응용통계연구 제 **15**권 2호, *2002*년, *pp. 201–211*

# 데이터마이닝 소프트웨어의 기능 및 효율성 비교에 관한 사례연구

한상태[1] 강현철[2] 이성건[3] 이덕기[4]

요 약

최근 정보기술 분야의 급속한 발전과 더불어 기업 및 사회 각 분야의 데이터베이스에 쌓이고 있는 데이터의 양도 급격히 증가하고 있다. 이러한 관점에서 데이터마이닝이 큰 주목을 받고 있으며, 다양한 데이터마이닝 기법들을 이용하여 여러 가지 분석을 보다 손쉽게 수행할 수 있는 상용화된 소프트웨어들이 많이 개발되고 있다. 이들 데이터마이닝 소프트웨어들은 여러 가지 관점에서 서로 다른 모습을 가지고 있는데, 따라서 이들의 기능과 성능은 많은 사용자들의 큰 관심이 되고 있다. 본 연구에서는 현재 널리 사용되고 있는 몇 가지 데이터마이닝 소프트웨어들에 대해 기능상의 차이점 및 실제 사용에 있어서의 효율 등을 비교하고자 한다.

주요용어: 데이터마이닝, 모델링 알고리즘, 정확도

## 1. 서론

최근 사회 각 분야에 쌓이고 있는 데이터의 기하급수적인 증가로 인해 데이터마이닝이 한층 그 중요성을 더해가고 있으며 상용화된 데이터마이닝 소프트웨어의 수도 계속 증가하고 있다. 이들 데이터마이닝 소프트웨어들은 가지고 있는 기능이나 성능에 있어서 서로 상이한 점들을 지니고 있으며, 따라서 일반 사용자들은 소프트웨어의 선택이나 실제 사용에 큰 어려움을 겪고 있다. 그러나 데이터마이닝 소프트웨어들을 다양한 관점에서 비교함으로서 사용자가 자신의 목적과 환경에 맞는 소프트웨어를 선택하고 그 기능들을 충분히 활용함에 있어서 도움을 줄만한 연구가 제대로 되어있지 않은 것이 현실이다.

King & Elder(1998)는 14개 데이터마이닝 소프트웨어들의 기능 및 효율을 비교한 바 있으며, Abbott et al.(1998)은 Clementine, Darwin, E-Miner, I-Miner, PRW 등 주요 다섯 개의 데이터마이닝 소프트웨들을 비교하고 있다(다만 이들 연구는 초기 버전의 소프트웨들을 비교대상으로 하고 있으며, 소프트웨어들이 매우 빠르게 발전하고 있다는 점을 고려하여 참조하기 바란다). 또한 Choi & Lee(2001) 및 Choi et al.(2001) 등의 연구에서도 본 연구에서 다루고 있는 소프트웨어들에 대해 몇 가지 측면에서 비교하고 있다.

1) (336-795) 충청남도 아산시 배방면 세출리 산 29-1, 호서대학교 자연과학부 정보통계학전공, 조교수
E-mail: sthan@office.hoseo.ac.kr
2) (336-795) 충청남도 아산시 배방면 세출리 산 29-1, 호서대학교 자연과학부 정보통계학전공, 전임강사
E-mail: hychkang@office.hoseo.ac.kr
3) (136-701) 서울시 성북구 안암동 5가 1, 고려대학교 대학원 통계학과 박사과정
4) (336-795) 충청남도 아산시 배방면 세출리 산 29-1, 호서대학교 대학원 수학과 석사과정

본 연구에서는 데이터마이닝 소프트웨어들에 대한 다양한 비교를 통하여 각 소프트웨어에 대한 구체적 성능과 효율을 파악함으로서 데이터마이닝 소프트웨어를 활용하고자 하는 일반 사용자 및 연구자들에게 유익한 기초 정보를 제공해 주고자 한다. 본 연구에서 비교대상에 포함시키고 있는 소프트웨어들은 SPSS Clementine(버전 6.02), SAS Enterprise Miner(버전 4.0 : 이후 E-Miner), IBM Intelligent Miner for Data(버전 6.1 : 이후 I-Miner)이며, 이들은 현재 가장 널리 사용되고 있는 대표적인 데이터마이닝 소프트웨어들이다.

## 2. 운영환경 및 특징에 대한 비교

이 절에서는 먼저 각 데이터마이닝 소프트웨어들의 운영환경 및 지원되는 데이터베이스 엔진 등에 대해서 간단히 비교해 보고자 한다.

