터과학(computer science), 인공지능(artificial intelligence : AI), 공학(engineering)과 같은 분야에서 개발된 특징을 가지고 있다. 그러나 실제로 이를 활용하는 전문가들은 경영, 경제, 정보기술 분야에 있는 사람들이다. [1]

2.3 데이터마이닝의 분석유형 및 기법

데이터마이닝은 기계학습, 패턴인식, 데이터베이스, 통계학, 인공지능, 지식획득, 데이터 시각화, 고성능 컴퓨팅 등 다양한 연구 분야와 관련을 가지면서 많은 기법들이 개발되었다. [그림 2.1]은 데이터마이닝의 분석유형, 분석기법, 적용 예를 나타낸 것이다.



[그림 2.1] 데이터마이닝 기법

분석유형은 Classification(분류), Clustering(군집화), Prediction(예측), Association Analysis(연관성 분석)으로 크게 5가지로 나누어 살펴볼 수 있다.

Classification(분류)은 미리 정의된 범주형 클래스가 있을 때, 어떤 특정 데이터 아이템이 어느 클래스에 속하는지 결정하는 것이다. 예를 들어 고객을 우량고객, 불량고객, 일반고객의 세 클래스로 구분할 때, 특정 고객이 어느 클래스에 속하는지 분류하는 것이다. 판별분석, 회귀분석, 의사결정나무, 신경망 등의 분석기법이 사용된다.

Clustering(군집화)은 데이터 아이템의 집합이 주어졌을 때, 비슷한 특징을 가진 아이템들이 함께 그룹 지어지도록 데이터 아이템의 집합을

몇 개의 소집단으로 분할하는 것으로, 고객 집단을 구매 형태가 비슷한 몇 개의 소집단으로 구분하는 것이 군집화의 예가 될 수 있다. Self organizing, K-mean clustering, K-nearest neighbor 등의 분석기법이 있다.

Prediction(예측)은 데이터 아이템과 예측모형이 주어졌을 때, 그 데이터 아이템의 특정속성 값을 예측하는 것이다. 신용카드의 경우, 남의 카드를 훔치거나 주워서 몰래 불법적으로 사용하는 것을 방지하는 것이 중요하다. 이를 위해 신용카드의 거래 당일 누적 회수, 거래 간 시간차이, 평균거래금액, 사용 장소 등을 이용하여 신용카드 불법사용 예측모형을 만든 후, 특정 카드거래가 불법일 가능성을 예측할 수 있다.

Association analysis(연관성분석)은 데이터 아이템에 어떤 패턴이 존재하는 것이 다른 패턴이 존재하는 것을 암시하는 것과 같은 관계를 찾아내는 것이다. 예를 들어 '신발을 구매하는 고객의 10%는 양말을 동시에 구입한다.'와 같이 하나의 거래나 사건에 포함되어 있는 항목들 간의 연관성을 파악하는 것이 그 예가 될 수 있다.

Sequence pattern(연속패턴분석)은 연관성 분석에 시간정보가 포함된 형태로서, 시간상에 순차적으로 나타나는 데이터 아이템간의 관계를 찾아내는 것이다. 주식 투작에서 '지난주에 X 주식이 5일 동안 10% 오르고 Y 주식이 같은 기간동안 10~20%올랐을 때, 금주에 Z 주식이 오를 확률은 70%이다.'와 같은 패턴을 찾아내는 것이 연속패턴분석의 예이다. [3]

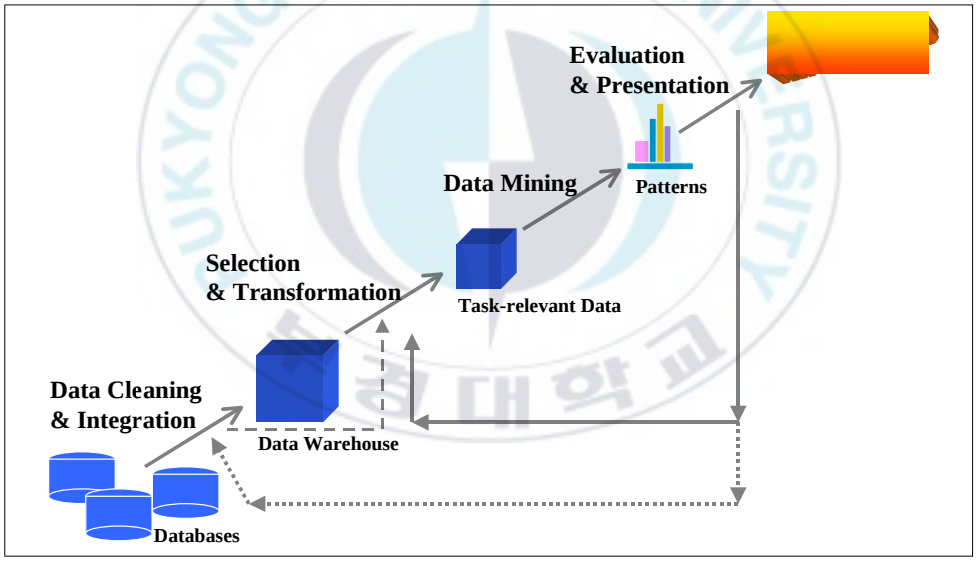
분석기법으로는 Decision trees, Neural networks, Genetic algorithms 등 많은 기법들이 있다. 이 장에서는 이러한 기법들에 대한 자세한 설명은 생략하기로 하며, 본 연구에서 사용한 의사결정나무와 군집분석에 대한 것만 이 후에 설명하고자 한다.

데이터마이닝의 활용분야는 매우 다양하며, 현재에도 그 활용분야

가 점차 늘어나고 있다. 그 중에서도 특히 기업의 의사결정과 향후 대처에 대한 문제에서 많이 활용되고 있으며 주요 활용분야는 데이터베이스 마케팅, 신용평가, 품질개선, 고객이탈방지, 도용사고방지, 수요 및 판매 예측, 고객이탈 및 연체점수 등 여러 분야에서 사용되고 있다.

2.4 데이터마이닝 수행과정

데이터 베이스로부터 추출된 데이터로부터 유용한 정보를 얻기 위해서는 [그림 2.2]와 같은 과정을 따르게 되며, 이러한 단계들은 반복적으로 수행하는 것이 일반적이다.



[그림 2.2] 데이터마이닝 수행과정

l Data Cleaning(데이터 정제) : 데이터베이스에는 일관성이 없고 불완전하며 오류가 있는 데이터가 존재할 수 있다. 따라서 데이터 정제과정을 통해 데이터의 무결성을 보장해주어야 한다.

l Data Integration(데이터 통합) : 데이터가 복수개의 이질적인 데이터

베이스에 분산되어 있는 경우에는 이들의 통합이 전제된다.

- Data Selection(데이터 선택) : 분석하고자 하는 작업과 관련된 데이터들이 데이터베이스로부터 검색된다.
- Data Transformation(데이터 변환) : 요약이나 집계 등과 같은 연산을 수행함으로써, 데이터마이닝을 위해 적합한 형태로 데이터를 변환하거나 합병정리 한다.
- Data Mining(데이터마이닝) : 데이터 패턴을 추출하기위하여 다양하고 상이한 기능별 데이터마이닝 기법들을 적용하는 과정이다. 동일한 데이터마이닝 기능을 수행할 수 있는 기법이 여러 가지 존재하므로 이중에서 적합한 방법을 선택해야 한다.
- Evaluation(패턴 평가) : 도메인 전문가와 협력하여 마이닝된 결과를 해석하고 발견된 패턴이 가치가 있는지 또는 적용 가능한지를 평가하는 단계이다. 경우에 따라서는 앞 단계로 돌아가 마이닝을 재수행하기도 한다.
- Presentation(지식 표현) : 실제로 발견된 패턴을 지식화하거나 문서화하고 관련 부서에 보고하는 단계이다. 또한 기존에 추출된 지식과 상충하는지 검사하고 해결하는 과정도 포함된다. [16]

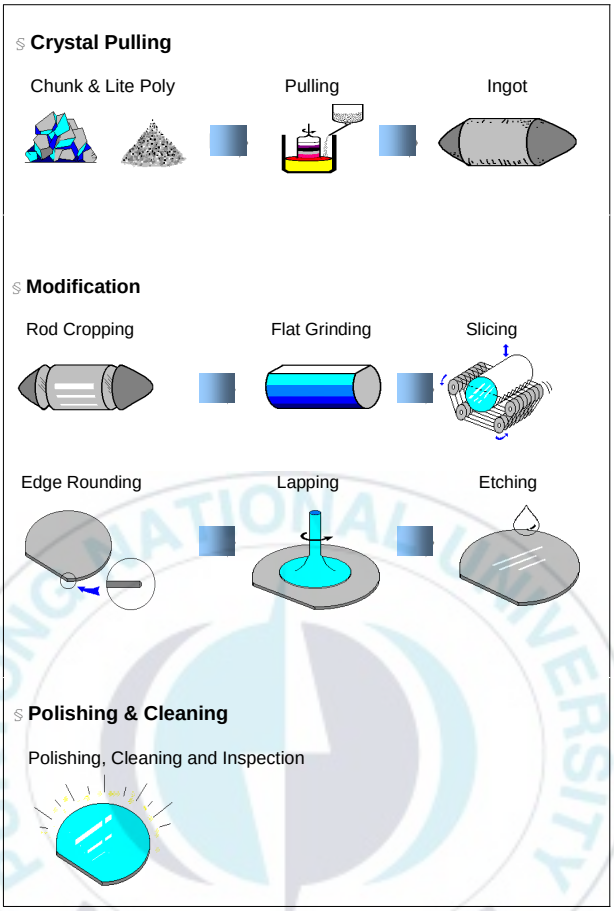
제 3 장 데이터 분석사례

본 장에서는 M사의 제조공정에서 얻어진 대용량 데이터를 사용하여 실리콘 웨이퍼의 중요품질특성 중 하나인 Warp의 불량원인을 분석한 사례를 보이고자 한다. 단순히 통계적인 분석에 의한 방법뿐만이 아니라 공정에 대한 경험과 기술적인 지식을 활용하여 의미 있는 정보를 도출하는 과정을 소개하며, 이후 4장에서는 데이터마이닝의 기법 중 군집분석과 의사결정나무를 적용한 사례를 소개하고자 한다.

