데이터 기술: 지식창조를 위한 새로운 융합과학기술

박성현†

한국연구재단 기초연구본부

Data Technology: New Interdisciplinary Science & Technology

Park Sung Hyun†

Directorate for Basic Research in Science and Engineering, National Research Foundation of Korea

Key Words : Statistics, Data Technology, Science & Technology, Knowledge Society, Information Technology, Six Sigma, Quality Management

Abstract

Data Technology (DT) is a new technology which deals with data collection, data analysis, information generation from data, knowledge generation from modelling and future prediction. DT is a newly emerged interdisciplinary science & technology in this 21st century knowledge society. Even though the main body of DT is applied statistics, it also contains management information system (MIS), quality management, process system analysis and so on. Therefore, it is an interdisciplinary science and technology of statistics, management science, industrial engineering, computer science and social science. In this paper, first of all, the definition of DT is given, and then the effects and the basic properties of DT, the differences between IT and DT, the 6 step process for DT application, and a DT example are provided. Finally, the relationship among DT, e-Statistics and Data Mining is explained, and the direction of DT development is proposed.

1. 서 론

통계학은 과학인가, 아니면 기술인가? 이러한 질문은 최근 학자들 간에서 자주 토론의 과제가 되고 있다. 통계학은 크게 이론 통계학과 응용 통계학으로 나눌 수 있으며, 일부 학자들의 편협한 사고로 인하여 오늘날 통계학이 상당히 위축되어 가고 있다고 생각한다. 일부 이론 통계학자들의 견해는 통계학은 수학에 가까운 과학(Science)이며, 이론 통계에 대한 연구논문이 가장 중요하며, 사회경제 발전을 위한 통계 응용은 부수적인 문제로 별로 중요하지 않다고 보고 있다. 한편 일부의 응용 통계학자들은 통계의 응용 없이는 이론연구도 의미가 없다고 주장하면서 통계 패키지를 잘 쓰는 것이 가장 중요하며, 통계학을 패키지에 의하여 데이터 분석을 다루는 시스템적인 기술(Technology)로 보는 견해가 강하다. 이 두 가지 견해는 극단적인 견해로, 오늘날 대부분의 통계학자들이 동의하는 것은, 통계학은 통계적 방법론에 근거한 의사결정과학(Decision-making Science)인 동시에, 지식정보화 사회에서 데이터로부터 지식을 창출하는 기술(Knowledge generating technology)이라고 보고 있다. 또한 통계학의 발전을 위해서는 이론연구와 응용연구가 공존하여야 하며, 이론연구는 응용을 염두에 두어야 하고, 응용연구는 이론적인 배경을 가져야 한다는 것이다.

통계학이 과학이냐, 기술이냐의 본질에 관한 논쟁은 오래전부터 있어온 것으로 Healy (1978), Friedman (2001), Straf (2003) 등에서도 찾아 볼 수 있다. 이들의 견해는 통계학이 발전하기 위해서는 과학에 안주하지 말고, 기술로 발전되어야 하며, 과학으로 무장한 기술이 바람직하다는 것이다. 즉, 데이터를 다루는 과학에서 지식을 창출하는 기술로 발전되어야 통계학이 미래 지향적으로 발전할 수 있다는 것이다. 통계학을 볼 때

† 교신저자 parksh@snu.ac.kr

과학과 기술의 비중이 과거에는 6:4 정도이었으나 최근에는 과학과 기술의 비중이 4:6으로 기술에 더 무게를 두고 있다고 판단되며, 이는 지식사회에서 지식창출 기술을 제공하는 학문에 대한 수요가 늘어나기 때문으로 보인다.

2. 데이터기술이란?

데이터기술(DT)이란 데이터의 측정, 수집, 축적 기술에서부터 시작하여, 데이터의 전송, 분석 및 해석 능력, 데이터로부터 정보와 지식을 창출하는 기술, 데이터로부터의 통계적 모형화 기술, 데이터로부터 미래를 예측하는 기술 등을 다루는 전반적인 과학기술적 방법론을 말한다. DT는 데이터의 취급, 소프트웨어의 구축, 모형화 및 미래 예측 기술을 주로 다루고 있으므로, 그 진행과 결과가 눈에 잘 띄지 않으며, 보통 간과하기 쉽다. 그러나 국가 선진화와 통계학의 발전을 위해서는 DT는 필수적이라고 생각한다. 우리나라는 OECD 국가 중에서 DT 분야가 낙후되어 있으며, 조만간 크게 보완되지 못하면 국가 경쟁력에 큰 타격을 줄 것으로 생각한다. 데이터 기술이란 용어는 처음으로 박성현(2001a, 2001b)에 의하여 언급되었으며, 통계학 응용으로서의 DT의 발전방향이 Park과 Suh(2008)에 의하여 언급되었고, 그 하나의 적용 사례가 Erto, Pallotta와 Park (2008)에 의하여 제시되었다. 앞으로 DT의 개념을 이용한 적용 사례는 수없이 많이 나타낼 것이다.

정부에서 발표하는 각종의 과학기술 관련 계획에 의하면 우리나라가 향후 발전시켜야할 첨단과학기술 분야로 IT(정보기술), BT(생명공학), NT(나노기술), ST (항공우주기술), ET(환경기술), CT(문화기술)의 소위 “6T"를 들고 있고, 최근에는 이들 간의 융합기술 (예로, IT + BT, IT + ET 등)의 발전을 많이 언급하고 있다. 지식기반 정보화 사회에서 과학기술의 선진화, 국가경쟁력 제고, 전 국민 과학화 등을 위하여 반드시 포함시켜야할 분야로 데이터로부터 시작되는 DT가 매우 중요하며, ”6T"대신에 DT를 포함하여 “7T”가 첨단과학기술로 언급되어야 할 것이다.

3. DT의 발전으로 인한 효과

지식기반 정보화 사회는 데이터 홍수의 시대라고 할 수 있으며, 이러한 데이터로부터 필요한 정보를 순발력 있게 정확하게 추출할 수 있는 능력은 매우 중요하며, 또한 얻어진 데이터로부터 어떤 현상을 예측하기 위한 모델링과 이로부터의 예측은 과학기술의 발전 측면에서도 필요하다. 이러한 기능을 다루는 분야가 DT이며, 21세기에 매우 중요한 복합적인 학문분야이다. DT의 미비로 발생할 수 있는 국가적 손실은 매우 크며, 역으로 DT의 발전에 의한 효과도 엄청나다. 몇 가지 예를 들어보자.

3.1 국가경제 지표의 과학적 관리 운영

첫째로, 국가경제를 다루는 경제지표의 과학적인 관리가 안 될 때 발생하는 국가적 손실을 예방할 수 있다. 우리나라는 1997년 외환위기 때, 외환보유고를 포함한 각종의 경제지표의 변화를 소홀히 생각하는 가운데 스스로 위기를 자초하게 되었다. 외환보유고와 관련이 있는 데이터의 적절한 수집, 정리 및 분석을 통하여 외환보유고에 대한 예측 모델을 만들고, 이를 통하여 외환보유고의 변화를 예측할 수 있었다면 IMF 위기를 사전에 대비할 수 있었을 것이다.

3.2 품질과 생산성의 최적화

둘째로, 산업에서는 많은 종류의 최적화 문제가 따른다. 품질 최적화, 생산성 최적화 등은 필수적인 요소이며, 품질과 생산성은 모두 많은 변수들의 지배를 받는다. 영향을 주는 변수들의 최적조건을 찾아서 운영하는 것은 산업 경쟁력을 위하여 필수적인 요소이다. 예를 들면, 제품 품질은 생산과정에서의 많은 운전변수들 (operating variables)의 영향을 받는다. 운전변수들의 최적 조건을 찾아주는 것은 품질고급화의 선결과제이다. DT는 최적 운전변수들의 조건을 찾는데 유용하게 사용될 수 있다.

3.3 품질비용의 최소화

세 번째로 산업에서 불량품 발생 등으로 인한 품질비용(quality cost)은 매출액의 20-30% 수준에 이른다는 보고가 있다. 이 품질비용 중에서 예방비용, 평가비용, 내부 실패비용, 외부 실패비용은 각각 어느 정도인지 객관적 데이터로 평가한 후에, 필요한 데이터의 수집, 분석, 평가, 예측을 통하여 품질비용은 최소화하는 방안을 강구하고 실행한다면, 품질비용은 매출액 대비 10%

수준으로 충분히 낮출 수 있다고 한다. 그러나 대부분의 우리 기업들은 아직도 품질비용을 제대로 계산하지 못하고 있다. DT의 적절한 활용은 적자기업을 흑자 기업으로 바꾸는 중요한 처방이 될 수 있다.

3.4 정책 결정의 합리성 추가

네 번째로 정부의 각종 공공기관에서 정책결정을 할 때에, 현상을 정확히 파악할 수 있는 객관적 현황 데이터나 과학적 분석결과가 있다면, 가장 합리적인 결정을 내릴 수 있을 것이다. 그러나 데이터기술의 부족으로 이러한 정책결정이 제대로 안 이루어지는 사례가 허다하며, 이로 인하여 국민에 주는 손실금액을 막대할 것이다.