Clementine(SPSS Institute, 2001)은 Windows 95/98, Windows NT, AIX, Solaris, HP-UX, Digital UNIX, IRIX, DGUX, NCR UNIX SVR5 등에 설치하여 사용할 수 있다. 또한 ODBC(open database connectivity)를 지원하는 어떤 형태의 데이터베이스와도 연결이 가능하며 SPSS, Oracle, SAS, MS Excel 데이터들도 사용할 수 있다. Clementine의 특징 중 하나는 개방형 시스템으로서 추가적인 분석을 위하여 SPSS뿐만 아니라 Excel, SAS 등 다른 분석프로그램과도 연결이 가능하고 외부에서 작성된 알고리즘을 추가할 수 있다는 것이다.

E-Miner(SAS Institute, 1999)는 Windows 95/98(클라이언트), Windows NT, AIX, Solaris, HP-UX, Digital Compaq UNIX, OS/390, OS/400, NCR UNIX SVR5, MVS 등 매우 다양한 OS 환경에서 설치하여 사용할 수 있다. 또한 DB2, IMS, ADABAS/MVS, Sybase/UNIX Oracle/UNIX, Informix/UNIX, MS-SQL Sever, Non-Stop SQL, VSAM, MS Excle, SPSS 등 대부분의 관계형 데이터베이스와 직접 연결이 가능하다. E-Miner는 데이터마이닝을 위한 가장 풍부하고 다양한 모델링 기법 및 알고리즘들을 포함하고 있다고 할 수 있다.

I-Miner(IBM Institute, 1999)는 Windows 95/98(클라이언트), Windows NT, AIX, Solaris, OS/390, OS/400, MVS 등에 설치하여 사용할 수 있으며, DB2 및 UDB 등 IBM 계열의 데이터베이스들과 연결하여 사용할 수 있다. I-Miner의 특징 중 하나는 병렬 데이터마이닝을 지원한다는 것이다.

## 3. 구조 및 노드(오브젝트)의 기능에 대한 비교

세 개의 데이터마이닝 소프트웨어들은 메인화면의 구조에 있어서 유사한 형태를 가지고 있다. 즉 데이터베이스와의 연결, 데이터 탐색, 변형, 모델링, 모형평가 등 데이터마이닝에 필요한 여러 기능들이 노드(node) 또는 오브젝트(object)라는 이름으로 모듈화되어 소프트웨어 내에 내장되어 있으며, 사용자는 특별한 작업공간에서 이러한 모듈들을 선택하고 연결함으로서 각자의 목적에 맞는 데이터마이닝 작업을 수행할 수 있다.

### 3.1. 메인화면의 구조

Clementine의 메인화면은 그림 3.1에서와 같이 여러 개의 패널로 구성되어 있다. 이들 중 Palettes 패널은 6개로 구분된 노드들의 모임으로 이루어져 있는데, 각 노드들은 데이터의 입출력, 탐색, 변형, 모델링 등을 수행할 수 있는 기능들을 가지고 있다. 이 6개의 노드들 중 Source 노드는 처음 노드이고, 레코드 Ops 노드와 필드 Ops 노드는 중간 노드이며, 그래픽 노드와 모델링 노드 및 Output 노드는 최종 노드이다. Stream 패널은 Palettes 패널로부터 노드들을 선택하여 연결하는 일종의 작업공간으로, 이 때 노드들은 처음 노드, 중간 노드, 최종 노드를 순서대로 연결해야 하고 노드들의 연결과정을 간략하게 해주는 슈퍼노드를 사용할 수 있다. 그리고 Generated Model 패널에서는 모델링의 결과를 볼 수 있다.

E-Miner의 메인화면은 그림 3.2에서와 같이 프로젝트 윈도우, 툴 윈도우, 분석흐름도 윈도우 등 여러 개의 윈도우들로 이루어져 있다. 이들 윈도우 중 프로젝트 윈도우는 프로젝트 및 분석흐름도를 전반적으로 관리하는 일종의 탐색기이고, 툴 윈도우는 표본추출(sampling), 탐색(exploration), 변형(modification), 모델링(modeling), 평가(assessment) 등 일련의 데이터마이닝 작업들을 수행할 수 있는 노드들로 구성되어 있다. 그리고 분석흐름도 윈도우는 툴 윈도우로부터 노드들을 선택하여 데이터마이닝 분석흐름도를 작성하는 일종의 작업공간이다.

I-Miner의 메인화면은 그림 3.3에서와 같이 마이닝베이스에 저장되어 있는 오브젝트 유형을 보여주는 마이닝베이스 컨테이너와 오브젝트의 각 유형에 대한 폴더의 내용을 볼 수 있는 폴더내용 컨테이너, 그리고 오브젝트 폴더의 작업을 관할하는 작업영역으로 구성되어 있다.