3.1 제품 소개 및 공정소개

실리콘 웨이퍼는 반도체 칩의 핵심 소재로서 다결정의 실리콘을 원재료로 하여 만들어진 결정 실리콘 박판을 말한다. 실리콘은 반도체 산업에 매우 안정적으로 공급될 수 있는 재료일 뿐 아니라 독성이 전혀 없어 환경적으로 매우 우수한 재료이다. 또한, 실리콘으로 만들어진 웨이퍼는 비교적 고온에서도 소자가 작동하는 장점이 있다. 이러한 장점 때문에 실리콘 웨이퍼는 반도체 산업에서 DRAM, ASIC, HTR(Transistor), CMOS, ROM 등 다양한 형태의 반도체 소자를 만드는데 이용되며, 이들 소자는 컴퓨터, 전자제품, 산업용기계 등 산업분야에 없어서는 안 될 중요한 부품들이다. [13]

실리콘 웨이퍼의 제조공정은 [그림 3.1]에서 보는 바와 같이 크게 Crystal Pulling, Modification, Polishing & Cleaning의 세 부분으로 구성되어 있다.



[그림 3.1] 실리콘 웨이퍼의 제조공정

Crystal Pulling은 ingot을 형성하는 공정으로 Chunk Poly와 Lite Poly를 석영도가니에 담은 후 소량의 Dopant를 첨가하여 섭씨 1400도 이상의 고온에서 녹인 후 ingot을 생성하게 된다.

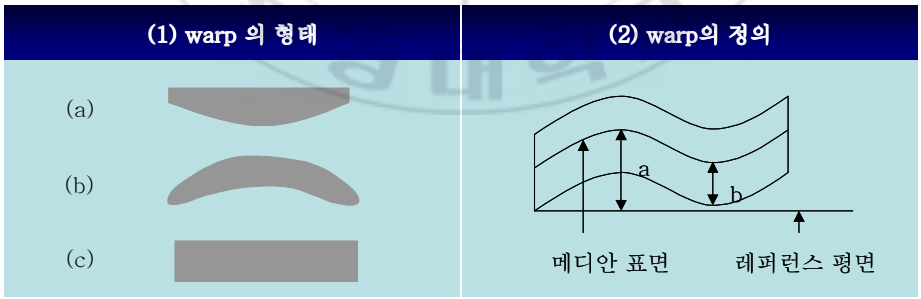
Modification공정은 다시 Rod Cropping, Flat Grinding, Slicing, Edge Grinding, lapping, Etching 등 여러 세부 공정으로 나누어진다. Rod Cropping은 ingot의 불필요한 부분을 잘라내어 Rod를 만드는 공정이며, Flat Grinding은 Rod의 표면을 Grinding한 후 결정 방향성을 알 수 있도록 flat을 형성하는 공정이다. 이렇게 얻어진 Rod는 Slicing공정

을 통해 얇게 절단됨으로써 비로소 웨이퍼의 형태를 가지게 된다. 절단된 웨이퍼는 테두리를 둥글게 연마하는 Edge Grinding 공정을 거쳐 Lapping공정에서 표면의 saw mark와 scratch 등을 제거하게 된다. 마지막으로 Etching공정을 통해 Lapping공정에서 발생한 미세한 기계적 결함을 화학적 식각을 통해 제거한다.

Polishing & Cleaning 공정에서는 경면 연마공정(Polishing)과 세정공정(Cleaning) 및 최종 검사(Inspection)를 거쳐 제품을 완성하게 된다.

3.2 Warp의 정의 및 불량원인 도출

Warp는 웨이퍼가 휘어진 정도를 말하며 [그림 3.2]의 (1)과같이 여러 형태가 있을 수 있다. 그림에서 (a), (b)는 심하게 휘어진 경우이며 (c)는 이상적으로 편평하게 절단된 경우이다. 현장에서는 Warp의 측도로서 [그림 3.2]의 (2)와같이 웨이퍼의 메디안 표면과 레퍼런스 평면 사이의 거리 중 최대값(a)과 최소값(b)의 차이를 사용한다. [13]

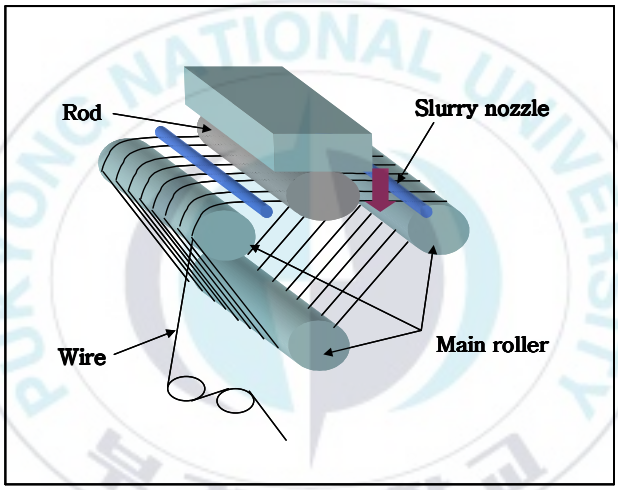


[그림 3.2] Warp의 정의 및 형태

Warp에 직접적인 영향을 주는 공정은 Slicing공정으로서 Rod를 절단할 때 각 웨이퍼의 두께가 균일하게 유지되도록 하는 것이 중요하다.

절단방식은 예전에는 다이아몬드를 사용한 Blade Type의 절단 방식을 사용하였으나 절단시간의 감소, 두께 편차의 감소, Warp감소 등 여러 가지 장점으로 인해 Wire를 이용한 절단방식이 사용되고 있다.

Slicing 공정에서 Rod를 절단하는 작업은 Multi-wire saw 장비에 의해 이루어지므로 Warp 불량 발생 메커니즘을 이해하기 위해서는 장비의 구성과 작동원리를 알 필요가 있다. [그림 3.3]은 장비의 구성 요소를 보여 주며, Rod는 감겨지는 Wire와 Wire위에 뿌려진 Slurry의 마찰을 이용하여 여러 개의 웨이퍼로 잘려진다.



[그림 3.3] Multi wire saw 기계

공정과 장비에 대한 이해를 바탕으로 특성요인도, XY matrix, FMEA 등의 도구를 이용하여 Warp 불량의 잠재적인 중요원인들을 도출하였다.

3.3 통계 및 기술적 지식에 의한 분석

일반적으로 많은 양의 데이터를 이용하여 통계적인 분석을 실시할

경우, 많은 정보를 확보하고 있다는 장점도 있지만 다음과 같은 문제점들도 있다.

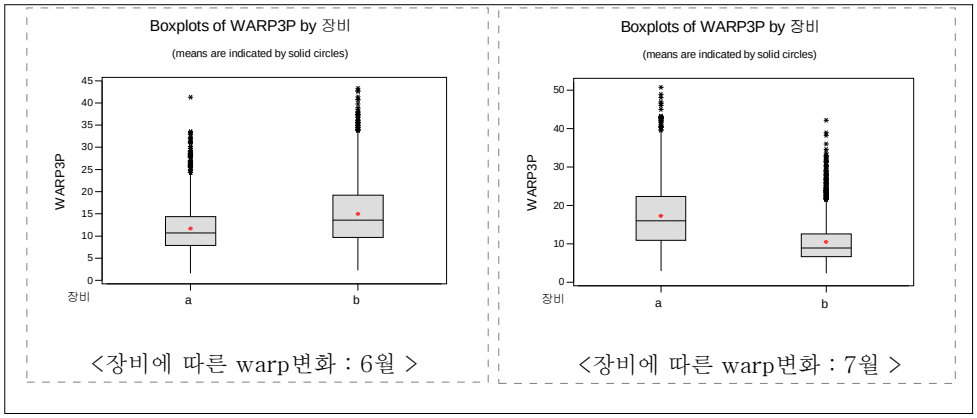
첫째, 많은 데이터들이 모두 신뢰할 수 있는 데이터인지 확신이 없을 경우, 어떤 데이터를 배제해야 할 것인지 결정하기 어렵다.

둘째, 실제로는 별 의미 없는 작은 차이들을 통계적인 분석결과로는 매우 유의한 차이가 있는 것으로 나타나게 된다. [표 3.1]의 경우, 실제로 두 종류의 장비에 대해 Warp 평균 및 표준편차 차이에 대한 가설 검정을 실시한 결과 p값은 모두 0.000으로 나왔지만 실제 차이는 의미를 부여할 수 있는 정도의 값이 아니었다.

[표 3.1] 장비별 Warp의 평균 및 표준편차 비교 (단위 : μm)

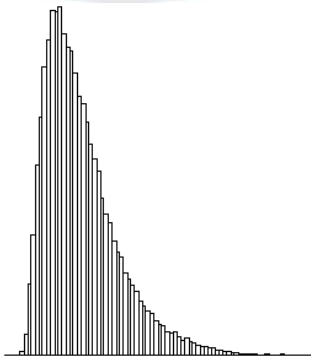
가설	검정방법	분석결과	
장비 c, d별 차이	One-way ANOVA	p-value	평균차이
		0.000	0.671
	Test for equal variances	p-value	표준편차 차이
		0.000	0.307

셋째, 각 인자별로 Warp에는 영향을 주지만 일정한 패턴이 나타나지 않을 수도 있다. [그림 3.4]의 경우, 6월에는 장비 a가 Warp의 값이 더 작은데 비하여 7월에는 장비 b가 Warp값이 더 작게 나타나서 월별로 Warp가 일정한 패턴이 나타나지 않았다.



