한국수학사학회지 제18권 제2호(2005년 5월), 95-106

데이터 마이닝의 수학적 배경과 교육방법론

서경대학교 소프트웨어학과 이승우
swlee@skuniv.ac.kr

본 논문에서는 수학을 기반으로 한 데이터베이스의 지식탐사 절차를 통하여 데이터의 선택, 정제, 통합, 변환, 축소, 데이터 마이닝 기법의 선택과 적용 및 모형의 평가에 관한 개념과 방법론을 소개하고 수학의 한 분야로서 통계학의 역할과 적용 방법에 관하여 연구하고자 한다.

또한 오늘날 관심이 대상이 되고 있는 데이터 마이닝의 역사와 수학적 배경, 통계 및 정보 기술을 이용한 데이터 마이닝의 주요 모델링 기법, 실용적 응용 분야 및 적용 사례 그리고 데이터 마이닝과 통계의 차이점에 관하여 조사하고 논하고자 한다.

주제어 : 지식탐사, 데이터 마이닝, 통계

0. 서론

데이터로부터 유용한 정보를 추출하는 과정을 지식탐사(knowledge discovery in database)라고 한다. 일반적으로 지식탐사 절차는 데이터를 선택한 후 정제 단계에서 도메인의 일관성을 유지하며 중복된 데이터를 제거하고 모호성을 제거한다. 그리고 통합, 변환, 축소의 단계를 거쳐 군집·분석·예측을 이용한 데이터 마이닝 단계를 거쳐 모형을 설정하는 단계로 수행된다.

오늘날 국가와 기업들은 사용 가능한 풍부한 정보를 가지는 대규모의 데이터베이스를 보유하고 있으며 거대한 양의 데이터를 축적해왔고 향후 이러한 추세는 변함이 없을 것이다. 그러나 데이터의 무제한적인 증가로 인하여 정확한 정보 분석은 더욱더 어려워졌다. 또한 정제되지 않은 데이터는 지식탐사 절차에 혼란을 초래하여 의사결정을 신뢰할 수 없는 경우를 초래한다.

이러한 불가피하게 어려운 상황에서 많은 기업들은 단순한 통계 정보만을 얻는 것이 아니라 지식탐사 절차를 통하여 정보분석작업을 지원하는 주요한 정보기술인 데이

터 마이닝이라는 새로운 분야에 관심을 가지는 시대적 배경에 놓이게 되었다.

지식탐사 절차에서 중요한 실제 탐사 단계는 데이터 마이닝 단계이다. 그러나 데이터의 선택, 정제, 통합, 변환, 축소의 단계를 간과해서는 데이터 마이닝을 통한 숨겨진 정보를 정확히 찾을 수 없기에 의사결정도 부정확하다.

그 이유는 대규모의 데이터베이스에 오류값[1](incorrect value)이나 특이값[2](outlier)이 포함된 부적절한 데이터들로 인하여 객관적인 의사결정을 유도할 수 없기 때문이다. 그러나 부적절한 데이터를 특이값으로만 판단할 경우, 통계학의 회귀분석을 이용하여 특이값이 영향 관측값[3](influential observation)인가를 신중하게 판단해서 결정해야 한다.

이러한 데이터 처리기법들을 데이터 마이닝 전에 적용하면, 지식탐사 절차의 전반적인 질과 최소한의 시간으로 데이터 마이닝을 시행함으로서 실질적으로 지식탐사 절차의 패턴을 개선시킬 수 있다.

1. 역사적 배경

21세기는 국가 및 기업 간의 경쟁이 심화되고 있기에 정보처리 분야에서 고부가가치를 생산해 낼 수 있는 정보기술 분야 전문가를 양산해서 국제 경쟁력을 확보해야 한다. 오늘날 정보의 중요성에 대한 인식이 확산됨에 따라 컴퓨터를 이용한 대규모 데이터 관리 경험을 통하여 대용량의 데이터에서 유용한 정보를 캐내는 데이터 마이닝(data mining)이라는 분야가 주목받고 있다.