3.5 의료산업, 질병관리 등을 위한 예측 모델의 개발과 활용

다섯 번째로, 의료산업이나 질병관리에서 개개인에게도 DT는 유용하게 사용될 수 있다. 예를 들면, 당뇨병이란 결과는 당뇨병의 원인이 되는 여러 가지 원인변수들(예로, 개인의 생활습관, 개인의 음식섭취 습관, 유전적 특성, 체질적 특이성, 운동량 등)에 영향을 받는다. 이런 원인변수들을 반영한 예측모델을 각종의 통계적 방법론(회귀분석, 데이터 마이닝, 자료 분석기법 등)을 이용하여 개발할 수 있다면, 개인의 건강을 위한 맞춤의학을 실현할 수 있을 것이다.

3.6 타 학문분야의 선진화에 기여

여섯 번째로, 기존의 학문분야의 선진화나 6T의 첨단 과학기술의 발전에 DT는 매우 중요한 인프라 역할을 할 수 있다. 각 학문분야나 융합과학에서 사전에 인과관계 데이터를 표본을 활용하여 신뢰성 있는 측정으로 수집 정리한 후 통계적인 S/W를 활용하여 사전 정보화하는 것은 매우 시급한 과제이다. 따라서 각 학문분야나 6T와 연계된 DT를 분야별로 도입 적용한다면 시너지 효과가 극대화 될 것이다.

3.7 경영정보의 선진화에 기여

마지막으로, 지금은 과히 데이터 홍수시대라고 말할 수 있다. 예를 들어, 수백만 명의 고객을 가지고 있는 회사에서 고객 관련 데이터의 순발력 있는 적절한 처리로 고객만족을 도모할 수 있다면 이 회사는 선진기업이라고 말 할 수 있을 것이다. CRM(고객관계경영), SCM(공급사슬관리), SPC(통계적 공정관리), ERP(전사적 자원관리), DBMS(데이터베이스 관리 시스템) 등은 기업 경영에 사용되는 DT 제품이라고 볼 수 있으며, 이처럼 DT의 적절한 활용은 기업의 선진화를 촉진할 수 있을 것이다.

4. DT의 본질

4.1 DT는 소프트웨어적인 인프라

앞에서 언급된 IT, BT, NT, ST, ET CT는 대부분 눈에 보이는 하드웨어적인 기술과 결과물을 생산해 낸다. 그러나 DT는 눈에 보이지 않으며 밑에 깔려 있는 소프트웨어적인 인프라에 해당함으로 보통 때 무시하기 쉽다. 그러나 데이터에 의하여 현상을 정확히 파악하지 못하고, 문제점을 객관적 데이터에 의하여 찾지 못하고, 앞으로 발생될 현상을 수리적 모형을 사용하여 예측할 수 없다면 첨단과학기술의 발전에 한계가 있고, 국가의 합리적 행정체계, 기업의 경쟁력 강화 등에 어려움이 있을 수밖에 없다.

DT는 모든 첨단과학기술의 기초에 해당하는 원천적 기초과학 기술이다. 또한 개인이든, 기업이든, 공공기관이든 모든 조직에서의 과학적, 합리적 운영에 중요한 인프라이다.

4.2 DT는 융합과학기술

DT의 기본적인 시작은 효율적인 데이터의 수집방법을 연구하고 이를 실행하여 데이터를 축적하고 보관하는 것이다. 이를 위하여 통계적 방법론으로 자주 사용되는 것은 표본설계, 실험계획법 등이다. 컴퓨터과학에서는 데이터베이스(DB)의 연구와 활용이 필요하다. 다음 단계에서 다량의 수집된 데이터로부터 정보를 추출해내는 방법으로 데이터 마이닝(data mining), 뉴럴 넷웍(neural network), 회귀분석(regression analysis) 등이 필요하다. 따라서 통계학과 컴퓨터과학은 기본적으로 관련된 학문이다.

그 다음 단계로 가장 중요한 것은 예측은 위한 모델링 작업으로 여기에서 산업공학, 경영학, 사회과학, 의약학, 경제학 등의 모든 학문의 지식이 바탕을 이루어

야 한다. 모델링은 DT의 핵심이며, 모델링이 이루어지면 현 상황의 분석과 미래의 예측이 가능하여 진다. 이처럼 DT는 많은 학문이 연결된 융합과학기술이라고 볼 수 있다.

4.3 DT는 IT와 다르다

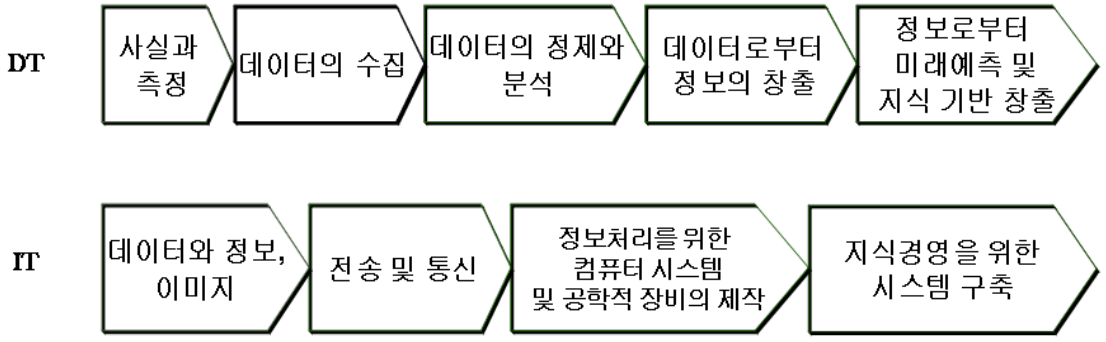
월드컵 축구에서 한국이 4강에 들어갈 수 있었던 근본 원인은 필자의 생각에는 히딩크와 그의 협력자들이 DT를 잘 활용하여 얻은 결과라고 볼 수 있다. 선수들의 강점과 약점을 데이터를 통하여 과학적으로 분석하고, 이를 체계적인 훈련프로그램을 통하여 인내심을 가지고 선수들의 기량을 향상시킨 것이 주요 원인이 아니겠는가? 혹자는 DT가 IT(정보기술)의 일부라고 말하며 IT의 발전은 DT의 자동적인 발전을 가져올 것이라고 말하고 있으나, 이는 잘못된 견해이다. IT와 DT의 차이점을 요약하여 보면 <표 1>과 같다.

DT를 다루는 학문 중에서 응용수학에서는 암호 수학, 금융수학 등이며, 통계학에서는 표본설계, 실험계획, 여론조사, 통계적 공정관리(SPC), 시계열분석, 데이터 마이닝 (data mining) 등이 관계가 있고, 계산과학에서는 수치해석, 시뮬레이션 기법, 최적화 기법 등이 관계가 있다. 전산이나 정보과학에서는 소프트웨어 공학, 뉴럴넷웍 (neural network) 등이 관계가 있으며, 산업공학과 경영과학에서는 품질경영, 고객관계경영 (CRM), 전사적 자원관리(EPR), 시스템 공학적인 접근방법 등이 관계가 있다. 특별히 DT의 발전은 국가 소프트웨어의 발전과 고부가가치 IT 산업의 발전에 심대한 영향을 준다. DT와 IT의 정보 흐름의 차이를 보면 <그림 1>과 같은 차이점이 있다.

DT는 주로 자료의 수집, 통계 수리적 분석에 의한 정보의 창출, 수리적 모델링에 의한 미래 예측, 그리고 지식기반(knowledge base)의 구축 등이 정보의 주요 흐름도이다. 그러나 IT에서는 데이터/정보/이미지 등의

<표 1> IT와 DT의 차이점

분류	IT	DT
관련된 주학문	컴퓨터공학, 전기전자공학, 통신공학, 제어계측공학, 정보공학 등	응용수학, 통계학, 계산과학, 산업공학, 정보과학, 경영학, 전산과학, 융합과학 등
주요 제품	통신장비, 전자장비, 반도체, 가전제품, 무선휴대폰 등의 하드웨어	암호시스템, 생산관리시스템, 통계패키지, DBMS, SPC, SCM, CRM 패키지, Data-mining 패키지, ERP 패키지(SAP-R3, Oracle) 등의 소프트웨어
주요 특징	▪ 주로 눈에 보이는 제품 ▪ 정보(글, 그림, 소리 등)를 전달하는 공학적 기술이나 제품.	▪ 주로 눈에 안 보이는 제품 ▪ 다량의 데이터로부터 현상을 파악하고 층별하여 효율성 극대화 하며, 미래를 예측하는 과학적기술
우리나라 수준	상	중하



<그림 1> DT와 IT의 정보 흐름도

공학적 전송이나 통신이 주요 관심사이며, 이를 위한 컴퓨터 시스템 구축, 장비의 제작 등이 다음 단계의 활동이며, 궁극적으로 기업에서는 지식경영을 위한 시스템 구축을 목적으로 하고 있다.