[그림 3.4] 장비에 따른 Warp의 변화

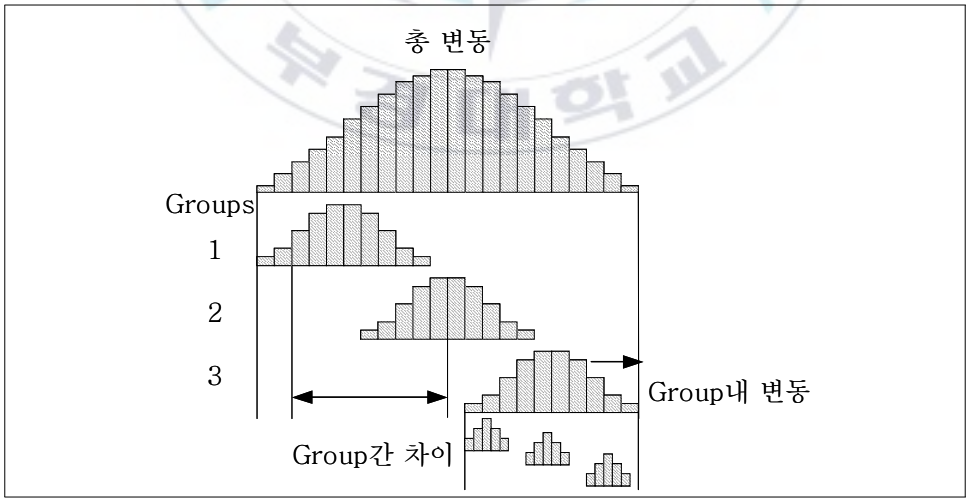
넷째, 안정되지 않은 공정에서 나온 수많은 데이터들이 뒤섞임으로써 전체적인 모양은 마치 정상적인 공정에서 나온 데이터처럼 보일 수 있다. [그림 3.5]는 전술한 Warp데이터를 히스토그램으로 나타낸 것으로 모양이 품질 특성치에 한쪽규격이 주어져 있는 경우의 정상적인 공정에서 나온 데이터의 전형적인 형태를 취하고 있다. 그러나 평균에 비해 표준편차가 지나치게 큰 것으로 미루어 적정 크기의 부분군으로 나누어 관리도를 작성하여 공정을 해석하면 공정에 이상이 있는 것으로 나타날 가능성이 크다.



[그림 3.5] Warp 데이터의 히스토그램

산포가 지나치게 큰 공정의 변동을 구성 요소별 부분군으로 분해해 보면 [그림 3.6]과 같이 부분군 간 평균의 변동이 표준편차의 크기에 반영되어 있을 경우가 많다. 문제는 어떤 기준으로 부분군을 구분해야 의미 있는 변동의 패턴을 식별할 수 있는가 이며, 이것은 현장의 경험과 기술적인 지식을 기초로 여러 가지 각도에서 데이터를 분석해봄으로써 알아낼 수밖에 없다.

여기서 분석 대상으로 하는 Slicing 공정의 경우 현장에서 두 모델의 장비를 사용하고 있으나, 모델별로는 실질적인 유의차를 보여주지 않았으므로 장비별로 차이를 분석해 보았다. 결과 장비별로 상당한 차이를 보였으나 월별로 일관성 있는 결과를 보여주지는 않았다. 즉, 제일 좋은 결과 혹은 제일 좋지 않은 결과를 주는 장비가 매월 달라지므로 변동의 원인이 장비에 있는 것이 아니라 주기적으로 교환해 주는 소모품들과 작업조건에 있을 것으로 추정되었다. 따라서 Slurry, Main Roller, Wire 등 장비에 장착되는 소모품들에 의한 영향을 순서대로 분석하였다.

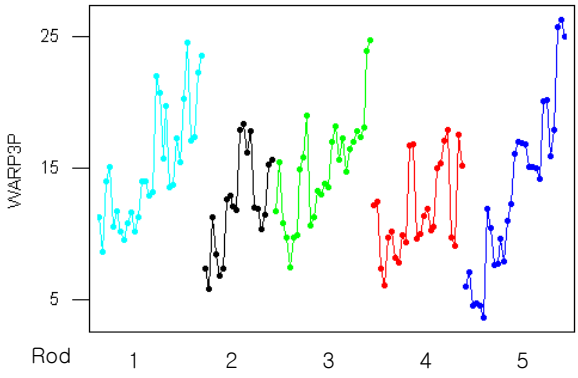


[그림 3.6] 공정 총 변동의 구성

3.3.1 인자별 분석 결과

Slurry는 보통 SiC(실리콘카바이드)로 구성되어 있으며 그 크기는 지름 10~15μm정도이다. 이러한 SiC알갱이는 연마제 역할을 하는 동안 점점 닳기 때문에 Slurry를 교체하거나 새로운 Slurry를 공급할 필요가 있다. 또한 Slurry는 Wire와의 마찰을 통해 Rod를 자르기 때문에 Slurry의 움직임이 중요하며 Slurry의 밀도와 온도는 이러한 움직임에 중요한 영향을 주는 요인이다. 최근 환경문제가 대두되면서 사용한 Slurry를 재생하여 사용하고 있으며, 재활용 시 Slurry내 이물질을 제거하는 것이 중요하다. 분석 결과 Warp는 Slurry의 밀도, 사용횟수, 공급량, 온도에 영향을 받는 것으로 나타났으나 그 패턴에 일관성이 없어서 다른 변수에 의한 영향을 강하게 받고 있을 것이라는 유추 외에 의미 있는 정보를 얻을 수 없었다.

Main Roller의 패임과 샤프트 마모, Wire의 장력, 뒤틀림, 정렬 상태 등은 Main Roller의 부위별로 달라질 수 있으므로 이들 중 일부 요인이 Warp에 영향을 준다면 같은 Rod를 잘라 제작된 웨이퍼 간에도 큰 편차가 있을 것이다. [그림 3.7]은 몇 개의 Rod에 대해 위치별로 Warp 변화를 살펴보기 위해 Run Chart를 작성해 본 것이다. 그림으로부터 알 수 있듯이 Warp는 동시에 작업된 같은 Rod 내에서의 차이가 Rod간 차이보다 훨씬 더 크다는 것을 알 수 있다.



[그림 3.7] Rod 위치별 Warp의 변화

이러한 변화를 확인하기 위해 몇 대의 장비에서 작업된 총 903개 Rod 대해 추가 분석하여 [표 3.2]에 정리하였다.

[표 3.2] Rod의 패턴

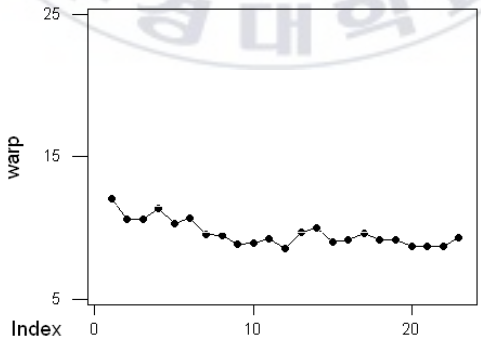
pattern	기계				계
	1	10	11	14	
↘	20	10	45	25	100
↗	97	89	61	38	283
∪	43	49	49	74	215
∩	26	16	10	4	56
→	7	23	36	25	91
∧	37	41	14	15	107
∨	11	18	12	10	51
계	241	244	227	191	903

같은 Rod 내 Warp의 변화 패턴은 [표 3.2]에서 보는바와 같이 총 7가지로 나타났다. 그 중 증가 형이 제일 많이 나타났으며 그 다음으로 U

자형이 많이 나타났다. 이러한 패턴이 생기는 근본 원인은 Wire Saw 장비의 작동 메커니즘을 검토하여 분석함으로써 어느 요인에 의한 것인지 규명할 수 있을 것으로 판단된다.

3.3.2 Slicing 공정의 작동 메커니즘 분석

같은 Rod를 잘라 형성된 웨이퍼간 품질특성이 균일하지 못하다는 것은 Rod를 장비에 장착하여 작업을 진행하는 과정에 무언가 문제가 있음을 의미한다. 더구나 품질특성의 변화 패턴이 [그림 3.7]과 같다는 것은 Rod 부위별로 다른 영향을 줄 수 있는 요소가 주된 원인임을 강하게 시사하고 있다. 절단작동 시 Rod 부위별로 달라질 수 있는 요소로는 Slicing 장비의 작동메커니즘작동원리로써 Main Roller의 패임, Wire 마모정도, Wire 정렬상태 등이 있다. 이러한 요인들의 영향을 완충할 수 있도록 Main Roller를 다시 설계하여 시험하고 데이터를 수집, 분석한 결과 [그림 3.8]과 같은 결과를 얻을 수 있었다. Rod 위치별 웨이퍼의 품질 특성이 이전에 비해 훨씬 더 균일함을 보여주고 있다.



[그림 3.8] 개선 후 Rod 위치별 Warp의 변화

제 4 장 데이터마이닝 기법의 적용사례

4.1 서론

앞장에서는 통계적인 분석과 기술적인 지식을 활용하여 의미 있는 정보를 도출하는 과정을 살펴보았다. 본 장에서는 앞에서 수행한 데이터 분석사례에 데이터마이닝을 적용함으로써 그 결과를 비교분석해 보았다. 데이터마이닝 기법은 군집분석과 의사결정나무를 이용하여 비교하고자 한다.