국내에서도 데이터베이스 시스템의 도입과 운영을 통하여 21세기 기업경영에서 중요성이 부각되고 있는 데이터베이스마케팅[4](database marketing), 고객관계관리[5](CRM: customer relationship management), 위험관리(risk management) 등의 분야가 어느 정도 진행되고 있다. 그러므로 그와 관련된 데이터 웨어하우스[6]와 데이터 마이닝 기술이 적용될 단계가 되었다고 판단할 수 있다. 그래서 최근에 이와 관련된 상업

1) 오류값: 측정이나 입력오류로 인해 잘못 기록된 값.
2) 특이값: 조사의 대상이 되는 모집단에 속하지 않는다고 의심이 될 정도로 정상 범위 밖으로 아주 동떨어진 관측값.
3) 영향 관측값: 어떤 관측값이 모수추정을 위한 통계량의 값에 큰 영향을 주어서 모수의 추정량에 큰 차이의 결과를 발생하는 관측값.
4) 데이터베이스마케팅: 고객에 관한 데이터베이스를 구축·활용하는 판매전략.
5) 고객관계관리: 고객과 관련된 기업의 내외부 자료를 분석, 통합하여 고객 특성에 기초한 마케팅 활동을 계획하고, 지원하며, 평가하는 과정.
6) 데이터 웨어하우스: 사용자의 의사 결정에 도움을 주기 위해 다양한 운영 시스템에서 추출, 변환, 통합되고 요약된 데이터베이스.

용 도구들이 활발히 소개되고 있으며 국내에서도 다양한 분야에서 데이터 마이닝 작업이 이루어질 전망이다.

정보화 시대의 기업경영에서 정보란 전통적인 자원들을 효과적으로 운영·관리하며 새로운 제품이나 서비스를 창출하는 역할을 담당하는 또 다른 자원으로서, 이것을 제대로 수집하고 활용할 수 있는 기업만이 경쟁에서 살아남고 우위를 확보하게 되는 것이다. 데이터 관점에서 엄밀하게 정의하면 정보(information)란 데이터(data)를 가공·처리해서 얻어지는 산출물이고, 지식(knowledge)이란 정보의 체계적인 활용을 통해 축적된 노하우이다[2, p. 21-22].

1960년대부터 수학적 개념을 기초로 하여 컴퓨터가 등장한 초기부터, 데이터를 과학적으로 저장하고 조작할 수 있는 데이터베이스가 창조되었다. 최초의 범용 데이터베이스시스템[7](DBMS)은 1960년대 초반 General Electric사의 Charles Bachman에 의해 설계되었고 Integrated Data Store라고 불렸으며 네트워크 데이터 모델의 근간이 되었다. 1960년대 말, IBM은 Information Management System이라는 DBMS를 개발하였으며 계층 데이터 모델의 근간이 되었다. 1970년대에 IBM의 산 호세 연구소(San Jose Research Laboratory)의 Edgar Codd는 수학 개념을 이용하여 관계 데이터 모델을 제시했으며 이것으로 인하여 DBMS는 학문 분야로 성숙하게 되었다.

즉, 1970년대부터 체계적으로 진화되어서 1980년 후반부터 데이터 웨어하우징과 데이터 마이닝으로 발전되었다. 데이터 마이닝의 등장배경으로서 정보화 시대의 도래로 인한 정보기술의 가속화 발전, 그것으로 인한 전통적인 전문가 시스템의 한계, 수많은 데이터와 그 속에서의 필요한 정보의 부재를 통하여 기업들은 보다 더 전문화된 정보기술을 이용하여 데이터를 여과하고, 분석하며, 결과를 해석하는 자동화 된 데이터 분석방안에 높은 관심을 갖게 되었다.

2. 수학적 개념을 이용한 데이터베이스의 지식탐사 절차

데이터 모델은 구조와 제약을 정의해야 될 뿐 아니라 데이터를 조작하기 위한 연산(operation)의 정의도 포함해야 된다. 관계 데이터 모델에서의 릴레이션(relation)을 조작하기 위한 기본 연산에는 관계 대수(relational algebra)와 관계 해석(relational calculus)이 있다.

관계 대수는 원하는 목표 데이터를 얻기 위해서 어떻게 해야 되는지 일련의 연산을 순서적으로 명시해야 되는 반면에, 관계 해석은 무슨 데이터를 원하는지만 선언하면

7) 데이터베이스시스템: 컴퓨터에 수록한 수많은 자료들을 쉽고 빠르게 추가·수정·삭제할 수 있도록 해주는 소프트웨어.

