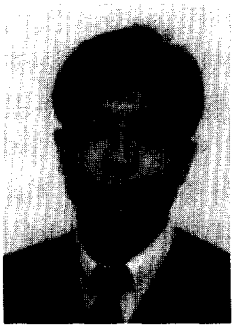
데이터마이닝과 산업공학



조 성 준
서울대학교 산업공학과 교수

사회가 점차 "정보화"되어가고 있다는 것은 우리 모두가 일상생활에서 느낄 수 있다. 특히, 봉급 생활자의 경우, 대부분의 수입과 지출 상황이 데이터화 되어 어떤 조직의 컴퓨터에 저장된다. 우리가 직장에서 받는 봉급은 직장 컴퓨터에, 이 가운데 원천징수 되는 부분은 국세청 컴퓨터에, 급여 지급일 거래 은행으로 계좌 이체 되면 거래은행의 컴퓨터에, 이 중 일부를 신용카드로 소비하면 신용카드 회사의 컴퓨터에, 휴대전화 요금 내면 휴대전화 회사 컴퓨터에, 항공편을 이용하여 여행을 하면 항공사 컴퓨터에, 증권 투자를 하면 증권사 컴퓨터에 모두 저장되고 있다. 최근에는 대중교통 이용 시, 언제 어디서 무슨 교통수단을 이용했는지가 교통카드를 통하여 컴퓨터에 저장된다. 즉, 현금 거래가 점차 축소되면서 대부분 중요 경제 활동 내역이 컴퓨터에 기록된다. 개인 차원의 데이터 생성 뿐 아니라 기업 차원의 데이터 생성도 활발하다. 특히, 최근 센서(sensor) 가격의 하락과 정확도 상승 덕분에 많은 제조 현장의 상황이 실시간으로 저장되고 있다. 또한, 증권 시장에서는 실시간으로 엄청난 양의 데이터가 생성되고 저장된다.

데이터마이닝(data mining)은 데이터의 산더미를 열심히 파서 그 안에 숨어있는 유용한 정보를 찾아내고자 하는 활동이다. 대규모 데이터베이스로부터 탐색과 분석을 통하여 의미 있는 패턴이나 규칙을 찾아내는 과정이다. 여기에는 데이터의 군집화(clustering), 분류(classification), 예측(prediction), 규칙 발견(rule discovery)와 같은 다양한 태스크가 있다. 데이터마이닝이 90년대 들어와서 각광을 받고 있는 배경에는 디스크와 같은 데이터 저장장치 가격의 하락, 이에 따른 데이터의 대량 생성(예: 신용카드, 슈퍼, 은행, 홈 쇼핑, 제조 현장, 증권 거래소 등), 데이터웨어하우스와 같은 데이터 저장 기술 개발 등이 있다. 이에 힘입어 대규모 데이터를 처리할 수 있는 강력한 알고리즘이 개발되고 사용 편의성 및 데이터 처리 능력이 획기적으로 개선된 소프트웨어 패키지가 출현하게 되었다.

데이터마이닝이라는 용어는 데이터베이스 분야를 연구하던 컴퓨터 과학자들이 만든 것이고 현재 미국 등지에서는 컴퓨터 과학자들의 주도하에 ACM(association for computing machinery) 의 SIGKDD(special interest group on knowledge discovery in database) 에서 매년 학술대회가 개최되고 있다. 그러나 실제 데이터

마이닝의 기본 개념은 통계학에 그 바탕을 두고 있으며, 인공지능(artificial intelligence)에서 개발된 모델과 알고리즘을 많이 사용하고 있어서 근본적으로 학제적인(inter-disciplinary) 분야이다. 분야별 해당 기법들을 구체적으로 보면 다음과 같다. 먼저, 통계학 분야의 기법들은 다양한 전처리 기법, 선형/비선형 회귀분석, K-means 군집화 알고리즘, 분류회기분석나무 모형(Classification and Regression Tree: CART), 부트스트래핑(bootstrapping), 선형 /비선형 주성분 분석(linear/nonlinear PCA) 등이 있다. 또한 인공지능의 기계학습과 뉴로컴퓨팅 기법들은 C4.5 나무 모형, 다층퍼셉트론 뉴럴네트웍, 레이디얼 베이시스 함수 네트웍(radial-basis function network), 자기조직특징지도(Self-organizing feature map) 네트웍 등이 있다. 끝으로 데이터베이스 분야에서 개발된 연관 규칙 발견 알고리즘("A priori"), 순서분석 알고리즘 등이 많이 사용되고 있다.