4.2 군집분석

4.2.1 군집분석의 개념

군집분석이란 주어진 관찰치 중에서 유사한 것들을 몇몇의 집단으로 그룹화하여, 각 집단의 성격을 파악함으로써 데이터 전체의 구조에 대한 이해를 돕고자 하는 분석방법이다. 즉, 대용량 데이터에서 개개의 관찰치를 요약하는 것보다는 전체를 유사한 관찰치들의 군집(cluster)으로 구분하여 복잡한 전체보다는 그를 잘 대표하는 군집들을 관찰함으로써 전체 데이터에 대한 의미 있는 정보를 얻어낼 수 있을 것이다. 군집분석은 분명한 분류기준이 없거나 알려져 있지 않은 상태에서 활용될 수 있는 기법이며, 또한 제조 공정의 품질 개선을 위한 데이터마이닝 적용에 가장 많이 사용되는 기법 중의 하나이기도 하다. [1]

군집분석에서는 주어진 변수들을 이용하여 각 관찰치가 서로 얼마나 유사한지 또는 유사하지 않은지를 측정할 수 있는 척도가 필요하다.

유사성은 동일한 관찰치에 대한 정의가 모호하므로, 군집분석에서는 일반적으로 비유사성을 기준으로 하고, 그 척도로서 거리(distance)를 주로 사용한다.

두 관찰치 X_i, X_j 사이의 거리를 d_{ij} 라 하면 이는 일반적으로 다음의 조건을 만족시키는 것을 기본으로 한다.

(1) $d_{ij} \geq 0; i = j$ 이면 $d_{ij} = 0$

(2) $d_{ij} = d_{ji}$

(3) $d_{ik} + d_{jk} \geq d_{ij}$

대부분의 군집분석 알고리즘들은 총 n개의 관찰치들 사이의 거리를 이용해 상대적으로 거리가 가까운 관찰치들을 같은 군집에 포함시켜서 군집간(서로 다른 군집에 속하는 관찰치들 사이)변이에 비해 군집 내(동일한 군집에 속하는 관찰치들 사이)변이를 줄이는 점에 기초하고 있다. 일반적으로 p개의 변수를 갖는 두 개체 X_i와 X_j를 각각 다음과 같이 표현할 경우,

$$X_i = (X_{i1}, X_{i2}, \cdots, X_{ip})$$
$$X_j = (X_{j1}, X_{j2}, \cdots, X_{jp})$$

두 개체사이의 거리 d_{ij}를 정의하는 방법은 여러 가지가 있는데, 그 중에서 민코브스키(Minkowski)거리는 다음과 같이 정의되며 특히 m = 2인 경우에 가장 널리 사용되는 유클리드(Euclidean)거리가 된다.

민코브스키 거리 : $d_{ij} = \left[\sum_{k=1}^{p} | X_{ik} - X_{jk} |^{m} \right]^{1/m}$, m은 실수

유클리드 거리 : $d_{ij} = \left[\sum_{k=1}^{p} | X_{ik} - X_{jk} |^{2} \right]^{1/2}$

그런데 이와 같은 거리의 측도들은 일반적으로 자료의 척도에 민감하므로 주의해야한다. 즉, 동일한 거리를 측정한 경우에도 측도에 따라서 그 값에 있어 차이가 크다(예를 들어, cm와 mm). 이에 대한 문제를 해결한 거리정의 방법으로는 표준화 거리와 마할라노비스(Mahalanobis)거리 등이 있다. [1]

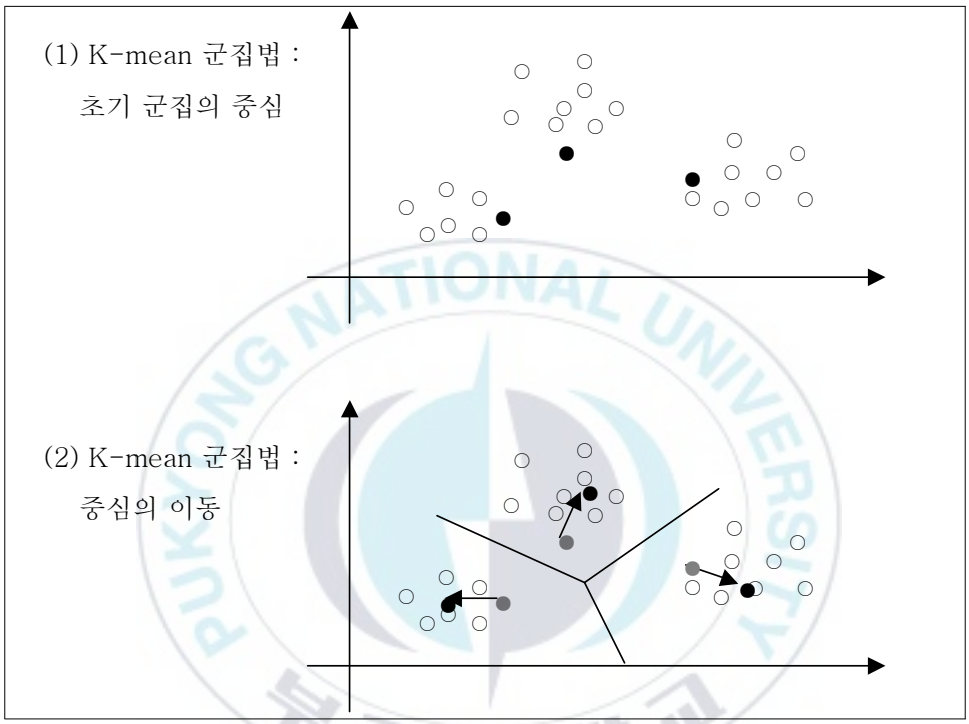
4.2.2 K-mean 군집분석

여러 군집방법 중에 대용량 데이터를 빠르게 처리할 수 있으며, 알고리즘도 비교적 간단한 방법으로 K-mean 군집방법이 있다. 이는 관찰치들 사이의 거리를 이용해 주어진 기준을 최적화 하도록 구현되므로 최적분리 군집방법이라고도 한다. 이는 사전에 결정된 군집수 k에 기초하여 전체 데이터를 상대적으로 유사한 k개의 군집으로 구분하는 방법이다. 기본적으로 다음의 과정을 따른다.

(1) 군집의 수 k를 정한다.
(2) 초기 k개 군집의 중심을 선택한다.
(3) 각 관찰치를 그 중심과 가장 가까운 거리에 있는 군집에 할당한다.
(4) 각 군집별로 그에 속하는 관찰치를 이용해 새로운 중심을 계산한다.

(5) (3)~(4)의 과정을 기존의 중심과 새로운 중심의 차이가 없을 때까지 반복한다.

앞의 과정을 그림으로 살펴보면 [그림 4.1]과 같다.



[그림 4.1]K-mean 군집분석

[그림 4.1]의 (1)은 거리를 표현하는 변수가 X_1, X_2로 하여 2차원으로 표현한 각 관찰치의 위치이다. 그림을 보고 군집 수는 3으로 결정하고 각 군집의 중심을 전체 데이터를 임의로 3등분해 각각의 평균으로 정하였다. [그림 4.1]의 (2)는 주어진 군집의 중심을 기준으로 관찰치를 가장 가까운 군집에 할당한 결과이고, 새로이 할당된 관찰치를 기준으로 그 군집의 중심을 다시 계산한 결과이다. 이 결과에서 군집의 평균이 이

동된 것을 볼 수 있다. 이제 이동된 군집의 평균을 통해 관찰치를 재 할당한 후 다시 새로운 군집의 평균을 계산한다. 이러한 과정을 군집의 평균이 거의 변화하지 않을 때까지 계속하면 전체자료를 가장 잘 구분하는 3개의 군집을 얻을 수 있다.

K-mean 군집방법은 군집수를 사전에 결정해야하며, 적절한 군집수를 택해야만 의미 있는 분석결과도 얻을 수 있다. 예를 들어 [그림 4.1]을 사전에 관찰하지 못하고 군집수를 2로 정한다면 그 결과는 데이터를 잘 요약하리라 기대하기 어렵다. 보통 군집수를 결정하는 방법은 적절한 다변량 통계분석법을 이용해 관찰치의 위치를 시각화해 관찰해서 결정하는 방법이 있다. 이는 보통 주성분 분석을 통해서 변수의 수를 줄이고 이를 2차원 혹은 3차원의 그래프를 통해서 살펴본다. [1]

4.2.3 군집분석의 장점 및 단점

군집분석의 장점은 다음과 같이 크게 3가지로 요약할 수 있다.

첫째, 탐색적인 기법이다. 군집분석은 그 자체가 대용량 데이터에 대한 기법으로서, 주어진 데이터의 내부구조에 대한 사전적인 정보 없이 의미 있는 자료구조를 찾아낼 수 있는 방법이다. 따라서 군집분석의 결과는 이후의 여러 가지 모형화를 위한 분석에 사용될 수 있다. 둘째, 다양한 형태의 데이터에 적용이 가능하다. 분석을 위해서는 기본적으로 관찰치간의 거리를 데이터형태에 맞게만 정의하면, 거의 모든 형태의 데이터에 대하여 적용이 가능한 방법이다. 예를 들어서, 신문기사 모음이나 보험에 대한 불만청구서 자료에 대해서도, 관찰치들 사이의 비유사성(혹은 유사성)의 정도만을 잘 표현해 낼 수 있다면 분석이 가능하다. 셋째, 분석방법의 적용이 용이하다. 대부분의 군집방법이 분석대상 데이터에 대해 사전정보를 거의 요구하지 않으므로 적용하는 데에 큰 어려움이 없

다. 즉 모형화를 위한 분석과 같이 사전에 특정변수들에 대한 역할정의가 필요하지 않고, 다만 관찰치들 사이의 거리만이 분석에 필요한 입력자료가 된다.