된다. 즉 관계 대수는 절차 언어이며 관계 해석은 비절차 언어이다.

기본적인 관계 대수 연산은 수학적 집합 이론으로부터 나온 일반 집합 연산(set operations)과 관계 데이터베이스에 적용할 수 있도록 특별히 개발한 순수 관계 연산들이 있다.

관계 데이터 베이스에서 사용되는 수학적 집합 연산은 합집합(union), 교집합(intersection), 차집합(difference), 카티션 프로덕트(cartesian product)가 있으며 관계 데이터 베이스에서 적용할 수 있는 순수 관계 연산자(relational operations)들로서 실렉트(select), 프로젝트(project), 조인(join), 디비젼(division) 등이 있다. 그리고 관계 대수를 확장하여 세미조인(semijion)과 외부조인(outerjoin), 외부합집합(outer-union), 집단연산 등이 있다. 수학적인 집단연산으로서 SUM, AVG, MAX, MIN, COUNT 등이 있다.

위의 제시한 수학적인 개념과 통계를 적절히 이용하여 데이터의 선택, 데이터의 정제, 데이터의 통합, 데이터의 변환, 데이터의 축소를 통하여 원하는 정보를 얻을 수 있다[1, p. 154-186].

2.1 데이터의 선택(data selection)

대규모 데이터베이스와 데이터 웨어하우스의 지식탐사 절차에서 정확한 데이터를 선택하는 과정은 데이터의 잡음[8](noise) 발생과 일부 데이터의 분실로 인하여 모순을 내포하고 일치성이 없기에 불완전성을 내포하는 특징이 있다.

불완전한 데이터는 주로 데이터를 입력할 때, 여러 가지 이유로 발생할 수 있으며 이해부족, 장비 고장, 데이터 전송 時 오류발생 또는 부적절한 수집도구로 인하여 누락·분실되어 발생된다.

데이터를 선택할 때, 일부 데이터의 파손에 대한 저항성을 가져야 하므로 통계학의 탐색적 자료분석(EDA: exploratory data analysis)을 이용하여 단점을 극복할 수 있다고 판단된다. 대용량의 데이터베이스에서 상관분석(correlation analysis) 또는 회귀분석(regression analysis)의 변수선택방법, 즉 최대결정계수선택법, C_p통계량, 모든 가능한 회귀(all possible regression), 뒤로부터 제거(backward elimination), 앞으로부터 제거(forward selection), 단계별 회귀(stepwise regression)를 이용하여 필요한 데이터들의 항목들을 선택할 수 있다고 판단된다.

8) 잡음: 공학에서 유래된 용어로서 자료계열 안에 존재하는 랜덤성을 말한다. 계열의 실질적인 유형을 식별할 수 있도록 하기 위해서는 잡음을 제거해야 한다.

2.2 데이터의 정제(data cleaning)

지식탐사 절차에서 실세계의 데이터들은 잡음(noise)이 섞여있고 일부가 분실될 수 있으며 일관성이 없고 불완전하므로 데이터의 정제 단계를 거쳐 교정해야 한다.

결측치(missing data)를 정제하는 방법으로 수동으로 채워 넣는 방법, 전역상수(global constant)를 사용하여 채워 넣는 방법, 속성 평균을 사용하여 채워 넣는 방법, 가장 가능성이 높은 값을 사용하여 채워 넣는 방법 등이 있다[5, p. 152].

잡음 섞인 데이터를 정제하는 방법으로 시계열 평활법(smoothing method)을 이용하여 잡음으로 어지럽혀진 데이터베이스로부터 잡음부분을 제거하여 모수(parameter)를 매끈하게 분리해내는 통계적 기법이 있다.

탐색적 자료분석에서 윈도[9](window)와 가중최소제곱법(weighted least squares method)으로 구성된 산점도 평활(scatterplot smoothing)을 이용하여 잡음을 제거할 수 있다. 비모수 통계학을 이용한 비모수회귀 분석(nonparametric regression)에서 극소 띠너비(local bandwidth)를 사용하는 방법으로서 극소적응평활(local adaptive smoothing)을 이용하여 잡음의 제거가 가능하다.

