Proceedings of the Autumn Conference in 2001
and the 30th Anniversary of the Korean Statistical Society

데이터마이닝 소프트웨어의 효율성에 관한 비교 연구

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 있으며, 이러한 관점에서 데이터마이닝이 큰 주목을 받고 있다. 따라서 다양한 데이터마이닝 기법들이 연구되고 있으며 데이터마이닝을 보다 손쉽게 수행할 수 있는 여러 상용화된 소프트웨어들이 개발되어 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.

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

2.1 Clementine의 운영환경과 특징

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 가능하

1) 호서대학교 자연과학부 수학전공 교수, (336-795) 충남 아산시 배방면 세출리 산 29-1
2) 호서대학교 자연과학부 수학전공 교수, (336-795) 충남 아산시 배방면 세출리 산 29-1
3) 고려대학교 대학원 통계학과 박사수료, (136-701) 서울시 성북구 안암동 5가 1번지
4) 호서대학교 대학원 수학과 석사과정, (336-795) 충남 아산시 배방면 세출리 산 29-1

며 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.

2.2 E-Miner의 운영환경과 특징

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.

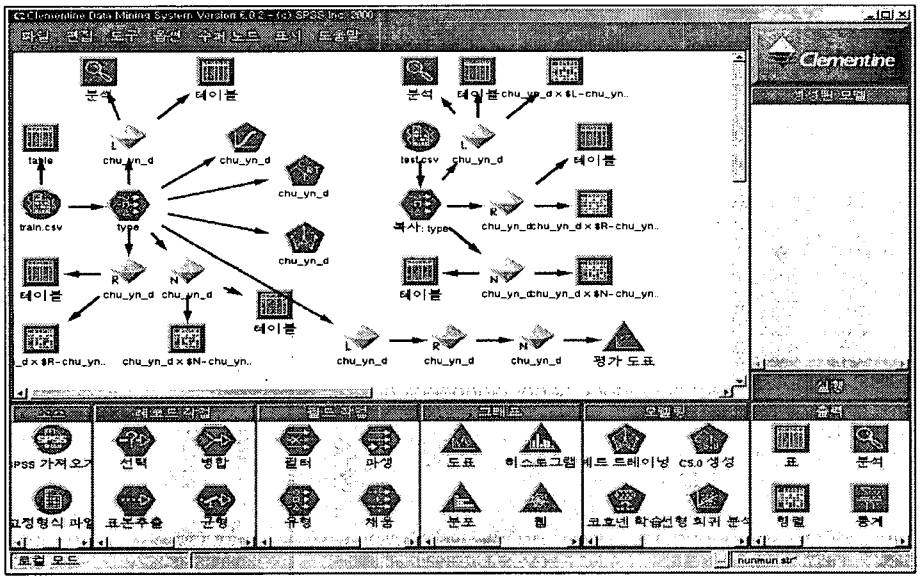
2.3 I-Miner의 운영환경과 특징

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.