그러나 군집분석의 결과는 관찰치 사이의 비유사성인 거리 또는 유사성을 어떻게 정의하는가에 크게 좌우된다. 따라서 특히 여러 가지 자료유형을 포함하는 데이터의 경우, 관찰치들 사이의 거리를 정의하고 각 변수에 대한 가중치를 결정하는 것은 매우 어려운 문제이다. 또한, K-mean 군집분석에서는 사전에 정의된 군집수를 기준으로 동일한 수의 군집을 찾게 되므로 만일 군집수 k가 원 데이터구조에 적합하지 않으면 좋은 결과를 얻을 수 없으므로 여러 번의 탐색적인 군집분석의 절차가 이어져야 한다. K-mean 군집분석은 탐색적인 분석방법으로서의 장점을 가지고 있는 반면에, 사전에 주어진 목적이 없으므로 결과를 해석하는데 있어서 어려움이 있다. 따라서 주어진 변수에 따라 잘 구분된 군집이라 하여도 그 결과를 충분히 이해하고 실제적으로 활용하기는 쉽지 않다.

군집분석은 구조가 복잡하고 변수가 많은 대용량 데이터를 분석하기 위한 좋은 방법이다. 따라서 데이터마이닝을 수행하기 위한 초기 작업으로 유용하지만, 이는 복잡한 데이터를 요약하는 도구일 뿐이고, 유사성을 갖는 군집내의 관찰치들을 통한 변수들 사이의 규칙이나 패턴을 찾는 또 다른 데이터마이닝 기법을 사용하여 더욱 유용한 결과를 얻도록 노력해야 할 것이다. [1]

4.2.4 군집분석의 적용

이번 절에서는 3장의 대용량 데이터 분석사례를 군집분석으로 분석해본 결과를 소개하겠다. 데이터 분석은 반응변수 Y에 영향을 주는 X인

자들 중 데이터가 수집되어있는 Main roller run-out, Slurry density, Slurry 유량, Slurry 온도 in, Slurry 온도 out을 분석하였다. 군집분석의 전반적인 절차는 다음과 같다.

(1) 주요 X요인들을 표준화 한 후 유사한 것들을 몇몇의 집단으로 그룹화 한다. 이때, 표준화는 측도차이에 의해 군집 형성 시 잘못된 군집이 형성되는 것을 막기 위한 것이다.
(2) 형성된 군집 중 반응변수 Y인 Warp의 값이 가장 작은 군집을 파악한다. 이는 Warp가 작으면 작을수록 좋은 값을 나타내며 가장 이상적인 값은 0이기 때문이다.
(3) 마지막으로 선택된 군집의 X요인들의 패턴을 파악하고 반응변수Y가 작게 나오도록 하는 X요인들의 조건을 이끌어 낸다.

[그림 4.2]을 살펴보면 총 4개의 군집이 형성된 것을 알 수 있다. 이 중 Warp값이 가장 작은 군집은 군집 4라는 것을 알 수 있다. 따라서 군집 4의 데이터 패턴을 파악함으로써 반응변수 Y를 좋게 하는 X요인의 수준을 파악할 수가 있다. [표 4.1]는 X요인들의 패턴을 정리한 것이다.

Cluster	Name	Mean	Std Dev.	Minimum	Maximum	N
.	WARP3P	11.7120765	6.1107851466	2.63593	56.6643	2000
.	RunOut	13.069	4.0244673149	5	60	2000
.	DENSITY	1.4963481741	0.0119337022	1.45	1.53	1999
.	Flow	96.457358112	5.2337986293	80.1	133	1991
.	Tempin	23.015245737	0.3186434253	21.8	29.9	1994
.	Tempout	25.695588972	0.809317971	23.1	30	1995
1	WARP3P	8.6874928571	2.5449866547	5.12916	11.6569	7
1	RunOut	13	1.1547005384	11	15	7
1	DENSITY	1.4985714286	0.0069006556	1.49	1.51	7
1	Flow	98.833333333	2.1397819204	96.1	100.8	6
1	Tempin	28.74	0.8502940668	27.9	29.9	5
1	Tempout	26.716666667	1.5829297731	25.3	28.7	6
2	WARP3P	16.928627948	5.6636133411	7.9404	56.6643	853
2	RunOut	13.41031653	3.2613811582	5	21	853
2	DENSITY	1.4962323944	0.0121972616	1.45	1.53	852
2	Flow	96.725516921	5.2101260627	81.2	133	851
2	Tempin	22.99212691	0.1077974697	21.8	24	851
2	Tempout	25.539717979	0.7788811979	23.2	30	851
3	WARP3P	8.8079328571	3.4534814934	4.39251	14.6757	14
3	RunOut	38.571428571	14.064216928	30	60	14
3	DENSITY	1.5	0	1.5	1.5	14
3	Flow	89.978571429	3.1248824154	87.6	95	14
3	Tempin	22.985714286	0.036313652	22.9	23	14
3	Tempout	25.678571429	0.2259290952	25.3	25.9	14
4	WARP3P	7.8151952496	2.4275405383	2.63593	15.569	1126
4	RunOut	12.493783304	3.1492442103	5	21	1126
4	DENSITY	1.4963765542	0.0118290167	1.45	1.53	1126
4	Flow	96.321785714	5.2280084043	80.1	130	1120
4	Tempin	23.007651246	0.149722805	22.3	25	1124
4	Tempout	25.808362989	0.8095497401	23.1	29	1124

[그림 4.2] 군집분석 결과

[표 4.1] X요인의 패턴정리

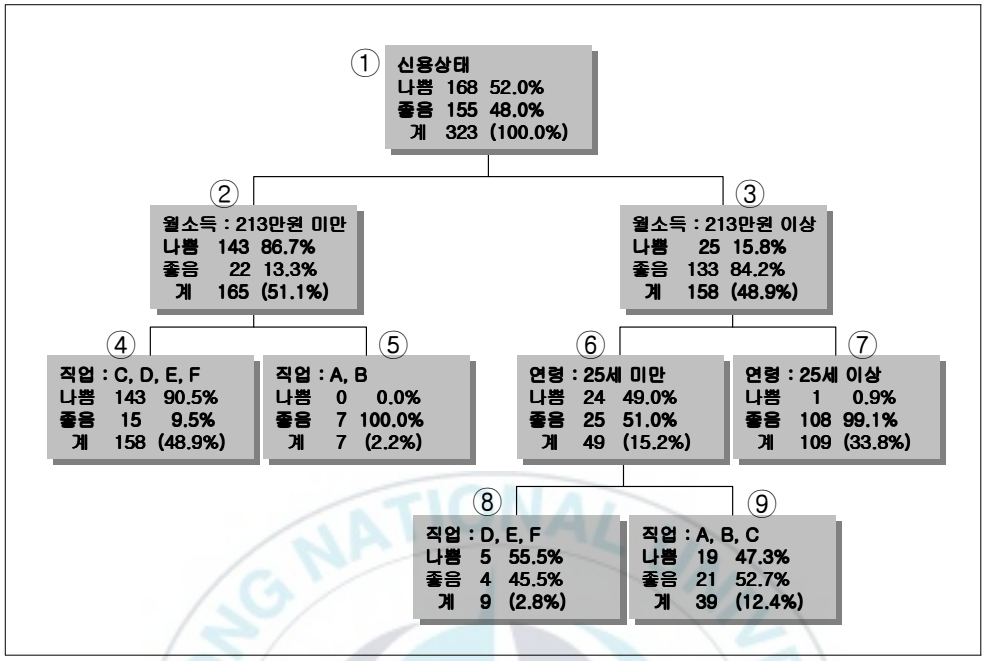
X요인	평균	표준편차
Main roller run-out	12.49	2.43
Slurry density	1.50	0.01
Slurry flow	96.32	5.23
Slurry 온도 in	23.01	0.15
Slurry 온도 out	25.81	0.81

4.3 의사결정나무 분석

4.3.1 의사결정나무의 개념

의사결정나무는 의사결정규칙을 [그림 4.3]과 같은 나무구조로 도표화하여 분류와 예측을 수행하는 분석방법이다. 의사결정나무는 분류 또는 예측의 과정이 나무구조에 의한 추론규칙에 의해서 표현되기 때문에, 다른 방법들(예를 들면, 신경망, 회귀분석 등)에 비해서 연구자가 그 과정을 쉽게 이해하고 설명할 수 있다는 장점을 가지고 있다. 의사결정나무는 분류 또는 예측을 목적으로 하는 어떤 경우에도 사용될 수 있으나, 분석의 정확도보다는 분석과정의 설명이 필요한 경우에도 사용된다. 의사결정나무 분석이 유용하게 활용되는 응용분야는 다음과 같다.

- 세분화(Segmentation) : 데이터를 비슷한 특성을 갖는 몇 개의 그룹으로 분할하여 각 그룹별 특성을 발견하는 경우
- 분류(Classification) : 관측개개를 여러 예측 변수들에 근거하여 목표변수의 범주를 몇 개의 등급으로 분류하고자 하는 경우
- 예측(Prediction) : 자료로부터 규칙을 찾아내고 이를 이용하여 미래의 사건을 예측하고자 하는 경우
- 차원축소 및 변수 선택(Data reduction and variable screening) : 매우 많은 수의 예측변수 중에서 목표변수에 큰 영향을 미치는 변수들을 골라내고자하는 경우
- 교호작용효과와 파악(Interaction effect identification) : 여러 개의 예측변수들이 결합하여 목표변수에 작용하는 교호작용을 파악하고자 하는 경우
- 범주의 병합 또는 연속형 변수의 이산화(Category merging and discrediting continuous variable) : 범주형 목표변수의 범주를 소수의 몇 개로 병합하거나, 연속형 목표변수를 몇 개의 등급으로 이산화 하고자 하는 경우



[그림 4.3] 의사결정나무의 예

의사결정나무는 하나의 나무구조를 이루고 있으며, 마디라고 불리는 구성요소들로 이루어져 있다. 마디는 그 기능에 따라서 다음과 같이 여러 가지로 분류할 수 있다.