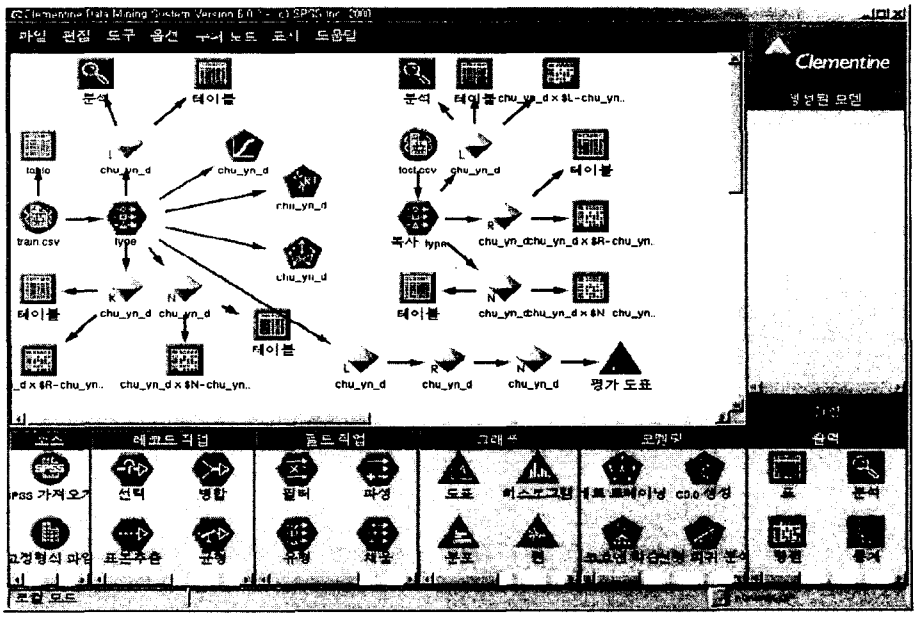


그림 3.1: Clementine의 메인화면

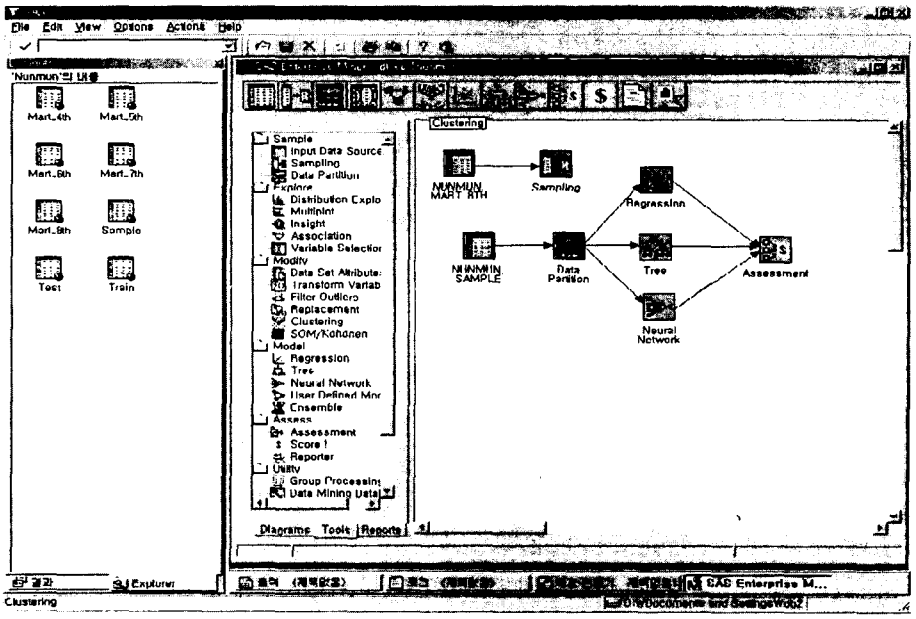


그림 3.2: E-Miner의 메인화면

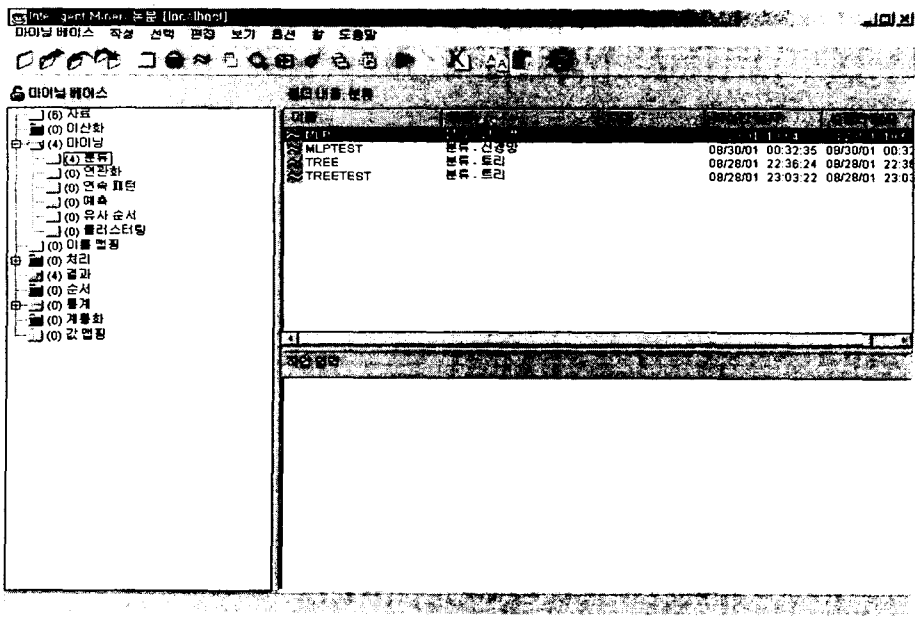


그림 3.3: I-Miner의 메인화면

표 3.1: 데이터마이닝 소프트웨어들의 기능 비교

| 노드 | | Clemen- | E-Miner | I-Miner |
|---|---|---|---|---|
| 데이터 추출 및 관리(변형) | Data Load | O | O | O |
| | Data Parition | SPSS | O | O |
| | Balance | O | SAS | O |
| | Merge | O | SAS | O |
| | Sample | O | O | O |
| | Select | O | O | O |
| | Sort | O | SAS | O |
| | Aggregate | O | SAS | O |
| | Append | O | SAS | O |
| | Distinct | O | SAS | O |
| | Imputation | O | O | O |
| | Filter | O | O | O |
| | Derive | O | O | O |
| | History | O | O | O |
| 데이터 탐색 | Distribution | O | O | × |
| | Histogram | O | O | × |
| | 3-D Plot | SPSS | O | × |
| 모델링 알고리즘 | Neural Network | O | O | O |
| | Decision Tree | O | O | O |
| | Clustering | O | O | O |
| | Regression | O | O | O |
| | Association Rule | O | O | O |
| | Factor Analysis | O | SAS | O |
| 모형의 평가 및 결과보기 | Assessment | O | O | O |
| | Table | O | O | O |
| | Report | O | O | × |