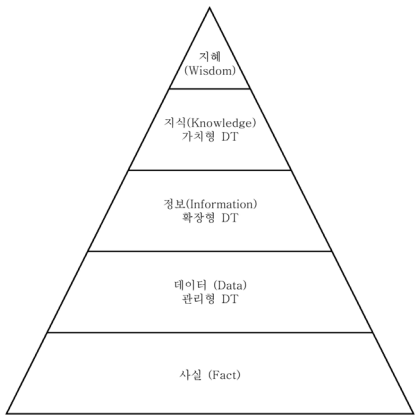
우리나라의 주력 수출제품을 보면 반도체, 조선, 자동차, 휴대폰 등 눈에 보이는 제품이 대부분이다. 눈에 잘 보이지 않는 소프트한 제품들은 국제 경쟁력이 미약하다. 예를 들면, 통계분석용 소프트웨어는 SAS, SPSS, Minitab 등 미국제품이 국내 시장을 석권하고 있으며, 이로 인한 외화의 유출은 엄청나다. 심지어 반도체, 조선 등에 사용되는 공정관리용 소프트웨어도 외국제품이다. 우리나라에 수없이 많이 들어와 있는 외국 컨설팅 회사들이 주로 하는 일이 사실상 DT와 관련된 경영자문이 대부분이다. 이제 우리산업도 DT에 더욱 눈을 떠서 고부가가치 산업에 투자할 때이다. DT의 발전은 21세기에 7대 지식강국이 되려는 우리나라에게 매우 중요하며, 국가 선진화에 핵심적인 요소가 될 것이다.

4.4 DT는 지식창출의 원동력

21세기 지식사회에서는 창의적 지식창출이 국가경쟁력의 원천이라고 말한다. 지식의 창출과정을 살펴보면 <그림 2>와 같은 피라미드 그림(참조: Park(2003))이 얻어 진다. 먼저 우리 주위의 사실(현상)을 정확히 파악하기 위하여 사실을 측정하는 계량화된 데이터가 필요하다. 이 단계의 DT는 관리형(management) DT라고 볼 수 있다. 다음으로 데이터로부터 분석 및 해석을 통하여 정보를 얻는다. 이 단계는 집행형(execution) DT의 단계이다. 이 정보들을 여러 가지 형태로 가공하여 필요한 지식을 얻게 되며, 이 단계는 가치 창출형(valuation) DT의 단계이다. 이러한 지식창출과정은 DT의 도움 없이는 불가능하다. 현재 우리는 정보사회에서 지식사회에 돌입하고 있으며, 앞으로 먼 훗날에는 지혜의 사회에 이르는 인류가 될 수 있을 것으로 생각한다.

4.5 DT는 품질경영 혁신전략의 요소

품질경영 혁신전략으로 흔히 사용되는 TQM(전사적 품질경영), 6 시그마 등은 개선활동에서 Define, Measure, Analyze 등의 단계를 거친다. 이러한 단계는 DT의 단계와 맥을 같이 한다. DT도 품질경영 혁신전략으로 사용될 수 있으며, DT의 활용은 품질경쟁력을 높이는 중요한 도구가 될 수 있다.



<그림 2> 지식의 창출과정 피라미드와 DT의 흐름

5. DT 적용 프로세스

DT의 정의에서 볼 때 DT의 활용 사이클은 DMAMPV (Define, Measure, Analyze, Model, Predict, Verify)로 볼 수 있다. 이 사이클은 6 시그마에서 다루는 DMAIC (Define, Measure, Analyze, Improve, Control)와 유사한 점이 있으나, DT에서는 모델 구축과 미래 예측을 강조하고 있다. 각 단계의 자세한 활동 내용은 다음과 같으며, 이들 단계의 적용 사례는 다음 절에서 찾을 수 있다.

Define(정의) 단계:
- DT와 관련된 프로젝트 선정(배경 및 당위성 설명)
- 프로젝트 정의(목표와 범위 등 설정)
- 프로젝트 승인(실행계획을 수립하고 상급 책임자로부터 승인 받음)

Measure(측정) 단계:
- DT와 관련된 y들의 확인(프로젝트의 결과변수에 대한 구체적 지표 선정)
- 잠재 원인변수(x들)의 발굴과 우선 순위화
- x와 y들에 대한 측정과 측정시스템 분석

Analyze(분석) 단계:
- 데이터 분석(몇 개의 중요(VitalFew) x를 확인하기 위한 통계적 분석)
- y에 영향을 주는 중요 x의 선정 (분석 결과 검토 및 개선 우선 순위화)
- 기타 중요한 정보의 추출

Model(모형화) 단계:
- y와 x 간의 함수관계를 설명하는 모형의 선정
- 통계적 함수추정으로 모형화
- 모형의 통계적 적합성 검토

Predict(예측) 단계:
- 주어진 모형으로부터 임의의 x의 값에서 y에 관한 예측 실시
- 예측값의 통계적 적절성 검토
- 예측값의 활용

Verify(검증) 단계:
- 적정한 y값을 유지하기 위한 x값 들에 대한 관리
- 모형의 적절성 검토
- 관리계획 수립과 문서화

6. DT 적용 사례

6.1 단계 1: Define (정의)

모 정유공장에서 프로젝트팀 활동으로 실시한 품질 개선 사례이다. 이 공장에서는 케로신 스위트닝(kerosene sweetening) 공정 제품인 U-1700 SW-kerosene의 색상품질을 I-MR(개개의 측정치와 이동범위, individual data-moving range) 관리도로 평소에 관리해 주고 있으나, 색상의 평균치가 만족스럽지 못하고 또한 산포도 큰 편이다. 이 색상품질의 평균치를 향상시키는 연구를 실시하였다.

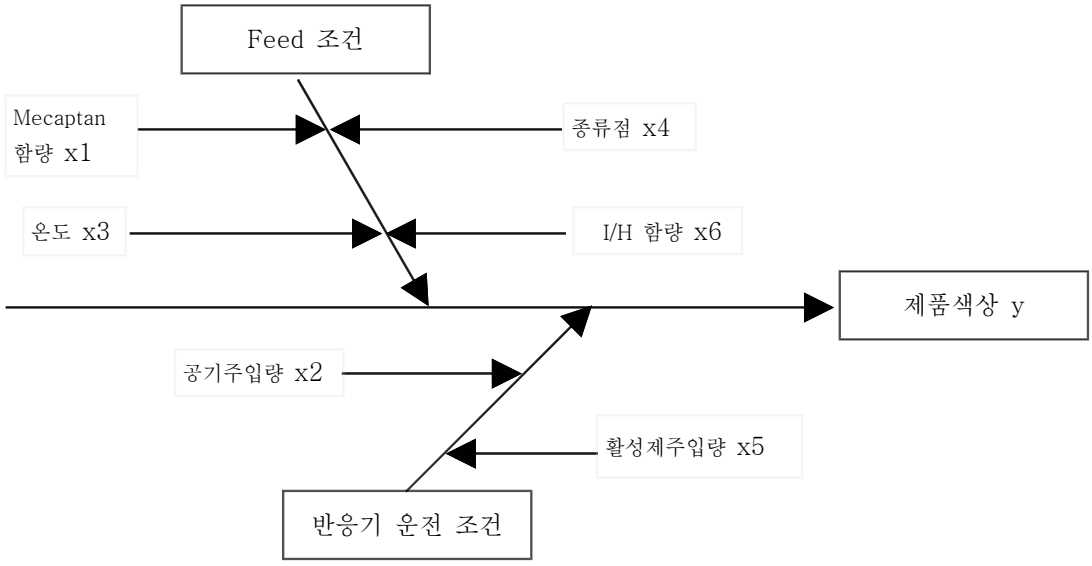
먼저 원인규명을 위하여 관련된 기술자, 연구원 등 팀이 모여서 BS (brain-storming)을 실시한 결과 색상품질을 나타내는 y로 sweetening 공정의 반응기 출구의 색상을 나타내는 세이볼트 값(saybolt number)을 잡아주기로 하였다. 이 y에 영향을 주는 x로 10여개의 변수를 거론하고 토론하였으나, 과거 기술적인 측면이나 외국문헌에 의하여 중요한 원인변수(인자)로 <그림 3>에서와 같이 sweetening 공정으로 들어가는 피드(feed) 조건으로 4가지, 반응기 운전조건으로 2가지를 선택하였다.

결과변수: y = 반응기 출구에서 샘플링된 제품을 측정하여 얻어지는 색상값으로 saybolt number로 얻어지며, 망대특성(크면 클수록 좋음)임.