- 뿌리마디(root node) : 나무구조가 시작되는 마디로써 전체 자료로 이루어져 있다([그림 4.3]의 ①번 마디).
- 자식마디(child node) : 하나의 마디로부터 분리되어 나간 2개 이상의 마디들을 의미한다. [그림 4.3]의 ②번과 ③번 마디는 ①번 마디의 자식마디이다.
- 부모마디(parent node) : 자식마디의 상위마디를 위미한다. [그림 4.3]의 ①번 마디는 ②번과 ③번 마디의 부모마디이다.

l 끝마디(terminal node) 또는 잎(leaf) : 각 나무줄기의 끝에 위치하고 있는 마디를 의미한다([그림 4.3]의 ④, ⑤, ⑦, ⑧, ⑨번 마디). 의사결정나무에서의 분류규칙은 끝마디의 개수만큼 생성된다고 할 수 있다. 예를 들어, 월소득이 213만원 이하이면서 직업이 C, D, E, F인 경우에는 신용상태가 나쁘다고 평가 할 수 있으며(④번 마디), 월소득이 213만원 이상이면서 연령이 25세 이상인 경우에는 신용상태가 좋다고 평가할 수 있다(⑦번 마디).

l 중간마디(internal node) : 나무구조의 중간에 있는 끝마디가 아닌 마디들을 의미한다([그림 4.3]의 ②, ③, ⑥번 마디).

l 가지(branch) : 하나의 마디로부터 끝마디까지 연결된 일련의 마디들을 의미하며, 이 때 가지를 이루고 있는 마디의 개수를 깊이(depth)라고 한다. [그림 4.3]의 ③, ⑥, ⑦, ⑧, ⑨는 하나의 가지를 이루고 있으며, 이 가지의 깊이는 3이다. [1]

4.3.2 의사결정나무의 형성과정

의사결정나무의 형성과정은 [그림 4.4]와 같이 4단계로 구성된다.

[그림 4.4] 의사결정나무 형성과정

각 단계별로 살펴보면 다음과 같다.

l 의사결정 나무 구축 : 분석의 목적과 자료구조에 따라서 적절한 분리 기준과 정지규칙을 지정하여 의사결정 나무를 형성한다.

l 가지치기 : 가지를 너무 구체적으로 분할하였을 경우 전체적인 나무의 규모가 커지고 분할효율이 떨어지는 문제점이 발생한다. 너무 세분화된 의사결정나무의 경우 새로운 자료에 대해서 반드시 좋은 결과를 가져오지는 않는다. 이러한 문제를 해결하기 위해서는 적당한 크기의 의사결정나무를 설정하는 것이 중요하다. 이 단계에서 분류오류를 크게 할 위험이 높거나 부적절한 추론규칙을 가지고 있는 가지를 제거해야 한다. 즉, 의사결정 나무를 생성하는 과정에서 적당한 종료 마디 수를 갖게 되면 마디의 분할을 정지하거나 가지가 많은 나무로부터 분류 성능 향상에 큰 도움이 되지 못하는 가지를 잘라내어야 한

다.

- 타당성 평가 : 적당한 크기의 의사결정 나무를 설정한 뒤, 이익도표나 위험도포를 사용하거나 검증용 자료에 의한 교차 타당성 등을 이용하여 의사결정나무를 평가한다.
- 해석 및 예측 : 생성된 의사결정 나무를 규칙에 따라 해석하고 예측모형을 설정한다. [11]

위의 과정은 분리기준, 정지 규칙, 평가기준 등을 어떻게 지정하느냐에 따라서 서로 다른 의사결정나무가 형성된다.

먼저 분리기준은 하나의 부모마디로부터 자식마디들이 형성될 때, 입력변수(input variable)의 선택과 범주(category)의 병합이 이루어질 기준을 의미한다. 즉, 어떤 입력변수를 이용하여 어떻게 분리하는 것이 목표변수의 분포를 가장 잘 구별해 주는지를 파악하여 자식마디가 형성되는데, 목표변수의 분포를 구별하는 정도를 순수도(purity) 또는 불순도(impurity)에 의해서 측정하는 것이다. 순수도는 일반적으로 목표변수의 측도에 따라서 다음과 같은 분리기준들이 사용된다.

1) 이산형 목표변수에 사용되는 분리기준 : 목표변수가 이산형인 경우에는 목표변수의 각 범주에 속하는 빈도에 기초하여 분리가 일어나며, 이를 분류나무를 구성한다고 한다.
   - 카이제곱 통계량(Chi-Square statistic)의 p-값 : p-값이 가장 작은 예측변수와 그 때의 최적분리에 의해서 자식마를 형성한다.
   - 지니 지수(Gini index) : 불순도를 측정하는 하나의 지수로서 지니 지수를 가장 감소시켜 주는 예측변수와 그 때의 최적분리에 의해서 자식마디를 선태 한다.
   - 엔트로피 지수(Entropy index) : 다항분포에서의 우도비 검정통계

량을 사용하는 것돠 같은 것으로 알려져 있다. 이 지수가 가장 작은 예측변수와 그 때의 최적분리에 의해서 자식마디를 형성한다.

2) 연속형 목표변수에 사용되는 분리기준 : 목표변수가 연속형인 경우에는 목표변수의 평균(mean)에 기초하여 분리가 일어나며, 이를 회귀나무를 구성한다고 한다.

- 분산분석에서의 F 통계량 : p-값이 가장 작은 예측변수와 그 때의 최적분리에 의해서 자식마디가 형성된다.
- 분산의 감소량(Variance reduction) : 예측오차를 최소화하는 것과 동일한 기준으로 분산의 감소량을 최대화하는 기준의 최적분리에 의해서 자식마디가 형성된다.

정지규칙은 더 이상 분리가 일어나지 않고 현재의 마디가 끝마디가 되도록 하는 여러 가지 규칙을 의미하며, 가지치기는 형성된 의사결정나무에서 적절하지 않은 마디를 제거하여, 적당한 크기의 부나무(subtree) 구조를 가지는 의사결정나무를 선택하는 것이다. [1]

4.3.3 의사결정나무의 장점 및 단점

의사결정나무의 장점은 다음과 같이 3가지로 살펴볼 수 있다. 첫째, 나무구조에 의해서 모형이 표현되기 때문에 모형을 사용자가 쉽게 이해할 수 있다는 것이다. 또한 새로운 개체에 대한 분류 도는 예측을 하기 위해서 뿌리마디로부터 끝마디까지를 단순히 따라가면 되기 때문에, 새로운 자료에 모형을 적합 시키기가 매우 쉽다. 그리고 나무구조로부터 어떤 입력변수가 목표변수를 설명하기 위해서 더 중요한지를 쉽게 파악

할 수 있다. 둘째, 의사결정나무는 두 개 이상의 변수가 결합하여 목표변수에 어떻게 영향을 주는지를 쉽게 알 수 있다. 예를 들어, [그림 4.3]에서 월 소득이 많고 적음에 따라서 신용상태에 영향을 주는 변수와 그 형태가 서로 다르다는 것을 알 수 있다. 의사결정나무는 유용한 입력변수나 교호효과 또는 비선형성을 자동적으로 찾아내는 알고리즘이라고 할 수 있다. 셋째, 의사결정나무는 선형성(linearity)이나 정규성(normality) 또는 등분산성(equal variance)등의 가정을 필요로 하지 않는 비모수적인(nonparametric)방법이다. 또한 의사결정나무에서는 순서형 또는 연속형 변수는 단지 순위(rank)만 분석에 영향을 주기 때문에 이상치(outlier)에 민감하지 않다는 장점이 있다.

의사결정나무의 단점으로는 비연속성을 들 수 있다. 실제로 의사결정나무에서는 연속형 변수를 비연속적인 값으로 취급하기 때문에 분리의 경계점 근방에서는 예측 오류가 클 가능성이 있다. 예를 들어, [그림 4.3]에서 어떤 고객의 월 소득이 210만원인 경우에, 월 소득이 213만원보다 적다고 해서 신용상태가 나쁠 것으로 단순히 분류하는 것은 옳지 않을 수 있다. 최근에 이러한 단점을 극복하기 위하여, 앞서 논의한 장점을 해치지 않고 모수적 모형이나 신경망 등을 의사결정나무와 결합하는 방법들이 연구되고 있다. 또한 의사결정나무에서는 선형 또는 주효과 모형에서와 같은 결과를 얻을 수 없다는 한계점이 있다. 뿐만 아니라 분석용 자료(training data)에만 의존하기 때문에 새로운 자료의 예측에서는 불안정할 가능성이 높다. [1]

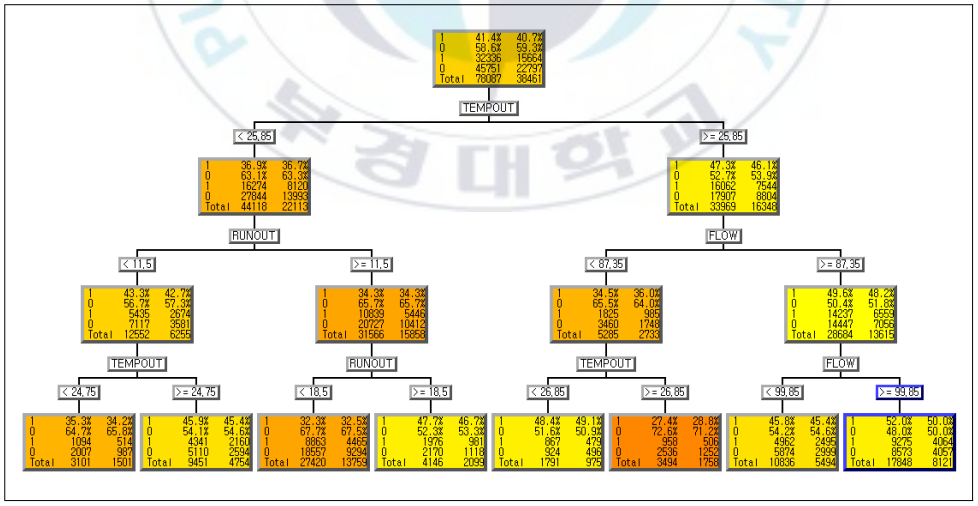
4.3.4 의사결정나무의 적용

이번 장에서는 3장에서 소개된 데이터를 의사결정나무를 이용하여 분석해 보기로 한다. 분석하기에 앞서 반응변수 Y를 이진화 시켜 입력변

수와의 관계를 파악해보았다. 반응변수 Warp의 값이 9보다 작으면 범주 1, 그렇지 않으면 범주 0으로 이진화 시켰으며, 9라는 수치는 M사에서 목표로 하는 값으로 적어도 9보다 작아야 Warp값이 좋다는 의미이다.