| | Data Export | O | SAS | DB2 |

### 3.2. 메인화면의 구조와 노드(오브젝트)들의 기능 비교

먼저 메인화면의 구조를 살펴보면 세 가지 소프트웨어 모두 각각의 기능을 담당하는 노드(또는 오브젝트)들과 이들을 이용하여 필요한 작업을 수행하는 작업공간으로 구성되어 있으며, 작업공간에는 분석흐름도를 시각적으로 표현할 수 있다. 특히 Clementine과 E-Miner는 분석흐름도를 한눈에 파악할 수 있는 장점이 있었으며, I-Miner는 여러 오브젝트 유형을 하나의 오브젝트로 단일화하여 데이터마이닝 작업을 수행 할 수 있는 편리함을 가지고 있다.

다음으로 각 소프트웨어들에 내장되어 있는 노드들의 기능에 대하여 살펴보면, I-Miner가 데이터 분포탐색을 포함하지 않는다는 것만을 제외하고는 데이터마이닝을 수행하기 위해 필요한 기능들을 폭넓게 가지고 있는 것을 알 수 있다(표 3.1 참조). 따라서 이들 소프트웨어들이 가지고 있는 기능상의 차이점은 크지 않은 것으로 보인다.

표 3.2: 데이터마이닝 소프트웨어들의 모델링 알고리즘 비교

| 모델링 기법 및 알고리즘 | | 소프트웨어 | | |
|---|---|---|---|---|
| | | Clementine | E-Miner | I-Miner |
| 연관성분석 | Association Rule | O | O | O |
| | Sequential Association | × | O | O |
| 군집분석 | k-Means | O | O | × |
| | Kohonen Map | O | O | O |
| | TwoStep | O | × | × |
| | Demographic | × | × | O |
| 의사결정 나무분석 | CHAID | × | O | × |
| | CART | O | O | × |
| | C4.5(C5.0) | O | O | × |
| | SPRINT | × | × | O |
| 신경망분석 | MLP | O | O | O |
| | RBF | O | O | O |
| 회귀분석 | Linear Regression | O | O | O |
| | Logistic Regression | O | O | × |