다양한 분야의 연구자들이 데이터마이닝 연구를 수행하고 있는데 데이터마이닝에 대한 생각도 다양하다 할 수 있다. 통계학자들은 "데이터마이닝은 고전적 데이터 분석이므로 통계학의 범주에 속한다"고 생각한다. 그러나 차이가 엄연히 존재한다. 기존의 통계모델에서 가정하고 있는 데이터의 수 보다 몇 천, 몇 만 배 많은 데이터를 다루어야 하고, 실험계획에 의거하여 계획적으로 관찰된 데이터가 아니고 비즈니스 상황에 따라 주어진 데이터가 대부분이며, 이들은 정규분포를 따르지 않는 경우가 많다. 인공지능 분야의 연구자들은 "데이터마이닝은 기존의 귀납적 학습(inductive learning) 을 비즈니스 문제에 적용한 것" 이라고 생각하며, 인공지능 본연의 "완전 지능 시스템" 구축에 비하여 문제의 난이도가 훨씬 낮다고 생각한다. 데이터베이스 관련 연구자들은 "데이터마이닝은 발전된 형태의 데이터베이스 질의(query)"라고 생각한다. 그런 의미에서 통계학이나 인공지능 분야에서 이미 알고리즘이 존재하는 문제에 대한 해법을 재발견하는 경우도 상당히 존재한다. 사실, 관련된 세 분야의 연구자들이 다른 분야의 연구에 대하여 더 많이 이해할수록 각자가 수행하는 연구결과의 질이 높아질 수 있다.

물론 해법을 제시하는 세 분야의 전문가 만으로 데이터마이닝이 성공적으로 수행될 수는 없다. 문제 영역에 대한 이해 또한 매우 중요하다. 특히 응용분야가 생산 및 공정관리, 마케팅, 재무의 비즈니스 영역 뿐 아니라, 일반 공학 및 과학의 영역에까지 넓어지고 있기 때문이다. 또한 이 모든 분야에서는 기존의 방법론에 의하여 문제를 해결해 왔으므로 데이터마이닝 방식의 문제 해결방식이 효과나 효율 면에서 뛰어나다는 것을 보여주기 위해서는 문제에 대한 확고한 이해가 필수적이다. 이 점에서 데이터마이닝은 단순 과학이 아니라 공학적인 접근 방법으로 이해할 필요가 있다. 특히, 비즈니스 문제 해결에 사용되는 경우 비즈니스 프로세스와의 통합이 필수적이다. 즉, 데이터마이닝의 목적이 "흥미로운" 패턴이나 정보의 발견이 아닌 구체적인 비즈니스 문제 해결이어야 한다는 것이다. 특히 문제 파악, 데이터마이닝 적용, 발견한 패턴이나 정보에 의한 비즈니스 액션, 이에 대한 결과 및 성과 측정의 사이클의 한 부품으로서의 역할을 충실히 수행해야 한다. 다음의 실제 사례를 들어보자.

고객이 감소하는 미국의 어느 은행은 이를 타개하기 위한 대안으로 예금 이자를 올리고 대출 이자를 내리는 방안을 고려하였다. 즉, "경쟁력 있는 제품"을 제공한다는 것이다. 그러나 이 대안은 은행 수익 자체를 감소시키고, 소소한 금리 차이에 매우 민감한 기회주의적인 고객들만을 주로 유치하게 된다는 문제점을 가지고 있다. 따라서, 장기적 대안이 아니며, 6개월 후 또는 1년 후에 예대마진이 증가하게 되면 다시 고객들은 떠나버리는 원래의 상태로 돌아오게 되는 것이다. 다른 대안은 "데이터마이닝"을 적용하는 것이었는데, 이 방법은 고객이 감소하는 원인을 파악하고 이에 대한 근본적인 조치를 취하여 고객 감소를 방지하는 것이다. 실제 이 은행의 고객에 대하여 "군집화" 알고리즘을 수행하여 세분화 하였더니, 특정 군집의 이탈율이 특히 높다는 사실을 파악하게 되었다. 이 군집에 속한

고객들의 특징을 살펴보니 주로 저녁 시간에 폰 뱅킹을 이용하여 은행 업무를 처리하고 입출금은 타행 ATM을 이용하고 있었다. 이들은 은행의 자원을 최소한으로 사용하면서도 높은 수수료를 물고 있어서 은행 수익에 크게 기여하는 고수익 고객 군인 것이다. 그런데 이들이 저녁 시간에 전화를 하였을 때에 대기 시간이 매우 길다는 사실 또한 발견하였다. 따라서 이것이 이들의 이탈 원인이 아닌 가하는 의문을 가지게 되었으며, 이에 따라 이들에게만 따로 특별 전화번호를 부여하여 대기 시간을 크게 단축하는 마케팅 액션을 취한 결과, 이들의 이탈율이 크게 감소하게 되었다. 즉, DM을 이용하여 문제의 원인을 파악하고 이에 대한 해결책을 통하여 문제를 해결하고 원래의 비즈니스 목표를 달성한 것이다. 물론 특별 전화번호 부여에 드는 비용과 고객 유지로부터 발생하는 매출 증가에 따른 수익 증가에 대한 철저한 수익 비용 분석도 필요하다.