입력변수는 군집분석에서와 동일하게 Main roller run-out, Slurry density, Slurry 유량, Slurry 온도 in, Slurry 온도 out으로 두었으며, [그림 4.5]는 의사결정나무분석 결과를 Tree노드로 나타낸 것이다.

Tree노드를 보면 반응변수 Y를 구분 짓는데 영향을 주는 입력변수로는 Slurry 온도 out, Main roller run-out, Slurry 유량이 됨을 알 수 있다. [그림 4.5]에서 분석할 수 있는 한 예로, Slurry 온도 out(TEMPOUT)이 25.85보다 크거나 같고, Slurry 유량(FLOW)이 99.85보다는 크거나 같으면, 52.0%의 반응변수 Y가 범주 1에 속함을 알 수 있다. 이런 방식으로 [그림 4.5]의 결과를 분석해 정리해 놓은 것이 [표 4.2]이다.



[그림 4.5] 의사결정나무분석의 Tree노드 결과

[표 4.2] 의사결정나무 분석 결과

X요인	의사결정나무 분석 결과
Main roller run-out	-
Slurry density	-
Slurry flow	99.85보다 높은 수준
Slurry 온도 in	-
Slurry 온도 out	25.85보다 높은 수준

[표 4.2]를 보면 Slurry flow와 Slurry 온도 out에 대해서는 분석결과가 나왔으나 그 외 Main roller run-out, Slurry density, Slurry 온도 in에 대해서는 의사결정나무를 통해서는 결론을 내릴 수가 없었다.

4.4 결과 비교

이번 장에서는 앞서 수행한 데이터마이닝 기법 중 군집분석과 의사결정나무분석을 비교해 보았다. [표 4.3]은 분석결과를 정리한 것이다.

[표 4.3] 분석결과 비교

X요인	M사 기준	군집분석결과 (평균값)	의사결정나무 분석결과
Main roller run-out	12.2/+1.5 (Max 20)	12.49	-
Slurry density	1.50±0.03	1.50	-
Slurry flow	기계A : 93±2 기계B : 101±1 (Min 80)	96.32	99.85보다 높은 수준
Slurry 온도 in	-	23.01	-
Slurry 온도 out	-	25.81	25.85보다 높은 수준

[표 4.3]에서 M사 기준은 M사에서 실리콘 웨이퍼를 생산할 때 각 X요인들의 생산조건을 나타내고 있다. 3장에서 소개한 데이터 분석결과 일반적인 통계 분석만으로는 각 X요인들이 영향정도와 최적조건을 찾아낼 수 없어 현재 M사에서 웨이퍼를 생산할 때의 X요인의 생산조건과 데이터마이닝 결과를 비교하고자 한다. M사 기준과 비교해볼 때 군집분석의 결과를 살펴보면, Main roller run-out, Slurry density는 M사에서 제시하고 있는 기준과 거의 같다는 것을 알 수 있었다. 또한 Slurry flow의 경우 M사에는 각 기계종류별로 그 기준을 달리하여 군집분석의 결과와는 차이가 있었으나 Slurry flow의 최소값인 80보다 크고 기계 B에 적용하고 있는 102보다는 작은 값을 나타냈다. Slurry 온도 in과 Slurry 온도 out의 경우는 그 기준을 나타내지 않아서 비교할 수 없었다.

의사결정나무를 이용하여 분석한 결과를 살펴보면, 반응변수에 영향

을 주는 X요인이 Main roller run-out, Slurry 유량, Slurry 온도 out으로 나타났으며 이 중 Slurry flow, Slurry 온도 out에 대한 결과만 분석할 수 있었다. Slurry flow의 경우 99.85보다 높은 수준을 요구했으며, Slurry 온도 out은 25.85보다 높은 수준일 경우 반응변수에 좋은 영향을 준다고 분석되었다. 이는 군집분석결과와도 비슷한 결론이었으며, 또한 Slurry flow의 경우 M사의 기준인 최소값 80보다는 크며, 102보다는 작다는 것을 알 수 있었다.

본 연구를 살펴볼 때, 데이터마이닝 분석을 통하여 다음과 같이 두 가지 면에서 그 의의를 살펴볼 수 있다.

첫째, 기존 공정에서 제품의 생산조건이 정해져 있을 경우, 그 조건이 맞는지 데이터마이닝을 통하여 확인해 볼 수 있다. 군집분석의 경우 반응변수 Warp의 값이 가장 작은 군집의 데이터 패턴을 조사해봄으로써 X요인들의 패턴을 확인할 수 있었다.

둘째, 제품의 생산조건이 정해져있지 않을 경우, 데이터마이닝을 통하여 제품의 생산조건을 설정할 수 있다. 즉, DOE 등과 같은 추가실험을 하지 않고도 생산조건을 설정할 수 있으며, DOE 실험에 따른 추가 비용과 시간도 최소화 할 수 있다. 본 연구의 경우, M사에서 제시하고 있지 않은 Slurry 온도 in, Slurry 온도 out의 생산조건을 데이터마이닝을 통해 분석된 결과를 이용하여 설정할 수 있을 것으로 판단된다.

제 5 장 결론 및 향후 연구과제

정보기술의 발달로 하루에도 엄청난 데이터가 축적되고 있다. 이렇게 축적된 대용량의 데이터를 이용하여 공정개선을 위한 분석을 할 경우, 불필요한 정보 및 잡음요인으로 올바른 분석을 하지 못할 수도 있다. 이러한 경우 공정에 대한 기술적인 지식 및 경험을 활용하여 분석의 기본방향을 정하고 데이터를 분석한다면 보다 의미 있는 정보를 도출할 수 있다. 또한 대용량의 데이터를 효과적으로 분석해 주는 한 방법인 데이터마이닝을 이용한다면 이미 집적되어있는 데이터로부터 유용한 정보를 도출할 수 있다.

본 논문에서는 일반적인 통계적 분석과 공정에 대한 기술적인 지식을 활용하여 대용량 데이터를 분석한 사례를 소개하였으며, 또한 데이터마이닝 기법 중 군집분석과 의사결정나무를 이용하여 분석해보고 그 결과를 비교하였다. 그 결과 통계적인 분석과 기술적인 지식을 활용하여 반응변수 Warp의 불량률을 개선할 수 있었으며, 데이터마이닝의 경우 M사에서 제시한 X요인의 기준과 비슷한 수준을 도출할 수 있었다. 특히 데이터마이닝의 경우 다음과 같이 두 가지 면에서 그 의의를 살펴볼 수 있다. 첫째, 기존 공정에서 제품의 생산조건이 정해져 있을 경우, 그 조건이 맞는지 데이터마이닝을 통하여 확인해 볼 수 있다. 둘째, 새로운 제품을 생산하고자 할 경우, 데이터마이닝을 통하여 제품의 생산조건을 설정할 수 있다.

데이터마이닝은 현실적인 유용성으로 인하여 현재에도 계속 연구가 되고 있으며, 적용 분야 또한 영업/판매 부분에서 제조부분까지 확대되고 있다. 또한 군집분석, 의사결정나무뿐만이 아니라 신경망분석, 회귀분석, 시장바구니분석 등 여러 가지 기법이 존재한다. 따라서 본 논문에서

는 군집분석과 의사결정나무 분석에 대한 사례만 소개하였으나 데이터마이닝의 다른 기법의 적용사례가 추가적으로 연구되어야 할 것이다.