원인변수:

<Feed 조건>
x1 = Mecaptan 함량(wt. ppm)
x2 = 공기주입량
x3 = feed 온도(0C)
x4 = feed end-point (종류점) (0C)

<반응기 운전조건>
x5 = 활성제 주입량(wt. ppm)
x6 = 처리 원유 중 I/H 함량(vol. %)



<표 2> 제품색상 관련 원 데이터

<표 2> 제품색상 관련 원 데이터

데이터 번호	x1	x2	x3	x4	x5	x6	y
1	111.6	130	46	247	62	100.0	21
2	91.5	135	50	251	70	94.2	21
3	90.8	135	50	245	70	94.1	21
4	80.8	175	43	247	80	16.3	13
5	82.8	145	46	248	80	22.0	20
6	78.6	145	46	255	80	22.0	20
7	93.2	120	43	254	80	22.0	21
8	93.7	120	38	253	80	16.4	23
9	95.9	120	38	253	80	10.9	21
10	97.6	115	42	256	70	10.8	27
11	100.0	115	42	254	70	10.8	26
12	100.0	115	50	260	50	10.0	26
13	100.0	115	51	264	40	0.9	26
14	100.0	119	47	268	40	0.9	27
15	98.7	119	46	259	40	1.0	28
16	100.0	119	44	268	40	1.4	27
17	94.0	135	45	261	50	2.8	23
18	138.7	160	53	268	70	96.9	17
19	157.0	143	56	264	36	95.1	20
20	160.0	150	55	265	36	94.4	18
21	160.0	137	47	261	65	94.4	21
22	119.0	128	59	274	40	86.2	21
23	116.0	139	53	273	52	13.8	19
24	118.6	139	52	265	52	13.8	17
25	117.7	139	50	270	40	13.8	17
26	117.5	150	50	273	40	13.8	21
27	114.3	173	40	259	40	1.9	18
28	133.2	115	58	270	60	9.6	19
29	129.5	107	53	274	40	9.6	23
30	128.4	128	59	262	40	9.6	21
31	121.4	128	57	268	40	11.4	21
32	150.0	110	51	269	40	2.7	23
33	150.8	110	50	276	40	3.0	23
34	150.0	110	49	275	40	3.0	24
35	106.9	110	49	271	40	3.0	26
36	124.3	120	53	271	40	3.0	20
37	139.2	105	52	274	40	2.2	24
38	139.7	119	53	270	40	1.2	22
39	140.4	119	52	266	40	1.2	22
40	140.4	119	48	276	40	1.2	23
41	138.9	119	50	270	40	1.3	26
42	139.5	119	51	271	40	1.8	23
43	105.9	133	52	260	40	4.9	20
44	110.0	133	52	269	40	3.9	23
45	101.6	133	53	273	40	3.9	22

6.3 단계 3: Analyze (분석)

<표 2>에 있는 데이터에 대하여 우선적으로 각 변수에 대한 평균과 표준편차를 계산하여 보니 다음과 같다. 평균에 비하여 상당히 큰 표준편차를 가지고 있는 변수는 x6가 가장 크고, 그 다음 순위의 변수로는 x5, x1, y, x2, x3, x4의 순서이다.

<표 3> 각 변수의 평균과 표준편차

변수	평균	표준편차	변동계수(%)
x1	117.29	22.81	19.4
x2	128.27	16.31	12.7
x3	49.33	5.08	10.3
x4	264.00	8.81	3.3
x5	50.96	15.61	30.6
x6	23.05	34.13	148.1
y	21.89	3.17	14.5

변수들간에 상관관계가 있으리라고 생각되며, 이러한 관계를 규명하여야만 원인변수가 결과변수에 영향을 주는 관계를 규명할 수 있으므로 상관분석을 실시하여 상관계수를 모두 구하여 보니 <표 4>와 같다.

<표 4>의 결과로부터 제품색상(y)에 상관관계가 높은 원인변수로는 공기주입량(x2), 처리원유 중 I/H 함량(x6), 촉매 활성제 주입량(x5) 등임을 알 수 있다. 또한 원인 변수들 간에도 상관관계가 높은 짝들이 여러 개 존재한다. 예를 들면, x1과 x4, x4와 x5, x1과 x3, x1과 x5, x3와 x4, x3와 x5 등이다.

<표 4> 변수들 간의 상관계수를 나타내는 상관행렬

	x1	x2	x3	x4	x5	x6	y
x1	1.0	-0.175	0.532	0.632	-0.553	0.154	-0.075
x2	-0.175	1.0	-0.077	-0.343	0.251	0.393	-0.748
x3	0.532	-0.077	1.0	0.553	-0.562	0.166	-0.188
x4	0.632	-0.343	0.553	1.0	-0.763	-0.340	0.171
x5	-0.553	0.251	-0.562	-0.763	1.0	0.285	-0.264
x6	0.154	0.393	0.166	-0.340	0.285	1.0	-0.360
y	-0.075	-0.748	-0.188	0.171	-0.264	-0.360	1.0

6.4 단계 4: Model (모형화)

제품색상(y)과 원인변수들과의 함수관계를 모형화(modeling) 시켜 보기 위하여 단계별 회귀분석(stepwise regression)을 실시하여 보니 <표 5>과 같은 결과를 얻었다. 이러한 회귀방정식은 모든 통계패키지에서 가능하다. 이 표는 각 단계별로 y의 변화를 가장 잘 설명하여 주는 원인변수들의 부분집합으로 구성된 다항 선형회귀모형을 보여주고 있다.

위의 결과에서 x4는 전혀 의미 없는 변수로 판명되었다. 실무적으로 볼 때 x1과 x6는 임의조절이 매우 어려운 변수이고, 단계별 회귀분석 결과에서 보면 단계 3에서의 모형이 상당히 좋은 추정 방정식(결정계수 0.680)을 주므로, 단계 3의 결과방정식을 y를 설명하는 최종의 최적 모형으로 채택 하였다

<표 5> 단계별 회귀분석의 결과

단계	입력변수(원인변수)	추정된 회귀방정식	결정계수(R^2)
1	x2	$\hat{y}$ = 40.538 - 0.145 x2	0.560
2	x2, x3	$\hat{y}$ = 48.697 - 0.149 x2 - 0.156 x3	0.620
3	x2, x3, x5	$\hat{y}$= 55.952 - 0.136 x2 - 0.269 x3 - 0.066 x5	0.680
4	x2, x3, x5, x1	$\hat{y}$ = 58.682 - 0.139 x2 - 0.218 x3 - 0.084 x5 - 0.339 x1	0.725
5	x2, x3, x5, x1, x6	$\hat{y}$ = 65.046 - 0.153 x2 - 0.265 x3 - 0.111 x5 - 0.045 x1 - 0.021 x6	0.755

6.5 단계 5: Predict (예측)

결과변수 y와 원인변수로 선정된 x2, x3, x5 간에 어떤 상관관계를 주는가를 <표 4>에서 살펴보니 모두 음의 상관관계(negative correlation)를 가지고 있다. 이들 원인변수가 실제로 취할 수 있는 가능한 값의 범위를 조사하여 보니 다음과 같다.

$$110 \leqq x2 \leqq 170$$
$$40 \leqq x3 \leqq 60$$
$$36 \leqq x5 \leqq 80$$

따라서 최적 회귀방정식에서 판단할 때, 각 변수들의 회귀계수가 모두 음이므로 이 변수들이 최소값을 취할 때 망대특성인 y가 최대가 된다. x2 = 110, x3 = 40, x5 = 36을 취할 때 방정식에서 y값을 예측하여 보면

y의 예측치 = 55.952 - 0.136 (110)
- 0.269 (40) - 0.066 (36)
= 27.9

가 된다. 이 정도의 값이면 색상의 세이볼트 값으로 매우 만족스러운 값이다. 따라서 결론으로 다음과 같이 feed 조건과 반응기 조건을 변경하기로 하였다.

x1 (Medaptan 함량) = 현재 수준 사용 (약 117 wt. ppm)
x2 (공기주입량) = 현재의 128 수준에서 110으로 낮춤.
x3 (feed 온도) = 현재의 49 0C 수준에서 40 0C 수준으로 낮춤.
x4 (feed end point) = 현재 수준 사용 (약 264 0C)
x5 (촉매활성제 주입량) = 현재의 51(wt. ppm) 수준에서 36 수준으로 낮춤.
x6 (처리원유 중 I/H 함량) = 현재 수준 사용 (약 23 vol. %)

6.6 단계 6: Verify (검증)

위에서 얻은 표준화 조건에서 색상이 만족스러운 값을 얻으므로 허용차에 대한 분석을 통하여 최적 운전조건을 다음과 같이 정하고, 이를 표준으로 문서화하기로 하였다.

x1 (Medaptan 함량) = 117 ± 3 wt. ppm
x2 (공기주입량) = 110 ± 2
x3 (feed 온도) = 40 ± 2 0C
x4 (feed end point) = 264 ±5 0C
x5 (촉매활성제 주입량) = 36 ± 3 wt. ppm
x6 (처리원유 중 I/H 함량) = 23 ± 1 vol. %

7. DT와 e-통계학와 데이터 마이닝

7.1 e-통계학이란?

e-통계학(Electronic-Statistics)은 Devillers (2002)에 의하여 제안된 통계학의 분야로, 컴퓨터를 활용하여 전자식 데이터의 수집, 저장, 검색, 정제를 실시하고, 컴퓨터를 통한 통계분석 및 모형화를 통하여 자동적인 지식의 창출 및 추론을 실시하며, 통계적 시뮬레이션과 예측을 실시하는 과학적 방법론이다. 따라서 e-통계학은 컴퓨터의 활용과 통계 소프트웨어의 활용을 강조하는 DT의 핵심적인 영역이라고 하겠다.