이제 데이터마이닝과 산업공학과의 관계를 생각해 보기위해 산업공학의 본질에 대하여 살펴보자. 역사적으로 보면 산업공학의 시작은 비즈니스 문제를 공학적인 방식으로 해결하고자 하는 노력에서 출발하였다. 20세기 초반에 노동 집약적인 대량 생산 공정에서의 효율성을 추구하기 위하여 시작된 시간 동작 관리, 20세기 중반 2차 세계대전 시 군수물자의 수송의 효율성을 추구하기 위하여 시작된 오퍼레이션즈 리서치(OR: Operations Research) 가 그 효시이다. 그 이후, 시간 동작 관리는 인간공학 및 생산공학으로, OR은 품질관리, 자동화 및 시스템 운용 등으로 발전하였다. 80, 90 년대에 들어 컴퓨터가 상용화되면서 인간 컴퓨터 인터페이스와 기업의 전 비즈니스의 정보화를 위한 ERP 등이 산업공학의 주 연구 과제로 등장하였다.

이제 21세기 산업공학의 미래를 생각해보면 두 가지 면에서 산업공학의 기회가 있다고 본다. 첫째, 산업공학은 정보 통신 분야에 대하여 더 많은 관심을 기울여야 한다. 그것은 타 산업에 비하여 성장 가능성 및 부가가치가 훨씬 높으며 실제로 대학에서 산업공학을 공부하고 사회에 진출하는 젊은이들이 이 분야로 많이 진출하기 때문이다. 또한 산업공학의 본질이 비즈니스 문제에 대한 공학적 해법이라고 했을 때에 이 공학의 범주에 정보기술은 점점 더 큰 자리를 차지하고 있다. 예를 들어 생산관리에 주로 사용되던 OR이 최근에는 통신망 설계 및 관리 등에 사용되는 것이 대상 변화의 대표적인 예라 할 수 있다. 산업공학이 기본적으로 도구를 제공해주는 학문이라면 시대상황에 맞게 산업공학의 관심도 자연스레 정보 통신 분야로 가고 있다. 둘째, 처음 출발선이었던 "비즈니스 또는 경영 문제의 공학적 접근"이라는 정의를 다시 돌아보면, 생산, 인사, 재무, 마케팅의 경영 4개 기능에서 주로 생산에 집중되었던 관심이 최근 재무와 마케팅으로도 확산되고 있다. 먼저, 재무, 마케팅 분야에서 스스로 공학적 방법론을 사용하고 있는데, 예를 들어 재무에서는 오래 전에 "재무 공학"이 탄생하였고, 마케팅도 "마케팅 공학"이라는 표현이나 교재가 널리 사용되고 있다. 앞에서 언급한 데이터마이닝에 기반을 둔 DB 마케팅이나 CRM 등이 실제 문제 해결에 사용되고 있다. 이런 상황에서 OR, 통계학, 컴퓨터에 밝은 우리 산업 공학도들이 기계학습, 뉴로컴퓨팅 및 데이터베이스 알고리즘에 대한 공부를 한다면 관련 분야 어느 누구보다도 기법에 대한 전문적 총괄적 지식을 쌓을 수 있다. 또한 통계학, 컴퓨터과학등을 전공하는 이들에 비하여 비즈니스 문제에 대한 이해가 높으므로 실제 문제 해결 능력면에서 앞서나갈 수 있다. 현재 데이터마이닝을 가장 활발히 적용하고 있는 국내 굴지의 기업들에서도 산업공학도에 대한 선호는 매우 크다 하겠다. 현재 국내 여러 대학의 산업공학과에서도 학부 대학원 과정에서 데이터마이닝을 전문으로 가르치고 연구하고 있어서 이 분야에 관심이 있는 학생들에게는 좋은 기회라 할 수 있다. 앞으로 데이터마이닝이 산업공학의 중요한 연구 분야로서 자리매김하고 산업공학 출신 가운데 훌륭한 데이터마이너들이 많이 배출되기를 기대해본다.