그러나 이와 같은 직접적인 비교에는 약간의 문제점이 있는데, 특히 데이터의 관리에 대해서 이들 소프트웨들은 다른 자료분석 소프트웨어나 데이터베이스 엔진과 밀접한 관련을 가지고 있기 때문이다. 이들 데이터마이닝 소프트웨어들이 모두 표준 SQL(structured query language)을 지원하고 있으며 앞절에서 서술한 바와 같이 다양한 데이터베이스와 연결하는 것이 가능하지만, E-Miner는 SAS 소프트웨어를 기반으로 하여 실행되며(즉, SAS와 독립적으로 사용할 수 없다) 데이터의 입출력, 변형, 결과의 출력 등이 SAS의 여러 모듈들과 연결되어 있다. 또한 Clementine 데이터의 입출력이나 결과의 출력 및 보고서 작성 등에 있어서 SPSS와, 그리고 I-Miner는 DB2와 밀접한 관련을 가지고 있다. 따라서 사용자의 작업환경이나 다른 소프트웨어 또는 데이터베이스에 대한 숙련도 등도 데이터마이닝 소프트웨어의 선택에 중요한 고려사항 중 하나라고 할 수 있다.

### 3.3. 모델링 알고리즘에 대한 비교

표 3.2는 세 가지 데이터마이닝 소프트웨어에 포함되어 있는 모델링 기법 알고리즘을 정리한 것이다. 이 표에서 알 수 있듯이 각 소프트웨어가 자신만의 독특한 알고리즘을 가지고 있는 경우도 있지만(예를 들어, TwoStep과 Demographic은 각각 SPSS와 IBM에서 개발한 군집분석 알고리즘이다), 전반적으로 모델링에 필요한 다양한 알고리즘을 가지고 있다. 다만 I-Miner의 경우 군집분석의 k-Means 알고리즘이나 로지스틱 회귀분석과 같은 통계학 분야에서 전통적으로 널리 사용되어져 온 모델링 알고리즘들이 포함되어 있지는 않다.

표 4.1: 모델링 정확도

| 모델링 기법 | | | 분석용 데이터 | | | 검증용 데이터 | | |
|---|---|---|---|---|---|---|---|---|
| | | | Clemen- | E-Miner | I-Miner | Clemen- | E-Miner | I-Miner |
| 로지스틱 회귀분석 | | 정분류율 | 73.87% | 73.89% | × | 73.95% | 73.73% | × |
| | | 민감도 | 87.64% | 87.91% | × | 87.40% | 88.11% | × |
| | | 특이도 | 45.32% | 44.83% | × | 63.52% | 43.84% | × |
| 의사 결정 나무 분석 | CHAID | 정분류율 | × | 79.26% | × | × | 79.61% | × |
| | | 민감도 | × | 91.10% | × | × | 91.16% | × |
| | | 특이도 | × | 54.70% | × | × | 55.61% | × |
| | CART SPRINT | 정분류율 | 78.32% | 79.22% | 85.02% | 77.21% | 79.41% | 78.89% |
| | | 민감도 | 91.37% | 91.41% | 91.00% | 83.79% | 91.31% | 86.11% |
| | | 특이도 | 51.27% | 53.94% | 72.63% | 63.52% | 54.69% | 63.88% |
| | C4.5 C5.0 | 정분류율 | 91.13% | 79.15% | × | 78.71% | 79.35% | × |
| | | 민감도 | 94.31% | 89.93% | × | 91.39% | 89.99% | × |
| | | 특이도 | 84.54% | 56.81% | × | 52.35% | 57.44% | × |
| 신경망 분석 | 은닉층 1 노드 2 | 정분류율 | 75.16% | 72.75% | 74.33% | 74.59% | 72.89% | 73.55% |
| | | 민감도 | 90.04% | 82.23% | 77.49% | 89.48% | 82.23% | 51.61% |
| | | 특이도 | 44.31% | 52.94% | 67.76% | 43.65% | 53.36% | 67.55% |
| | 은닉층 1 노드 3 | 정분류율 | 75.38% | 73.35% | 76.70% | 73.80% | 73.77% | 75.86% |
| | | 민감도 | 86.59% | 84.64% | 87.81% | 85.03% | 85.00% | 87.00% |
| | | 특이도 | 52.15% | 49.95% | 53.69% | 50.47% | 50.42% | 52.72% |

## 4. 실제 사례를 통한 데이터마이닝 소프트웨어의 성능 비교

본 사례분석에 사용된 데이터는 국내 한 은행의 데이터베이스로부터 추출된 것으로, 2000년 2월말 현재 거래 총금액이 10만원 이상으로 세금우대 여유한도가 있는 고객을 대상으로 총 50,000명의 고객을 추출하여 데이터를 구성하였다. 여기서 목표변수(target variable)는 2000년 2월에서 2000년 6월 사이에 세금우대상품을 가입하였는지의 여부이며, 입력변수(input variable)로는 고객의 인구사회적 속성 및 금융거래 속성 등 153개의 변수가 사용되었다.