7.2 데이터 마이닝이란?

데이터 마이닝(DM: data mining)이란 큰 규모의 데이터 베이스(DB: data base)를 탐사하여 알려지지 않은 유용한 정보와 지식을 만들어내는 기법 또는 과정을 말한다. 따라서 DM은 DT의 중요한 도구이다. 넓은 의미의 DM은 DB로부터 새로운 지식을 추출하는 전 과정을 의미하는 DB로부터의 지식발견(KDD: knowledge discovery in database)에 해당한다. DM의 발전 초기에 Chatfield (1995), Clark (1997), Kohovi와 Provost (2001) 등의 기여가 크다. DM의 중요 단계는 대략 다음과 같다.

- 데이터의 선정 - 데이터 베이스에 있는 대량의 운영 데이터로부터 필요한 부분의 데이터만을 선정함.
- 데이터의 정제 - 대량의 데이터를 취급할 경우에는 항상 데이터에는 오류, 누락, 중복 등으로 불완전한 상태이다. 따라서 데이터의 정제(cleaning) 과정이 필요함.
- 데이터 웨어하우스의 구축 - 필요에 따라 새로운 정보 또는 레코드가 쉽게 추가될 수 있어야 하며, 필요한 정보나 레코드를 쉽게 입출력할 수 있도록 데이터베이스를 구축할 필요가 있음. 이런 데이터베이스를 데이터 창고(warehouse)의 개념으로 해석하여 데이터 웨어하우스라고 부름.

■ 정보의 추출 - 여러 가지의 DM 기법을 통하여 정제된 데이터 웨어하우스로부터 필요한 정보를 추출하는 과정.

DM을 위하여 잘 알려진 패키지들은 CART (Salford Systems), Enterprise Miner (SAS Institute Inc.), AnswerTree (SPSS Inc.), Clementine (SPSS Inc.) 등이 있다.

7.3 e-통계학, DT와 DM 간의 관계

위의 정의에서 보면 e-통계학은 DT의 한 영역이며, DM과는 상당부분 중복되는 것은 있으나, 반드시 다량의 데이터와 데이터 베이스를 사용하는 DM과는 구별되는 측면이 있다. e-통계학은 데이터 마이닝(DM), 통계적 공정관리(SPC: statistical process control), CRM, 지역정보시스템(GIS: geographic information system) 등에 매우 유용하게 사용된다. 여기서 DT, IT, e-Statistics, DM 간의 관계를 간단히 그림으로 그려보면 <그림 4>와 같다.

8. 결 론

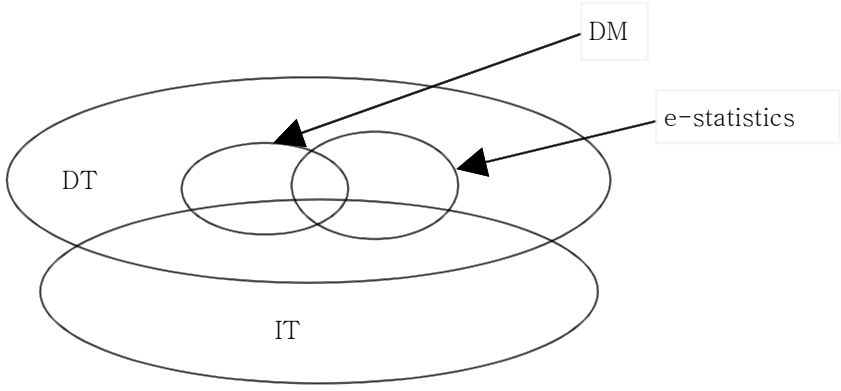
필자는 DT가 통계학이 살아나갈 중요한 방향이라고 생각하며, 산업공학이나 경영과학, 그리고 품질경영에 이르기까지 광범위하게 응용될 수 있는 중요한 방법론적인 인프라를 구성할 것을 확신한다. DT는 융합과학으로서 지식사회에서 하나의 학문분야로 발전해 나갈 것이라고 믿는다. 여기서 DT의 향후 연구 과제로 다음의 3가지를 제시하고 싶다. 이 3가지 과제가 성공적으로 연구될 때 DT는 굳건히 하나의 융합학문으로 자리잡을 수 있을 것이다.

8.1 DT와 DFSS가 연계된 DT-DFSS 개발

R&D 기관에서 체계적이고 과학적인 DFSS(Design For Six Sigma) 기법을 개발하여 적용하는 것은 매우 바람직하다. 아직까지 우리나라에 도입된 Six Sigma는 주로 제조부분에 치우쳐 있으며, 연구개발단계에서의 Six Sigma 적용 방법론이 잘 개발되어 있지 못하다. 우리나라의 연구개발 기관에 적절한 DT와 연계된 DFSS 방법론 (DT-DFSS)을 개발하여 반복적인 업무처리와 통계적인 분석이 실시간 (real time)으로 최적화 되도록 과학적이고 체계적인 시스템으로 통계프로그램화 및 문서화하고 매뉴얼로 발간해 내는 것이 향후 중요한 연구과제가 될 것이다.

8.2 Smart한 지능형 DT 패키지 개발

기업에서 시급히 필요한 것은 품질, 안전, 생산성, 수율 등을 동시에 통합하여 향상시키기 위한 'smart'한 데이터 분석용 통계패키지이다. 현재 Six Sigma 경영, 전사적 품질경영(TQM) 등을 위한 데이터 분석을 위하여 Minitab이 한국시장을 상당 부분 장악하고 있으나, 이는 단지 통계분석에 그칠 뿐 사용자에게 의사결정에 필요한 더욱 친밀하고 실용적인 정보를 제공하여 주지 못하고 있다. 우리 실정에 적합하고 사용자에게 더욱 가까이 다가갈 수 있는 지능형 데이터 분석 통계패키지를 한글로 개발하여 보급하는 것이 필요하다. 이는 데이터에서 정보를 추출하여 그 결과를 제공하는데 그치지 말고, 그 결과를 지식화 하여 사용하는데 고민하지 않도



<그림 4> DT, IT, e-Statistics와 DM 간의 관계

록 하는 내용을 가진 “Smart"한 패키지를 만드는 것이다. 이러한 패키지는 DT의 개념과 방법론이 밑바탕을 이루어야 한다.

8.3 각 영역별 다양한 모델의 개발 연구

DT의 과정에서 가장 중요한 부분은 모델링이다. 현상을 잘 설명하여주고, 이 모델로부터 미래의 변화를 예측하기 위해서는 신뢰성 높은 모델이 얻어져야 한다. 이를 위하여 DT에 잘 훈련된 전문가와 각 영역에서의 전문가들이 공동으로 작업하면 분야별로 좋은 모델을 만들 수 있다. 예를 들면, 주식시세 예측 모델, GDP 예측 모델, 당뇨병 증세 예측 모델 등은 모두 DT의 산출물이며, 정확성과 안정성을 가진 모델을 만들기 위한 연구 노력이 계속되어야 한다.

참고문헌

[1] 박성현 (2001a), “데이터 기술의 경제학”, 「한국경제신문 오피니언」, 2001년 12월 3일자.

[2] 박성현 (2001b), “지식기반 사회에서의 통계학 패러다임의 변화와 데이터 기술의 발전”, 「경영정보논총」, 서울대학교 경영대학, 제11권, p.53-59, 2001. 12.

[3] Chatfield, C. (1995), “Model uncertainty, Data Mining, and statistical inference”, *Journal of Royal Statistical Society*, *Series A*, Vol. 158, p. 419-466.

[4] Clark, G. (1997), “Statistical Themes and Lessons for Data Mining”, *Data Mining and Knowledge Discovery*, Vol. 1, p. 11-28.

[5] Devillers,J.C.(2002),“e-Statistics for Deriving QSAR Models”, *SAR and QSAR in Environmental Research*, Vol. 13, pp. 409-416.

[6] Erto, P., Pallotta, G. and Park, S. H. (2008), “An Example of Data Technology Product: a Control Chart for Weibull Processes”, *International Statistical Review*, Vol. 76, No. 2, pp. 157-166.

[7] Friedman,J. H. (2001),“The Role of Statistics in the Data Revolution”,*International Statistical Review*, 69, p. 5-10.

[8] Healy, M. J. R. (1978), “Is Statis-tics a Science?”, *Journal of the Royal Statistical Society*, *Series A*, 141, p. 385-393.

[9] Kohavi, R., and Provost, F. (2001), “Applications of Data Mining to Electronic Commerce”, *Data Mining and Knowledge Discovery*, Vol. 5, p. 5-10.

[10] Park,S. H. (2003), “Data Technol- ogy and Knowledge-based Six Sigma”, Asian Journal on Quality, Vol. 4, p. 40-45.