불일치 데이터를 정제하는 방법은 결측치가 없도록 임의의 데이터를 채워놓고 평활기법을 이용하여 잡음을 제거한 후, 외부참조를 통해 불일치성을 교정할 수 있다.

데이터의 정제 방법으로 결측치가 없는 데이터베이스로 이루어진 레코드만을 분석할 수 있는 완전 사례분석(complete case analysis) 또는 오류값이나 결측치를 다른 값으로 대체하여 사용하는 군집분석의 K-평균 군집화 기법 그리고 시각화 기법을 이용한 데이터 보정 방법이 있다[2, p. 88].

2.3 데이터의 통합(data integration)

데이터 통합 과정에서 모여진 데이터들을 데이터 웨어하우스에 하나의 통일된 데이터 저장소로 융합할 때 하나의 주어진 개념을 표현하고 있는 몇몇 속성들이 여러 데이터베이스에서 서로 다른 이름들을 갖고 있어서 불일치와 중복을 초래할 수 있다. 그리고 중복 데이터를 제거하는 작업을 통한 데이터의 항목별 정확도는 데이터 마이닝을 적용하려고 하는 조직의 성격에 따라 상당한 차이를 보이는 경우가 많다.

데이터 정제 이후 데이터 통합의 결과로 생길 수도 있는 중복들을 탐지하고 제거하기 위해서 추가적인 데이터 정제 작업이 수행될 수 있다.

그리고 중복은 상관분석(correlation analysis)을 이용하여 분석할 수 있다.

9) 윈도: 산점도의 일부만을 볼 수 있게 열어 놓은 창틀.

2.4 데이터의 변환(data transformation)

데이터 변환에서는 정제와 통합의 단계를 거쳐 데이터 마이닝에 적절한 형태로 데이터를 변환시키는 단계이다. 데이터의 변환은 보편적인 규칙이 존재하지 않기 때문에 전문가의 주관적인 판단과 경험에 의존한다.

데이터를 정규화(normalization)에 의하여 변환함으로서, 구간 [0,1]에서 일정비율로 척도화하여 신경망 기법, 최단이웃분류, 군집분석에 사용할 수 있다.

탐색적 자료분석의 박스-칵스 변환(Box-Cox transformation)을 이용한 데이터의 재표현(re-expression)을 통해서 자료분석을 용이하게 제시할 수 있다고 판단된다.

2.5 데이터의 축소(data reduction)

데이터의 축소(data reduction)는 중복 특징 제거, 군집화(clustering)등을 통하여 차원의 축소와 데이터의 압축 등을 이용하여 데이터의 크기를 줄일 수 있다.

데이터베이스의 차원축소는 다중공선성(multicollinearity)과 '2.1 데이터의 선택'에서 제시한 회귀분석의 변수선택방법을 이용하여 불필요한 데이터를 제거함으로 차원을 축소할 수 있다고 판단된다.

데이터의 압축은 웨이브렛 변환(wavelet transformation)과 확률변수사이의 분산-공분산 관계를 이용하여 확률변수들의 일차결합으로 만드는 다변량분석법인 주성분분석(principal component analysis)을 사용한다.

2.6 데이터 마이닝 기법의 선택 및 적용 (selection and application of data mining techniques)

(1) 고전적 기법: 통계를 이용한 데이터 마이닝의 주요 기법

데이터의 수집과 설명에 관련된 통계 기법을 통하여 데이터베이스에는 어떤 의미있는 패턴들이 내재되어 있으며 어떤 사건이 일어날 가능성을 예측함으로서 보다나은 의사결정을 내릴 수 있다.

데이터베이스에 대한 정보를 쉽게 이해할 수 있는 방법으로 히스토그램을 통하여 기술 통계량(descriptive statistics)을 얻을 수 있으며, 예측을 위해서는 회귀분석(regression analysis)이 사용된다.

데이터 마이닝 기법 중에서 예측되어야 할 임의의 레코드와 과거에 사용된 유사한 데이터베이스에서 레코드의 예측값을 발견하여 이것을 이용하는 예측기법인 근접이웃(nearest neighbor)과 n차원 공간 안에서 레코드들간의 위치와 연결성 등에 근거하여 비슷한 레코드를 하나의 그룹으로 만드는 예측기법인 군집분석(clustering analysis)이 있다. 군집분석은 동일한 성질과 가장 작은 수의 군집으로 이루어질 때가 최적이며 계층적 군집분석과 비계층적 군집분석이 있다.