PUKYONG NATIONAL UNIVERSITY
부경대학교

참 고 문 헌

[1] 강현철, 한상태, 최종후, 김은석, 김미경(2001), SAS Enterprise Miner 4.0을 이용한 데이터마이닝-방법론 및 활용-, 제3판, 자유아카데미

[2] 김만선(2005), 대용량 다차원 데이터의 패턴 분류를 위한 다중 SVM 분류기, 석사학위 논문, 공주대학교

[3] 김미정(2004), Six Sigma 분석 및 개선을 위한 데이터마이닝 기법의 이용, 석사학위 논문, 부경대학교

[4] 김영아(2004), 시장세분화를 위한 데이터마이닝 응용에 관한 연구 : 백화점 사례를 중심으로, 석사학위 논문, 서강대학교

[5] 김현민(2006), 데이터마이닝을 활용한 비즈니스 프로세스 관리시스템 구축에 관한 연구, 석사학위 논문, 전주대학교

[6] 권근섭, 최경현(2004), 교호효과를 고려한 향상된 의사결정나무 알고리듬에 관한 연구, 대한산업공학회지, 제30권, 제4호, pp.267-276

[7] 박재홍, 변재현(2002), 데이터 정제와 그래프 분석을 이용한 대용량 공정데이터 분석 방법, 품질경영학회지, 제30권, 제2호, pp.72-85

[8] 백동현, 한창희(2003), 데이터마이닝을 이용한 반도체 FAB공정의 수율개선 및 예측, 한국지능정보시스템학회논문지, 제9권, 제1호, pp.157-177

[9] 신선영(2005), 데이터마이닝을 이용한 제약회사 이탈고객 분석예측에 관한 연구, 석사학위 논문, 이화여자대학교

[10] 이승영(2003), 데이터마이닝을 이용한 데이터 활용에 관한 연구, 석사학위논문, 연세대학교

[11] 양승정, 이종태(2004), 퍼지이론을 이용한 의사결정나무의 분할개선

방법, 대한산업공학회, 추계학술대회논문집, pp.44-51

[12] 정영수, 강창욱, 변성규(2005), 데이터마이닝을 이용한 공정변수 확인 및 공정개선, 한국산업경영시스템학회지, 제28권, 제3호, pp.166-171

[13] 정철화(2004), 한국적 품질보전 적용사례-반도체 웨이퍼 생산업체 L사, FA저널, 11월호, pp.60-64

[14] 조혜정(2002), 고객세분화를 위한 데이터마이닝 기법 비교, 석사학위논문, 동아대학교

[15] 최종후, 한상태, 강현철, 김은석, 김미경, 이성건(2001), SAS Enterprise Miner 4.0을 이용한 데이터마이닝-기능과 사용법-, 제3판, 자유아카데미

[16] 캄버, 한(2003), 데이터 마이닝: 개념 및 기법, 자유아카데미

[17] Eugene L. Grant, Richard S. Leavenworth(1988), Statistical Quality control, McGRAW-HILL

[18] Jun Sugawara, Hiroshi Hara, Akira Mizoguchi(2004), Development of fixed-abrasive-grain wire saw with less cutting loss, SEI GECHNICAL REVIEW, NO.58, pp.7-11

[19] Ken-ichi ISHIKAWA, Hitoshi SUWABE, Shun-ichi Itoh, Michio UNEDA(2003), A Basic Study on Behavior of Slurry Actions at Multi-Wire Saw, Key Engineering Materials, Vols.238-239, pp.89-92

[20] Wayne A. Taylor(1991), Optimization and Variation Reduction in Quality, McGRAW-HILL

[21] http://database.sarang.net/database/datamining/outline/outline_back.htm

[22] http://dmlab.uos.ac.kr/dmlab/dmintro.files/frame.htm

[23] http://home.inje.ac.kr/~rihs/mining/mining.html

[24] http://myhome.naver.com/tongseob/menu1.html

Case study on the application of data mining and large-scale data analysis for process improvement

You-Jin Kwon

Department of Systems Management & Engineering,
The Graduate School,Pukyong National University

Abstract

Due to the recent development of information technology, a large amount of process data may be collected directly from shop-floor in real-time. For the purpose of process improvement, it may be useful to utilize the information derived from those data since it may relieve the burden of additional data collection. However, it is quite difficult to extract meaningful information from those massive data mainly because of the presence of disturbing noise factors. Engineering knowledge and technical experiences may be useful to identify those noise factors and analyze the data effectively. In addition, a variety of data mining techniques, including clustering and decision tree analysis, may be helpful to better understand the nature and characteristics of processes. The main purpose of this thesis is to outline analysis procedures for process improvement utilizing engineering knowledge combined with data-mining techniques. The usefulness and applicability of proposed procedures are demonstrated through the case study from a semi-conductor manufacturing process. More specifically, clustering and decision tree analysis are employed to analyze the process data.

keyword: Data Mining, Clustering, Decision tree analysis

감사의 글

대학원에 진학한지 벌써 2년이라는 시간이 흘렀습니다. 그동안 힘들고 어려운 일들도 많았지만 그만큼 기쁜 일 보람된 일 또한 많아, 진학당시 '과연 내가 잘 해 낼 수 있을까' 하는 걱정이 지금은 '역시 대학원에 진학하길 잘했다.'는 생각으로 바뀌었습니다. 먼저 이렇게 대학원을 졸업하게 곁에서 믿고 공부시켜주신 부모님께 감사드립니다. 딸의 일이라면 전적으로 믿고 지원해주시는 아버지, 항상 저의 건강을 챙겨주신 어머니, 그리고 평소 맛있는 것도 사주고 용돈도 주면서 많은 격려와 충고를 해준 오빠, 감사합니다.

학부시절부터 지도교수님으로 많은 관심과 조언을 해주신 권혁무 교수님 진심으로 감사합니다. 부족한 점이 많아 교수님의 가르침을 제대로 소화해 내지 못하고 꾸중을 들었던 적도 많았는데, 그렇게 꾸중 들으면서 배웠던 것들이 저에게 큰 디딤돌이 되었던 것 같습니다. 사회에 나가서도 교수님의 가르침, 항상 기억하겠습니다. 그리고 마지막 학기동안 저의 논문을 지도해 주신 김영진 교수님 감사합니다. 혹시나 논문 때문에 졸업하지 못할까봐(^^) 걱정해 주신 점, 그리고 논문뿐만이 아니라 취업문제 등 여러 가지 고민이 있을 때마다 진심어린 조언을 해주신 점, 정말 감사합니다. 저뿐만이 아니라 대학원생들 모두에게, 필요한 것이나 대학원생활을 하기에 어려운 점이 없냐는 말씀으로 항상 관심을 가져주신 고시근 교수님 감사합니다. 교수님이 계셔서 저희 대학원생들은 항상 마음이 든든했습니다. 그리고 저의 인사를 반갑게 맞아주시는 옥영석 교수님, 마주칠 때마다 정겹게 해주시는 한마디 한마디가 대학원 생활을 하는데 많은 힘이 되었습니다. 다만, 수업시간에 많이 졸아서 교수님을 당황하게 해드린 점(ㅠ_ㅠ)은 대단히 죄송합니다. 항상 노래를 흥얼거리시는 듯 즐거운 표정으로 대해주시는 구평회 교수님, 수업시간에 배운

내용 중 모르는 부분이 있어 찾아가도 귀찮으신 내색 없이 정성껏 가르쳐 주셔서 감사합니다. 멋쟁이라는 말이 너무 잘 어울리시는 이운식 교수님, 마주칠 때마다 저의 논문 진행상황을 물어봐 주시는 등 관심을 가져주셔서 감사합니다. 그리고 정도 많으시고 잘 웃으시는 김민수 교수님, 대학원 때 처음 교수님 수업을 들으면서 잘 적응하지 못하고 교수님을 귀찮게 해드렸는데 이해해주시고 조언도 해 주신 점 감사합니다. 수업시간에 열정으로 모든 학생들을 동참하게 만드시는 김병남 교수님, 교수님의 열정 본받겠습니다. 점잖으시고 침착하신 오수철 교수님, 옆방에서 떠드는 소리로 시끄럽게 해드린 점 대단히 죄송합니다. 마지막으로 기회가 없어서 수업은 받아보지 못했지만 마주칠 때마다 저의 인사를 따뜻하게 받아주시는 박병무 교수님 감사합니다.

웃음, 애교, 그리고 정이 많아서 친구를 즐겁게 해주고 잘 챙겨주는 아라.. 항상 신경써줘서 고마워. 초등학교 시절부터 곁에서 항상 힘이 되어준 소현이.. 지금은 서울에 있어서 자주 보지는 못하지만 생일 때마다 잊지 않고 챙겨줘서 너무 고마워~~~ 먼저 취업했다는 이유로 밥도 자주 사주고 간식도 사주면서 힘든 일이 있을 때마다 위로해준 경심이, 학부시절 곁에 없으면 불안할 정도로 붙어 다녔던 단짝 은희, 동기이자 지금은 조교 선생님으로 대학원 생활을 하는데 많은 도움을 준 혜란이, 어디를 가든지 웃음이 끊이지 않는 정미, 세심하게 배려해주는 수렴이, 믿음직한 제욱이, 너희들이 있어서 대학교생활을 즐겁게 보낸 것 같아서 고맙다. 그리고 언제나 믿고 의지할 수 있는 연이, 애교 많고 착한 태미, 바쁜 와중에도 먼저 연락해주는 영희, 너무 편한 친구들..고마워~~~

좋은 일이 있을 때는 누구보다 기뻐해주고 안 좋은 일이 있을 때는 자기 일처럼 걱정해주는 믿음직한 중완이.. 힘들 때 마다 진심어린 마음으로 위로와 격려를 해주고 대학원생활을 잘 할 수 있게끔 도와줘서 항상 고맙게 생각해^^

대학원생들의 정신적 지주이자 물질적 지주인 진아언니^^ 정말 친언니처럼 챙겨줘서 고맙습니다. 무엇이든지 열심히 하는 지현언니, 언제나 든든한 철호오빠, 구박 하는 듯 하면서도 챙겨주는 원일선배, 전공공부에 힘들어할 때마다 옆에서 도와준 정수선배, 비록 남자지만 언니처럼 편한 찬호선배, 학부시절부터 동기인 착한 희영이, 그리고 박사과정에 계신 황영제님과 박국제님, 감사합니다. 품질연구실 선배, 동기, 후배들.. 도움을 주기보다는 받는 쪽이어서 미안하면서도 고맙게 생각하고 있습니다. 마지막으로 학과사무실에서 언제나 따뜻한 차 한 잔을 대접해 주면서 푸근하게 맞아주었던 윤경순 선생님과 조민정 선생님에게도 감사의 말을 전하고 싶습니다.

권 유 진 드림

PUKYONG NATIONAL UNIVERSITY 부경대학교