[11] Park, S. H. and Suh, M. W. (2008), “Data Technology as a New Discipline for Broader Application of Statistics”, *Journal of Data Science*, Vol. 6, No. 3, pp. 357-368, July 2008 issue, ISSN 1683-8602(on-line).

[12] Straf, M. L. (2003), “Statistics: The Next Generation”, *Journal of the American Statistical Association*, 98, p. 1-6.

토론

임용빈

이화여자대학교 자연과학대학 통계학과

Discussion

Lim Yong Bin

Department of Statistics, Ewha Womans University

우리나라가 향후 발전시켜야할 첨단과학기술 분야로 IT(정보기술), BT(생명공학), NT(나노기술), ST(항공우주기술), ET(환경기술), CT(문화기술)를 들고 있고, 최근에는 이들 간의 융합기술 (예로, IT + BT, IT + ET 등)의 발전을 많이 언급하고 있는 것은 잘 알려진 사실이다. 이 논문에서는 데이터로부터 시작되는 데이터 기술(DT)을 정의하고, 거시적으로는 국가 경제 지표의 과학적 관리 운영과 정책 결정의 합리성을, 미시적으로는 산업에서의 품질과 생산성 최적화 등 DT의 발전으로 인한 효과를 구체적으로 나열하고, 이와 관련하여 DT의 중요성을 체계적으로 강조하고 있다. 또한, DT가 지식기반 정보화 사회에서 과학기술의 선진화, 국가경쟁력 제고, 전 국민 과학화 등을 위하여 첨단과학기술 분야로 포함되어야 하는 당위성을 잘 설명하고 있다. DT의 본질에서는 DT는 모든 첨단과학기술의 기초에 해당하는 원천적 기초과학 기술인 소프트웨어적 인프라에 해당함을 주목하고, DT와 IT의 정보흐름도를 비교하면서 IT와 DT의 차이점을 잘 부각시키고 있다.

DT의 정보흐름도에서 데이터의 수집은 이 논문에서 언급된 바와 같이 후 단계의 데이터로부터 정보의 창출과 정보로부터 미래 예측 및 지식기반 창출을 효율적으로 할 수 있도록, 즉 수리적 모델링에 의한 미래 예측을 정밀하게 할 수 있도록 설계되어야 한다. 이와 관련하여 DT가 IT와 차별성을 갖는 토론자가 경험한 예를 하나 들어보자. 초박막 액정표시장치인 TFT-LCD에서 TFT는 액정을 제어하기 위해 초박형 유리기판위에 반도체 막을 형성한 회로인데, 전기적 특성과 관련된 TFT 공정은 원장 글라스가 투입되면, GATE, SD, PAS & PXL Layer 단계를 거쳐서 TFT가 만들어 진다. TFT공정은 완전 자동화로 이루어 지고, 실처리 변수(입력변수)들은 주기적으로 자동 측정이 되어서 데이터베이스(DB)에 저장이 된다. 토론자는 DB에 저장된 실처리 변수들에 관한 자료를 활용하여 각 Layer 별로 불량에 영향을 미치는 실처리 변수들을 선별하여 직행수율을 향상시키기 위한 최적조건을 찾고, 수율 향상에 기여할 식스시그마 프로젝트의 목록을 제시하는 과제에 참여한 적이 있다.

이 과제를 수행하기 위해서는 데이터의 수집 단계에서 투입된 원장 글래스별로 실처리 변수들에 관한 이력 데이터 마트를 구축해야 한다. 그런데 현재의 데이터베이스에 저장된 실처리 변수들을 측정하는 목적은 장비가 제대로 작동하고 있는지를 감시하기 위함이다. 즉, 장비의 유지를 목적으로 실처리 변수들이 자동으로 측정이 되어서 저장이 되어 있기 때문에, 공정별로 측정된 실처리 변수의 원장 글라스가 앞 공정에서도 측정이 되었는 가의 여부는 관심사항이 아니다.

토론자의 연구팀은 공정별 실처리 변수들의 자료를 원장 글라스의 id 별로 정리하고, 병합하여 투입된 원장 글래스별로 실처리 변수들에 관한 이력 데이터 마트를 구축하기 위해서 많은 시간을 할애하고, 힘들게 작업한 기억이 난다. 힘들게 구축된 데이터 마트의 문제점은 이미 언급한 바와 같이 장비의 유지를 목적으로 실처리 변수들을 자동 측정하여, 원장 글라스 변수마다 많은 실처리 변수에서 결측치가 일어나고 있다는 사실이다. 이는 Layer 별로 직행수율에 관한 수리적 모델링에 제약조건을 초래하고, 정밀한 예측에 관한 장애 요인이 된다. DT의 관점에서 데이터를 수집했더라면, 동일한 글라스의 실처리 변수들이 선공정, 후공정 등에서 자동 측정이 되도록 설계되었을 것이다.

이 논문에서는 DT 적용 프로세스로 DMAMPV (Define, Measure, Analyze, Model, Predict, Verify)를 제안하고 있다. 토론자는 또 하나의 대안으로 DMAOV (Define, Measure, Analyze, Optimize, Varify)를 추천한다. 추천 이유는 다음과 같다. 실험계획법 전문 소

프트웨어인 Design Expert에서는 실험계획법의 단계를 Design, Analyze, Optimize 로 구분한다. Design 단계에서는 효율적인 실험점들을 생성하고, Analyze 단계에서는 회귀진단을 통해서 실험자료에 적합한 모형을 찾는 단계로 이 논문에서 제시한 적용프로세스의 A와 M단계를 포함한다. Optimize 단계에서는 결과 변수 y를 최적으로 하는 입력변수들의 최적조건을 최적화 알고리즘을 실행하여 찾고, 분석 결과의 재현성을 확인하기 위해서 최적조건에서의 미래의 예측치에 신뢰구간을 제공한다.

참고문헌

[1] Stat-Ease(2005). *Design-Expert, software for response surface methodology and mixture experiments, Version 7*, Minneapolis: Stat-Ease.

토론

장대흥

부경대학교 자연과학대학 통계학과

Discussion

Jang Dae Heung

Department of Statistics, Pukyong National University

통계학이 과학(science)인가, 아니면 기술(technology)인가에 대한 논의는 서양에서 일찍이 격렬하게 거론된 바가 있고 모두가 인정하지는 않더라도 대부분의 관련 학자들 사이에서 통계학은 과학이라고 결론이 났다. 이러한 시비가 일어나게 된 것 자체가 통계학이 다른 과학에 비교하여 응용성이 매우 강조되는 과학임을 암시한다고 하겠다. 본 논문을 읽고 두 가지 사항에 대하여 언급하고자 한다.

첫째, 데이터기술(DT:Data Technology)'이라는 용어 문제이다. 본 논문에서는 '데이터기술'에 대하여 '데이터의 측정, 수집, 축적 기술에서부터 시작하여, 데이터의 전송, 분석 및 해석 능력, 데이터로부터 정보와 지식을 창출하는 기술, 데이터로부터의 통계적 모형화 기술, 데이터로부터 미래를 예측하는 기술 등을 다루는 전반적인 과학기술적 방법론'이라고 정의하였다. 이 용어에 대하여 아직 관련 학자들 사이에 동의(consensus)가 필요할 뿐 만 아니라 이 용어의 사용/확산을 위한 전략 수립이 필요하다 하겠다. 전 세계적으로 DT라는 용어가 아직 일반화되어 있지 못하므로 DT가 IT와 다른 점들을 확실히 부각시키며 DT라는 용어를 학술적으로 정착시킬 필요가 있다. 이러한 학술용어의 정착을 위한 전략 수립이 대단히 중요하다고 생각한다. 이를 위하여 통계학, 산업공학, 경영학 등의 학제간 연구가 활발히 전개될 필요가 있다.

둘째, DT의 과정에서 가장 중요한 부분은 모델링이므로 이 모델링을 통한 데이터분석을 시행하기 위한 통계처리도구, 구체적으로 통계소프트웨어가 필수적으로 필요하다. 본 논문에서는 '데이터에서 정보를 추출하여 그 결과를 제공하는데 그치지 말고, 그 결과를 지식화하여 사용하는데 고민하지 않도록 하는 내용을 가진 "Smart"한 패키지를 만들어야 한다'고 언급하였다.