(2) 차세대 기법: 정보기술을 이용한 데이터 마이닝의 주요 기법

분류 목적으로 사용되는 지식발견(knowledge discovery) 기법으로서 예측모델과 같은 나무를 형성하는 일종의 데이터 마이닝 또는 통계적 방법의 한 부류로서 데이터 세분화를 위해 사용되는 의사결정나무(decision tree)가 있다.

주로 회귀 목적으로 사용되는 기계학습(machine learning) 기법으로서 인간의 뇌 그리고 신경세포가 반응하는 것과 유사하게 설계된 회로로서 다수의 마디를 네트웍으로 연결하고 각 마디들간의 연결의 세기로 정보를 표현하고 기억하는 신경망(neural network)이 있다.

비슷한 상품들을 찾아내는 친화 그룹화(affinity grouping)의 한 방법으로서 데이터베이스 테이블의 특성 속성들간의 통계적 상호 관련성을 나타내는 규칙으로 연관성 규칙(association rule)이 있다. 연관규칙의 일반적인 형식은 $X_1, X_2, \cdots, X_n \Rightarrow Y$로 표현되고, 속성 $X_1, X_2, \cdots, X_n$이 Y를 예측한다는 의미로 해석한다.

진화론에 기반을 둔 기계학습 알고리즘의 종류로서 다윈의 자연선택과 적자생존의 생물학적 모델에 근거한 병렬탐색을 이용하여 최적화 문제를 해결하는 방법으로 유전자 알고리즘(genetic algorithm)이 있다.

과거의 경험과 해법을 활용함으로서 주어진 문제를 해결하는 기법으로 사례기반추론(case-based reasoning), 특정분야에서 인간 전문가와 같은 수준의 지능과 경험을 가진 컴퓨터 시스템인 전문가 시스템, 패턴인식, 퍼지논리 시스템 등이 있다.

데이터 마이닝의 기술은 초기단계로서 데이터 마이닝 패키지는 기존의 OLAP이나 SAS, SPSS, S-PLUS와 같은 통계 기능에 의존하고 있으며 데이터 마이닝의 주요기법을 활용한 툴(tool)이 완전하게 개발되지 못한 상태여서 데이터 마이닝 패키지의 개발이 필요하다.

현재까지 개발되어 상용화되는 데이터 마이닝 툴은 IBM의 Intelligent Miner, AT & T Bell Lab의 IMACS, SAS의 Enterprise Miner, SPSS의 Clementine, 캐나다 Simon Fraser대학의 DBMiner 등이 있다.

그러나 데이터 마이닝 툴은 고가(高價)이며 데이터 마이닝의 주요 관점인 대용량의 데이터로부터 유용한 정보를 자동화 기법(technique)을 통하여 인간의 경험이나 논리로는 발견할 수 없었던 새로운 지식패턴을 추출하는 방법이지만 아직까지 해당분야의 분석전문가의 기술적인 지원이 필요하기에 이러한 단점을 해결해서 대기업뿐만 아니라 중소기업까지도 활용할 수 있는 접근 방법을 개발해야된다.

2.7 모형의 평가(model evaluation)

데이터베이스에서 기대보다 빈번히 발생하는 패턴을 발견하고 임의 표본추출을 통하여 데이터 마이닝에 의해 만들어진 예측 모델을 검증하는 것이다.

3. 데이터 마이닝의 응용분야 및 적용사례

데이터 마이닝은 다양한 응용 분야를 가진 새로운 학문영역으로서 활용 가능성이 무궁무진하다고 판단된다. 현재 데이터 마이닝의 개념 및 기법을 제품 등에 도입한 외국의 예를 살펴보면 다음과 같다.

국내의 경우 데이터 마이닝을 이용한 응용분야는 아직 초기 도입단계이며 앞으로 더욱 새로운 분야에 적용될 것으로 판단된다.