본 연구에서는 로지스틱 회귀분석, 의사결정나무분석, 신경망분석 등 3가지 모델링 기법을 각각 수행하여 그 결과를 비교하였다. 이 때 로지스틱 회귀분석의 경우 I-Miner는 해당 알고리즘을 가지고 있지 않으므로 비교대상에 포함하지 않았고, 신경망분석에 대해서는 은닉층의 개수를 하나로 하고 은닉노드의 개수가 2개인 경우와 3개인 경우에 대해 비교하였다. 참고로 본 사례분석을 수행한 컴퓨팅 환경은 CPU는 펜티엄 800 듀얼, RAM은 516M, 그리고 OS는 Windows 2000이 사용되었다.

또한 수행결과의 안정성을 비교하기 위하여 데이터의 60%인 30,000개를 분석용으로 사용하였고 나머지 40%인 20,000개를 검증용으로 사용하였으며, 표 4.1은 분석용 데이터와 검증용 데이터에 대해 정분류율(accurate rate), 민감도(sensitivity), 특이도(specificity) 등 정확도의 3가지 측면을 제시한 것이다(정분류율, 민감도, 특이도 등에 대해서는 강현철

표 4.2: 변수 및 개체의 수에 따른 모델링 수행시간 (시간:분)

| | 변수 수 | 개체 수 | 로지스틱 회귀분석 | 의사결정나무분석 | | | 신경망분석 | |
|---|---|---|---|---|---|---|---|---|
| | | | | CHAID | CART | C4.5 | 노드 2 | 노드 3 |
| Clementine | 20 | 10,000 | 0:01 | × | 0:01 | 0:01 | 0:01 | 0:01 |
| | | 20,000 | 0:01 | × | 0:01 | 0:01 | 0:01 | 0:01 |
| | | 50,000 | 0:01 | × | 0:01 | 0:01 | 0:02 | 0:03 |
| | 50 | 10,000 | 0:01 | × | 0:01 | 0:01 | 0:01 | 0:02 |
| | | 20,000 | 0:01 | × | 0:01 | 0:01 | 0:02 | 0:03 |
| | | 50,000 | 0:03 | × | 0:03 | 0:02 | 0:07 | 0:11 |
| | 100 | 10,000 | 0:12 | × | 0:02 | 0:01 | 0:02 | 0:02 |
| | | 20,000 | 0:25 | × | 0:03 | 0:01 | 0:05 | 0:06 |
| | | 50,000 | 1:19 | × | 0:04 | 0:03 | 0:19 | 0:25 |
| | 153 | 10,000 | 1:21 | × | 0:03 | 0:02 | 0:02 | 0:02 |
| | | 20,000 | 2:08 | × | 0:04 | 0:02 | 0:04 | 0:05 |
| | | 50,000 | 3:46 | × | 0:07 | 0:02 | 0:15 | 0:31 |
| E-Miner | 20 | 10,000 | 0:01 | 0:01 | 0:01 | 0:01 | 0:01 | 0:02 |
| | | 20,000 | 0:02 | 0:01 | 0:01 | 0:01 | 0:02 | 0:02 |
| | | 50,000 | 0:02 | 0:01 | 0:01 | 0:01 | 0:03 | 0:03 |
| | 50 | 10,000 | 0:01 | 0:01 | 0:01 | 0:01 | 0:01 | 0:02 |
| | | 20,000 | 0:02 | 0:01 | 0:01 | 0:01 | 0:03 | 0:03 |
| | | 50,000 | 0:02 | 0:01 | 0:01 | 0:01 | 0:04 | 0:05 |
| | 100 | 10,000 | 0:02 | 0:02 | 0:01 | 0:02 | 0:02 | 0:03 |
| | | 20,000 | 0:03 | 0:02 | 0:02 | 0:02 | 0:06 | 0:09 |
| | | 50,000 | 0:03 | 0:02 | 0:02 | 0:02 | 0:08 | 0:17 |
| | 153 | 10,000 | 0:03 | 0:02 | 0:02 | 0:02 | 0:04 | 0:05 |
| | | 20,000 | 0:05 | 0:02 | 0:03 | 0:02 | 0:06 | 0:09 |
| | | 50,000 | 0:06 | 0:03 | 0:03 | 0:03 | 0:14 | 0:18 |
| I-Miner | 20 | 10,000 | × | × | 0:01 | × | 0:07 | 0:08 |
| | | 20,000 | × | × | 0:01 | × | 0:15 | 0:16 |
| | | 50,000 | × | × | 0:01 | × | 0:05 | 0:14 |
| | 50 | 10,000 | × | × | 0:01 | × | 0:18 | 0:19 |
| | | 20,000 | × | × | 0:01 | × | 0:36 | 0:38 |
| | | 50,000 | × | × | 0:01 | × | 0:34 | 0:34 |
| | 100 | 10,000 | × | × | 0:01 | × | 0:39 | 0:40 |
| | | 20,000 | × | × | 0:01 | × | 1:01 | 1:14 |
| | | 50,000 | × | × | 0:02 | × | 1:12 | 1:21 |
| | 153 | 10,000 | × | × | 0:01 | × | 0:59 | 1:01 |
| | | 20,000 | × | × | 0:02 | × | 1:19 | 1:21 |
| | | 50,000 | × | × | 0:03 | × | 1:59 | 2:01 |