또한 '기존의 통계패키지들-예를 들어 Minitab-은 ' 단지 통계분석에 그칠 뿐 사용자에게 의사결정에 필요한 더욱 친밀하고 실용적인 정보를 제공하여 주지 못하고 있다'고 언급하고 '우리 실정에 적합하고 사용자에게 더욱 가까이 다가갈 수 있는 지능형 데이터 분석 통계패키지를 한글로 개발하여 보급하는 것이 필요하다'고 언급하여 새로운 한글패키지 개발에 대한 제안을 하였다. smart한 패키지라는 말과 연관하여 기억되는 한 가지 해프닝이 생각난다. 인공지능학자들과 통계학자들의 학제간 연구로서 1980년대에인공지능(artificial intelligen-ce)과 통계학의 연계성에 대하여 여러 논문들이 등장했었던 때가 있다. 물론 이 때는 컴퓨터과학계에서 인공지능 바람이 불던 때이다. 이 당시의 관련 학자들의 꿈은 '통계자료만 주어지면 컴퓨터가 알아서 자료의 성격을 파악하고, 이 자료에 적합한 통계모형을 선택하여 자료분석을 자동으로 행하고, 결론도 통계학자가 아닌 일반인들이 이해할 수 있는 용어로 출력이 되는' 패키지를 개발하자는 것이었다(예로, Gale(1986)). 이러한 작업들 중 하나의 결과가 회귀분석용 인공지능 소프트웨어 개발이었다. 이러한 연구들은 작은 성과들은 이루었지만 크게 보면 모두 실패로 돌아갔다(물론, 1980년대 이후에도 이러한 작업들은 지속되어 왔다(예로, Hand(1993)). 다양한 인간세계에서 나타나는 다양한 데이터들을 컴퓨터가 자동처리하기에는, 지금도 그러하지만 1980년대에는 더욱이 어려운 작업이었다. 그러므로 smart한 패키지를 개발하는 데 있어서 다음과 같은 여러 가지 고려 사항들을 염두에 두고 개발하여야 할 것이다.

1. smart하다는 것의 정의와 범위
2. smart한 패키지에 사용되는 컴퓨터언어
3. smart한 패키지를 개발하기 위한 조직구성과 프로세스

앞에서 언급한 2번과 관련하여 지능형 데이터 분석 통계패키지를 어떤 컴퓨터언어를 사용할 것인가 하는 문제를 생각해 볼 필요가 있다. 현재 대표적인 상업용 통계패키지로는 SAS, SPSS, Minitab, S-Plus 등이 있다. 또한 Excel을 이용하여 우리는 기초통계학 수준의 자료분석을 행할 수 있다. 그러나 이러한 통계패키지들은 모두 상업용 통계패키지이다. 즉, 사용자가 돈을 지불하고 이 들 통계패키지를 구입하여야 한다는 것이다. 그러나 R은 공개용(GPL, General Public License) 통계적 분석도구이다. 즉 무료라는 것이다. R의 사용이 무료이기 때문에 R을 사용한 통계분석이나 패키지 개발은 개인이나 단체에서 얼마든지 권장하거나 시행할 수 있다. R은 통계언어인 S 언어를 기반으로 하여 1995년 Auckland 대학 Robert Gentleman과 Ross Ihaka에 의하여 만들어진 통계용 언어이다. 대부분의 통계패키지가 일괄처리방식(batch mode)과 메뉴방식인 데 비하여 R은 대화식 처리방식(interactive mode)을 따르기 때문에 R console이라는 창에서 프롬프트(>) 다음에 R 명령문을 입력하고 ENTER 키를 치면 명령문이 시행이 되어 그 결과가 그 다음 줄에 바로 나타나게 된다. 통상 기업에서의 자료분석에서는 SAS, SPSS, Minitab 등이 주종을 이루고 연구자들의 자료분석에서는 SAS, SPSS, S-Plus 등이 주로 쓰이나 R이 최근에 폭발적으로 전 세계적인 인기를 얻고 있고 R 관련 통계서적들도 물밀듯이 쏟아져 나오고 있다.

전 세계 통계학자들의 최근기법들이 R로 작성되어 발표되고 있는 실정이다. R 공식홈페이지(www.r-project.org)에 가면 최근기법들을 R로 작성한 패키지들이 수천 개가 넘게 제시되어 있다. 이 들 패키지 중 통계적 품질관리에서 많이 쓰이는 관리도 작성 패키지로서는 'qcc'가 있다. 이 패키지를 이용하면 각종 계량형, 계수형 관리도들을 그릴 수 있고 공정능력지수를 계산하고 공정능력지수 관련 그림들을 그릴 수 있다. 생존분석(신뢰도공학)에서 쓰일 수 있는 'survival'이라는 패키지가 있고 실험계획법 관련 패키지들도 있다.

R은 1년에 수 차례 버전이 갱신되는 데 가장 최근 버전은 2010년 5월 31일 현재 R version 2.11.1이다. 허명회(2007)는 1장에서 'SPSS나 Minitab은 사용자 편의성을 극대화하여 거의 모든 것이 메뉴화되어 있어 원하는 것을 메뉴에서 골라 콕콕 찍기만 하면 된다. 반면, R은 언어이다. 즉 말하고 듣고 쓰는 방법을 새로운 외국어를 학습하듯 배워야 한다.'라고 서술하고 있다. S-Plus는 이러한 메뉴방식을 대화식 처리방식과 병행하고 있다. 반면, R은 메뉴화되어 있지 않다. 이것을 극복하고 R 패키지를 GUI화한 것이 R Commander이다. 이 R Commander는 Fox(McMaster 대학교수)가 개발한 R 통계GUI이다(Fox(2005) 참조.). R에서의RGUIproject (http://www.sciviews.org/_rgui/)는 여러 가지 방면으로 추진되고 있고 그 중의 하나가 바로 이 R Commander이다. 이 R Commander는 통계분석을 위한 R GUI이다. R이 무료 통계패키지인 것처럼 R Commander도 무료이므로 어느 장소, 어느 때나 컴퓨터에 저장하여 무료로 사용할 수 있다. 현재 버전은 R Commander version 1.6-0이다. 이 R Commander에서 관리도 작성 패키지 'qcc'나 생존분석(신뢰도공학)관련 'survival' 등을 불러 사용할 수가 있다. 지능형 데이터 분석 통계패키지를 개발할 때 이러한 유용한 R 패키지를 기반으로 하는 것이 좋은 전략이라고 생각한다.

참고문헌

[1] 허명회(2007).「R을 사용한 탐색적 자료분석」, 자유아카데미.

[2] Fox, J. (2005). "The R Commander: A Basic-Statistics Graphical User Interface to R", *Journal of Statistical Software*, Vol.14, pp.1-42.

[3] Gale, W. A.(1986). *Artificial Intelligence and Statistics*, Addison-Wesley, Reading.

[4] Hand, D. J.(1993). *Artificial Intelligence Frontiers in Statistics*, Chapman & Hall, London.

토론

박희준

연세대학교 공과대학 정보산업공학과

Discussion

Park Hee Jun

Department of Information and Industrial Engineering, School of Engineering, Yonsei University

산업화 사회에 접어들면서 활발해진 기업의 경제활동으로 인류가 생산하는 데이터의 총량이 두드러지게 증가하였으며, 20세기 후반에 이르러 정보기술이 급속하게 발달하면서 데이터를 저장할 수 있는 공간의 크기와 데이터를 공유할 수 있는 경로의 속도가 거의 무한대로 증가함으로써 데이터의 총량은 기하급수적으로 증가하게 된다.

데이터 총량의 증가와 데이터를 생산하고 활용하는 주체를 하나로 묶는 네트워크의 발달은 데이터에 근거해서 보다 계획적이고 과학적으로 경영 활동을 할 수 있는 기회를 기업에게 제공함으로써 기업의 생산성 향상에 기여해 왔다. 하지만 최근에 기하급수적으로 증가하고 있는 데이터와 정보의 총량 그리고 더욱 촘촘해진 네트워크는 기업이 경영 활동을 함에 있어서 경영 활동에 영향을 미치는 더 많은 요인들을 고려해야하는 필요성을 제공함으로써 계획적이고 과학적인 경영 활동에 어려움을 주고 있다. 기업들은 데이터와 정보의 범람으로 야기된 경영 환경의 불확실성 속에서 어떠한 데이터를 어떻게 활용하여 생산성을 향상시킬 수 있을지 고민하고 있다.

사용자들의 데이터에 대한 접근성과 데이터 활용 능력이 향상됨에 따라 시장의 요구는 더욱 높아지게 되었고, 높아진 시장의 요구에 기업들은 적절히 대응하면서 혁신적인 기술의 변화를 수용해야 하는 어려움에 직면하게 되었다. 또한 데이터를 생산하고 활용하는 주체들은 인터넷과 이동통신에 기반을 둔 네트워크를 통해서 연계되고, 그들 간의 상호작용은 세계 경제와 시장을 하나로 묶임으로써 기업은 지구촌 곳곳에서 발생하는 변화를 감지하고 그 변화에 반응해야 할 처지에 놓이게 되었다.

데이터 기술(DT: Data Technology)은 고조된 경영 환경의 불확실성에서 출발한 기업의 위기를 기회로 만들어 갈 수 있는 도구를 제공한다. 산업혁명 이후 계획과 통제 중심의 효율성 향상에 기반을 둔 기업의 경영 활동은 최근에 경영 환경의 불확실성이 증가함에 따라서 감지와 반응 중심의 효과성 향상에 기반을 둔 경영 활동으로 패러다임이 변화하고 있다. 데이터 기술은 정보통신기술을 통해서 수집되어 저장되고 공유되는 수많은 데이터를 분석하여 혁신의 기회를 발견하고 실천할 수 있는 도구로 활용될 수 있다.