- 생물의약과 DNA 자료분석을 위한 데이터 마이닝:

이질적, 분산된 게놈 데이터베이스의 의미적 통합, DNA 순서에서 유사성 조사와 비교, 연관분석을 이용한 동시 발생한 유전자 순차들의 확인, 경로분석을 이용한 질병 발달의 다른 단계를 연결하는 유전자들 확인, 시각화 도구와 유전학 데이터분석, 환자의 질병 진단 또는 질병의 예후 분석, 환자의 특성에 따른 의약품의 부작용 분석

- 금융 데이터 분석을 위한 데이터 마이닝:

다차원 데이터 분석과 데이터 마이닝을 위한 데이터 웨어하우스의 설계와 구축, 대부상환 예상과 고객 신용정책 분석, 목표 마케팅을 위한 고객의 분류와 군집화, 돈 세탁과 다른 금융 범죄 탐지, 고객분류를 통한 보험료 가격 정책 수립, 보험료 청구 사기 패턴 추적, 클래임 처리시간에 영향을 미치는 요소 발견, 신용카드 도용패턴 추적, 이탈 예상고객 선정 및 특성 분석, 우수고객 선정 및 특성 분석, 서비스별 홍보 대상고객 선정, 신용평가 모형개발, 신용평가 모형 개발, 주식 거래규칙 발견

- 소매업을 위한 데이터 마이닝:

데이터 마이닝의 이점에 기초를 둔 데이터 웨어하우스의 설계와 구축, 판매·고객·상품·시간과 지역의 다차원 분석, 판매 캠페인의 효과 분석, 고객 유지-고객 신용 분석, 구매추천과 상품의 교차 추천, 고객의 구매패턴과 선호도 발견, 제품/서비스 교차판매, 판매 실적에 영향을 미치는 요소 발견, 고객 분류, 그룹별 특성 발견, 광고·프로모션·이벤트의 효과 측정

- 제조 및 유통업을 위한 데이터 마이닝:

최종 생산품의 품질에 영향을 미치는 요인 발견, 경쟁사의 입찰액 예측, 제품의 수요예측, 대리점 여신평가 모형 개발, 매장진열 전략 수립, 상품 카달로그 디자인, 상품 교차판매

- 통신사업자를 위한 데이터 마이닝:

통신 데이터의 다차원 분석, 장거리 전화/무선 전화의 부정한 이용 패턴분석과 특이 패턴의 확인, 다차원 연관과 순차 패턴분석, 통신 데이터분석에서 시각화도구의 활용, 이탈 예상고객 선정 및 특성 분석, 서비스간의 연관관계 발견, 우수고객 선정 및 특성 분석([2, p. 40], [5, p. 545-552])

4. 수학적 관점에서 본 데이터 마이닝과 통계의 비교 분석

데이터 마이닝은 연관규칙탐사(association rules), 의사결정나무(decision trees), 신경망(neural networks) 모형의 기법을 이용한 기계학습(machine learning)·패턴인식(pattern recognition)·데이터베이스(database)에 관한 전산학의 영역이기도 하며, 군집분석(clustering analysis)에 관한 수학의 한 분야인 통계학의 영역이기도 하며, 기업과 고객간의 상호접촉을 통한 수익성 향상을 도모하는 고객관계관리(CRM)에 관한 경영학의 영역으로서 그 밖의 다양한 분야의 방법론들로 이루어져 있다.

통계학은 데이터의 계획, 수집, 처리 및 활용에 관한 인간의 지식 체계이다. 통계학의 학문영역은 데이터와 관련된 모든 것이다. 유용한 정보를 찾아내기 위해 데이터 마이닝 도구들이 사용하는 알고리즘들은 통계에서 사용되고 있는 기본적인 기법(회귀분석, 일반선형모형, 분산분석, 인자분석, 판별분석, 시계열분석, 생존분석과 품질관리 등)들로 유도되었다. 그러나 통계도구를 이용한 대용량의 데이터 분석은 실행시간이 오래 걸리고, 정확하지 않기에 통계전문가들에 의존해야 하는 단점이 있다.

모집단의 분포가 불투명하거나 정규분포가 아닌 경우, 측정된 데이터가 순서 또는

몇 개의 범주로 구분되는 정보만을 제공하는 경우, 양적으로 나타낼 수 없어서 순위만으로 관측된 경우, 주어진 자료에 특이점(outlier)이 있는 경우에 데이터를 관찰하여 데이터의 특성과 구조적 관계를 발견하여 정확히 정보를 찾아낼 수 있는 비모수 통계학의 방법론을 적용함으로서 통계를 이용한 보다 정확한 데이터 분석기법의 향상을 도모할 수 있다.