등(2001) 및 최종후 등(2001)을 참조하기 바란다). 이들 결과를 살펴보면 정확도의 측면에서 3가지 소프트웨어들 간에 큰 차이를 보이지 않았으며, 분석용 데이터에서의 결과와 검증용 데이터에서의 결과에 큰 차이가 없어 모두 안정된 결과를 보여주고 있다.

그러나 수행시간에 있어서는 소프트웨어 및 알고리즘에 따라 다소의 차이를 보이고 있었으며, 따라서 본 연구에서는 다음과 같이 변수의 수를 달리하여 보다 다양한 상황에서 분석을 수행하여 보았다; (1) 20개 [이산형(4개), 연속형(16개)], (2) 50개 [이산형(9개), 연속형(41개)], (3) 100개 [이산형(17개), 연속형(83개)], (4) 153개 [이산형(27개), 연속형(126개)]. 여기서는 먼저 E-Miner를 이용하여 로지스틱 회귀분석을 실시한 후 $p-$값이 작은 순서대로 변수를 선택하였으며, 또한 각각의 경우에 대해 10,000, 20,000, 50,000개의 개체를 랜덤추출하여 분석을 수행하여 보았다.

표 4.2는 각 경우에 대해 수행시간을 정리한 것인데, 이를 살펴보면 Clementine의 로지

스틱 회귀분석과 I-Miner의 신경망분석에서 변수의 수가 100개 이상인 경우 개체의 수가 증가함에 따라 수행시간이 다소 크게 증가하는 것을 볼 수 있다(이 경우에 대해서는 2~3회씩 반복수행하여 수행시간을 재검토하였다). 또한 다른 알고리즘에 비해 신경망분석에서 변수와 개체의 수가 증가함에 따라 전반적으로 수행시간이 증가하고 있는 것을 볼 수 있다. 한편 의사결정나무분석의 경우 모든 소프트웨어에서 다른 알고리즘에 비해 수행시간이 비교적 적게 소요됨을 볼 수 있으며, 변수의 수가 20~50개인 경우에는 소프트웨어 및 알고리즘에 따라 수행시간에 큰 차이를 보이지 않고 있다.

## 5. 결론

본 연구에서는 널리 사용되고 있는 세 가지 데이터마이닝 소프트웨어들의 구조와 노드의 기능을 여러 측면에서 비교하였고, 실제 사례를 통해 이들 소프트웨어에 포함되어 있는 모델링 알고리즘들의 성능을 비교하였다. 소프트웨어의 구조에 있어서는 I-Miner가 하나의 오브젝트로 원하는 데이터마이닝 작업을 수행 할 수 있다는 점에서 가장 편리함을 가지고 있으며, Clementine과 E-Miner는 시각적으로 작업의 흐름도를 쉽게 파악할 수 있는 장점을 가지고 있다고 할 수 있다. 또한 노드들의 기능에 있어서는 세 가지 소프트웨어 모두 각각의 기능을 가진 노드들의 모임으로 이루어져 있으며 전반적으로 데이터마이닝 수행에 있어 필요한 기능들을 폭넓게 포함하고 있다는 것을 알 수 있었다. 실제 사례를 통한 비교에서는 정확도의 측면에서는 소프트웨어들 간에 큰 차이가 없었으며 모두 안정적인 결과를 보여주고 있었으나, 수행시간에 있어서는 소프트웨어 및 알고리즘에 따라 다소의 차이를 보이고 있었다.