하지만 지금껏 활용되었던 선형적인 데이터의 활용 방법은 비선형적인 활용 방법으로 전환되어야 할 필요성이 있다. 데이터의 선형적인 활용에 의한 계획과 통제 중심의 경영 활동으로는 급변하는 환경의 변화에 기업이 적절하게 대응하면서 지속적인 성장을 만들어 가기 힘들기 때문이다. 또한 기업은 기업 내부 구성원 간의 데이터 공유뿐만 아니라 기업 외부에 존재하는 사용자, 협력 기업, 경쟁 기업 등과의 데이터 공유를 전략적으로 수행함으로써 기업 혁신 활동의 위험성을 줄이고 시장 지배력을 높여가야 한다.

21세기의 기업과 산업의 경쟁력은 데이터를 저장하고 유통할 수 있는 정보통신기술 인프라를 확보하는 데에 있지 않다. 저장되고 유통되는 데이터로부터 부가가치를 만들어 낼 수 있는 콘텐츠를 개발하여 확보하는 데에 있다. 국가 경쟁력 확보를 위해서 정부와 기업은 기하급수적으로 증가하는 데이터로부터 부가가치를 창출할 수 있는 열쇠를 제공하는 데이터 기술의 발전에 보다 많은 투자를 체계적이고 전략적으로 만들어 내야 할 것이다. 따라서 데이터 기술의 중요성을 언급하고, 데이터 기술에 대한 투자의 방향을 제시한 본 논문의 내용에 전적으로 동감한다.

참고문헌

[1] 찰스 리드비터 지음. 이순희 옮김, 「집단지성이란 무엇인가」.서울:21세기북스. 2009.

[2] Don Tapscott and Anthony D. Williams. 2006. *Wikinomics: How Mass Collaboration Changes Everything*. New York: a member of Penguin Group(USA) Inc. man & Hall, London.

저자 답변

박성현

한국연구재단 기초연구본부

Response

Park Sung Hyun

Directorate for Basic Research in Science & Technology National Research Foundation of Korea

세분의 토론자가 모두 미래에 지식창조를 위한 데이터 기술의 중요성을 인정하여 주신 것에 대하여 감사한다. 토론자들이 데이터 기술에 대한 추가적인 귀한 의견을 주셨으며, 이 토론에 대한 저자의 응답을 다음에 간단히 기술하기로 한다.

1. 임용빈 토론자:

임 교수님은 두 가지 사항을 토론하고 있다. 하나는 <그림 1>에 있는 정보 흐름도에서 DT의 데이터 수집 단계에서 IT의 데이터 수집과의 차별성을 강조하면서 "데이터 수집은 정보의 창출과 정보로부터 미래 예측 및 지식기반 창출을 효율적으로 할 수 있도록, 즉 수리적 모델링에 의한 미래 예측을 정밀하게 할 수 있도록 설계되어야 한다." 라고 언급하고 있다. 전적으로 동의한다. 이와 관련된 참고문헌으로 박성현(2003), Park and Suh(2008)을 참조하여 주기 바란다. 이런 것이 DT와 IT의 차이점이 될 것이다. 임 교수는 더 나아가서 다음과 같은 구체적인 사례를 들고 있다.

초박막 액정표시장치인 TFT-LCD 공정에서 이 공정에서 데이터 수집의 목적은 장비의 유지목적으로 각종의 입력 변수들이 자동적으로 측정되며, 이들 데이터가 허용차 안에 들어가서 장비가 제대로 작동하고 있는가 만을 판단하고 있는 것이다. 이런 데이터는 장비유지 목적의 IT 데이터이며, 모델링도 어려워서 DT 데이터라고 볼 수 없다. DT 데이터가 되려면 글라스 별로 식별번호(i.d.)가 있어야 하고, 각종 입력변수들에 대하여 선공정, 후공정을 구분하여 자동측정이 이루어져야 하며, 글라스별로 모델링이 되어 최적조건을 찾을 수 있어야 하는 것이다.

임교수의 의견에 동의한다. 따라서 DT 데이터가 되려면 데이터 수집단계의 설계가 매우 중요하며, 통계적 모델링(회귀분석 모형, 데이터 마이닝 모형 등)이 될 수 있도록 정제된 데이터가 작성되어야 한다. 모델링이 되어야 미래 예측이 가능하며 검증이 될 수 있는 것이다.

2. 장대흥 토론자:

장 교수님은 세 가지 사항을 토론하고 있다. 첫째로 "데이터 기술(DT)이란 용어를 학술적으로 정착시킬 필요가 있으며, 이 용어의 사용/확산을 위한 전략 수립이 필요하다" 라고 언급하고 있다. 전적으로 동의 한다. DT에 관심이 있는 학자들이 팀을 이루어 산업공학, 통계학, 경영학 등의 학제간 연구가 활발히 전개되기를 바라며, 한국품질경영학회가 그 중심에 서기를 기대한다. 두 번째로, 장 교수는 소위 "Smart"한 패키지의 개발은 바람직하나 smart의 정의와 범위, smart한 패키지에 사용되는 컴퓨터 언어, smart한 패키지를 개발하기 위한 조직 구성과 프로세스를 충분히 고려하여 개발하여야 한다고 지적하였다. 이에 대하여 전적으로 동의한다. DT의 핵심은 DT의 활용 사이클인 DMAMPV를 smart 하게 돌릴 수 있는 통계패키지의 개발이 필요하며, 특히 정보를 창출하여 주는 MP(model, predict)의 과정이 중요하다.

세 번째로 smart한 패키지를 개발할 때 장 교수는 컴퓨터 언로로서는 R을 추천하고 있다. 이것에 대해서도 동의한다. R은 GUI(Graphical User Interface)화 되어 있고, R-Commander에 의하여 실행이 가능하다. 앞으로 지능형 데이터 분석 통계패키지를 개발할 때 R패키지가 유용하게 사용될 것으로 믿는다.

3. 박희준 토론자:

박 교수님이 언급한 토론의 내용 중에서 특히 공감하는 문구로는 첫째로, "최근에 경영활동의 불확실성이

증가함에 따라서 감지와 반응 중심의 효과성 향상에 기반을 둔 경영 활동으로 패러다임이 변화하고 있으며, 데이터 기술은 정보통신기술을 통해서 수집되어 저장되고 공유되는 수많은 데이터를 분석하여 혁신의 기회를 발견하고 실천할 수 있는 도구"라고 언급한 부분이다. 전적으로 동감한다. DT는 IT로 획득한 데이터를 순발력 있게 고차원적으로 분석하여 고급 정보를 창출하는 도구로 활용될 수 있으며, 기업이든 국가든 간에 DT의 활용은 앞서가는 조직을 만드는 필수적인 도구라고 생각한다.

또한 박 교수는 "21세기의 기업과 산업의 경쟁력은 데이터를 저장하고 유통할 수 있는 정보통신기술 인프라를 확보하는데 있지 않고, 이들 데이터로부터 부가가치를 만들어 낼 수 있는 콘텐츠를 개발하여 확보하는데 있으며, 정부와 기업은 데이터로부터 부가가치를 창출할 수 있는 데이터 기술의 발전에 전략적으로 투자하여야 한다." 라고 언급하고 있는데, 매우 동감하는 의견이다. 우리나라는 IT 강국이라고 말하고 있으나, DT의 입장에서 보면 우리나라는 데이터를 저장하고 유통하는 기술은 발달되어 있으나, 데이터로부터 부가가치를 창출하는 기술은 부족하다. DT의 발전으로 부가가치를 창출하는 기술을 발전시킬 수 있으며, DT 연구를 활성화하기 위해서는 DT에 대한 국가적 투자가 필요하다.

DT는 통계학의 응용만의 측면에서 보면 충분하지 않다. DT는 정보통신기술, 전산과학, 통계학, 산업공학, 경영학 등이 결합된 미래 지향적 융합과학이며, 기업이나 국가운영의 과학화에 절대적으로 필요한 기술이 될 것이다. 교육과학기술부나 지식경제부 등에서 DT의 발전에 장기적으로 R&D 투자를 하여줄 것을 기대한다.

품질경영학자들이 DT에 관심을 가지는 것은 바람직하다. 품질관리 측면에서 원인변수와 결과변수 간에 함수관계를 규명하여 결과변수를 관리하는 것은 매우 중요하다. 또한 공정관리에서 최적 공정 조건을 탐색하여 최적의 조건을 유지하여 주는 것은 품질경영의 핵심이다. DT는 이러한 함수관계의 규명이나 최적 공정 조건의 탐색에 매우 유용하게 사용되는 기술이며, 향후 그 활동범위가 더욱 확대될 것으로 생각한다. 세 분 토론자들의 귀중한 의견에 감사한다.

참고문헌

[1] 박성현 (2003), "Visions of Data Technology and e-Statistics with their roles in industry and government",「경영정보논총」, Vol. 13, No. 1, p. 59-68.

[2] Park, S. H. and Suh, M. W. (2008), "Data Technology as a new discipline for broader application of statistics", *Journal of Data Science*, Vol. 6, No. 3, pp. 357-368.

2010년 7월 17일 접수, 2010년 9월 13일 수정