또한 통계학의 한 분야인 탐색적자료분석(exploratory data analysis)을 통하여 데이터 마이닝의 중요한 관점인 예외적인 데이터를 가정이나 모형에 의지하지 않고 데이터 패턴을 발견할 수 있다.

데이터 마이닝은 통계적인 과정을 효과적으로 자동화하여 통계전문가가 아닌 일반적인 기업에 종사하는 최종 사용자들이 겪을 수 있는 어려움을 최소화하여 쉽게 사용할 수 있는 도구들을 제공하고, 그것으로 인하여 데이터 분석 및 결과를 쉽게 이해할 수 있는 방법을 제시함을 목표로 한다. 그러기 위해서 보다 정확하고 과학적인 비모수통계(nonparametric statistics)를 이용하여 주어진 자료로부터 가정 없이 정보를 추출할 수 있는 데이터 마이닝이 필요하다고 판단된다.

5. 결론

데이터가 증가할수록 정보는 감소한다. 기존의 질의 도구 SQL[10](structured query language) 언어를 사용하여 데이터 집합에 적용하여도 개략적인 분석을 통해 관심 있는 정보의 80%정도를 얻을 수 있다. 숨겨진 데이터를 SQL을 이용하여 찾을 수 없기에 데이터베이스의 지식탐사 절차의 분석 알고리즘을 이용하여 숨겨진 정보 20%를 얻을 수 있으며 이 20%가 매우 중요한 정보라고 기업에서 판명된 사실이다.

오늘날 정보가 중요한 생산 요소이기 때문에 데이터의 기계적인 생산과 재생산으로 인하여 지식탐사 절차를 계속 수정·개발함으로서 정제된 데이터를 추출하여 자동화된 방법으로 데이터를 해석할 수 있는 기법이 시대적으로 요구되고 있는 상황이다.

그러므로 데이터의 패턴에 의존하는 통계적 기법을 적용하면, 업무 현장의 다양한 특성을 반영하고 기업들이 발견하지 못한 전략 수립에 필요한 귀중한 정보를 발견하여 과학적으로 입증되고 수학을 기반으로 한 통계적 방법론을 제공할 수 있을 것이라고 판단된다.

10) SQL: 관계형 데이터베이스의 조작과 관리에 사용하는 데이터베이스 하부 언어(sub-language).

참고 문헌

1. 이석호, 데이터베이스 시스템, 정익사, 2004.
2. 장남식 · 홍성완 · 장재호, 테이터 마이닝, 대청미디어, 1999.
3. 허명회 · 이용구, 테이터 마이닝 모델링과 사례, SPSS 아카데미, 2003.
4. Alex Berson · Stephen Smith · Kert Therling · Entrue Consulting CRM 그룹/ 홍성완 외 역, CRM을 위한 테이터 마이닝, 대청미디어, 2000.
5. Jiawei Han/박우창 · 승현우 · 용환승 · 치기헌 역, 테이터 마이닝 개념 및 기법, 자유아카데미, 2003.
6. Pieter Adriaans · Dolf Zantinge/용환승 역, 테이터 마이닝, 그린, 1998.
7. Ramakrishnan, Gehrke, *Database Management Systems*, McGraw-Hill, 2003.

The Korean Journal for History of Mathematics Vol. 18 No. 2 (May 2005), 95-106

Mathematical Foundations and Educational Methodology of Data Mining

Department of Software, Seokyeong University **Seung-Woo Lee**

This paper is investigated conception and methodology of data selection, cleaning, integration, transformation, reduction, selection and application of data mining techniques, and model evaluation during procedure of the knowledge discovery in database (KDD) based on Mathematics. Statistical role and methodology in KDD is studied as branch of Mathematics.

Also, we investigate the history, mathematical background, important modeling techniques using statistics and information, practical applied field and entire examples of data mining. Also we study the differences between data mining and statistics.

Key words: knowledge discovery in database, data mining, statistics

2000 Mathematics Subject Classification: 97C90, ZDM Subject Classification: K95

논문 접수 : 2005년 3월 2일, 심사 완료 : 2005년 4월