본 연구에서 비교대상으로 하고 있는 데이터마이닝 소프트웨어들은 가장 널리 사용되고 있는 것들로, 앞에서 살펴본 바와 같이 매우 폭넓은 기능들을 가지고 있다. 그러나 이들을 사용하기 위해서는 적지 않은 하드웨어 및 소프트웨어적 지원이 필요하고 구입비용도 다소 비싼 편이어서, 개인이나 작은 규모의 프로젝트를 수행하고자 하는 사용자들이 쉽게 사용하기에는 어려운 점도 있다. 현재 상용화되어 있는 소프트웨어들 중에는 이들에 비해 기능면에서는 다소 부족하지만 한 두 가지의 알고리즘만을 이용하여 간단한 분석을 수행하거나 특정 데이터베이스와의 연동을 고려해야 하는 사용자에게는 보다 효율적인 것들도 있다. 예를 들어 Oracle의 Darwin과 같은 시스템은 Oracle 계열의 데이터베이스를 사용하고 있는 업체들이 선호할 수도 있다. 또한 각 소프트웨어들의 기능이 매우 빠른 속도로 개선되고 있기 때문에 해당 소프트웨어에 대한 기술적 지원의 용이성 등도 고려해야 할 필요가 있을 것으로 생각된다.

본 연구는 서론에서도 언급한 바와 같이 데이터마이닝 소프트웨어에 대한 비교를 통해 각 소프트웨어의 장단점을 파악함으로서 소프트웨어 사용자들에게 조금이나마 유용한 정보를 제공하는데 목적을 두었다. 본 연구의 결과가 데이터마이닝 소프트웨어의 사용자들이 분석의 목적 및 자신의 환경에 맞는 소프트웨어를 선택하고 실제 사용함에 있어 조금이나마 도움이 되기를 바란다.

## 참고문헌

[1] 강현철, 한상태, 최종후, 김은석, 김미경 (2001). <데이터마이닝 -방법론 및 활용->, 자유아카데미, 서울.

[2] 최종후, 한상태, 강현철, 김은석, 김미경, 이성건 (2001). <데이터마이닝 -기능과 사용법->, 자유아카데미, 서울.

[3] Abbott, D.W., Matkovsky, I.P., and Elder, J.F. (1998). An Evaluation of High-end Data Mining Tools for Fraud Detection, *1998 IEEE International Conference on Systems, Man, and Cybernetics*, San Diego, CA, October 12-14, 1998.

[4] Choi, Y.S., Kim, J.G., and Lee, J.H. (2001). A Comparison of Capabilities of Data Mining Tools, *The Korean Communications in Statistics, Vol.* 8, 531-541.

[5] Choi, Y.S. and Lee, J.H. (2001). Comparisons of Clustering, Detection and Neural Network in E-Miner, Clementine and I-Miner, <한국통계학회 춘계 학술발표회논문집>, 113-118.

[6] IBM Institute (1999). *IBM DB2 Intelligent Miner for Data Using the Intelligent Miner for Data, Version 6 Release 1*, IBM Inc.

[7] King, M.A. and Elder, J.F. (1998). Evaluation of Fourteen Desktop Data Mining Tools, *1998 IEEE International Conference on Systems, Man, and Cybernetics*, San Diego, CA, October 12-14, 1998.

[8] SAS Institute (1999). *Enterprise Miner Reference*, SAS Institute Inc.

[9] SPSS Institute (2001). *Clementine 6.0 User's Guide*, SPSS Inc.

[ 2001년 10월 접수, 2002년 4월 채택 ]

# A Comparison on the Efficiency of Data Mining Softwares

Sang-Tae Han[1] Hyuncheol Kang[2] Seong-Keon Lee[3] Duck-Ki Lee[4]

ABSTRACT

Data is being generated at an ever increasing rate in recent years, mainly due to technological advances in system architecture, processor speed, and storage structures. In this respect, data mining has attracted considerable attention and many commercial softwares for data mining have been developed. In this study, we compare the differences of functions and efficiency of application about several commercial data mining softwares which are widely used in real field.

*Keywords:* Accuracy, Data Mining, Modeling Algorithm.

1) Assistant Professor, Department of Informational Statistics, Hoseo University, Asan, 336-795, Korea. E-mail: sthan@office.hoseo.ac.kr
2) Senior Lecturer, Department of Informational Statistics, Hoseo University, Asan, 336-795, Korea. E-mail: hychkang@office.hoseo.ac.kr
3) Department of Statistics, Korea University, Seoul, 136-701, Korea.
4) Department of Mathematics, Hoseo University, Asan, 336-795, Korea.