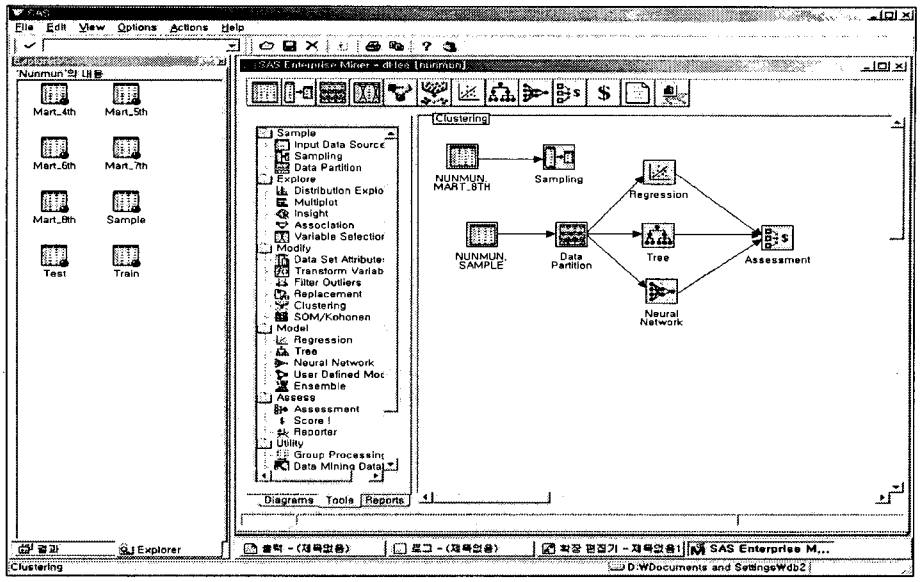
3. 구조 및 노드의 기능에 대한 비교

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.

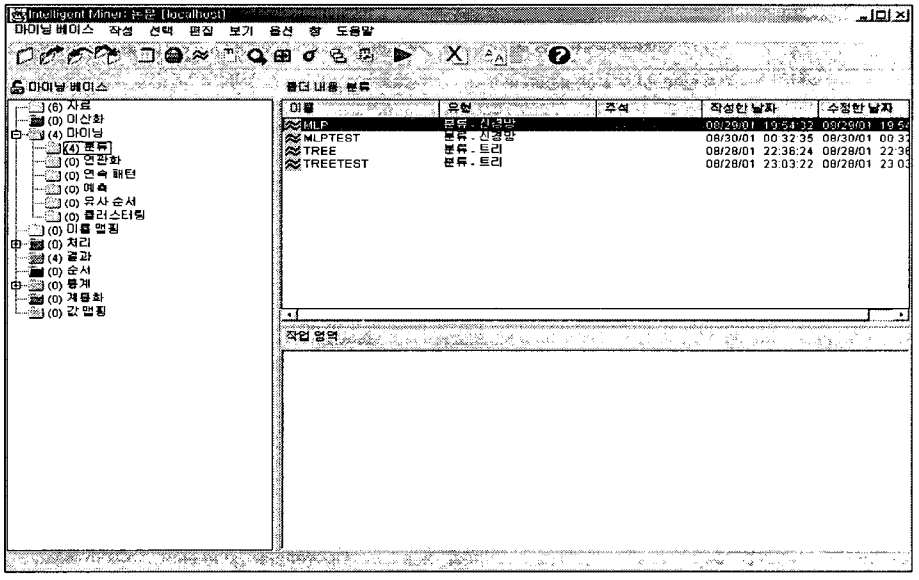
<그림 3-1> Clementine의 메인화면



<그림 3-2> E-Miner의 메인화면



<그림 3-3> I-Miner의 메인화면



특히 Clementine과 E-Miner는 분석흐름도를 한눈에 파악할 수 있는 장점이 있으며, I-Miner는 여러 오브젝트 유형을 하나의 오브젝트로 단일화하여 데이터마이닝 작업을 수행 할 수 있는 편리함을 가지고 있다. 또한 3가지 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. 따라서 3가지 소프트웨어들이 가지고 있는 기능상의 차이점은 크지 않은 것으로 보인다. 다음 <표 3-1>은 이들 소프트웨어들의 기능들을 비교·요약한 것이다.

<표 3-1> 데이터마이닝 소프트웨어들의 기능비교

노 드		Clementine	E-Miner	I-Miner
데이터 추출 및 관리	Data Partition	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
	History	O	O	O
데이터 탐색	Distribution	O	O	×
	Histogram	O	O	×
	3-D Plot	SPSS	O	×
모델링 기법	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	Sample Table
	Report	O	O	×
	Data Export	O	SAS	DB2

4. 모델링 알고리즘에 대한 비교

4.1 연관성분석

다음 <표 4-1>에서 볼 수 있는 바와 같이 일반적 연관성분석은 3가지 소프트웨어가 모두 포함하고 있으나, Clementine이 시차 연관성분석을 포함하고 있지 않은 것을 알 수 있다.

<표 4-1> 연관성분석 알고리즘

알고리즘	소프트웨어		
	Clementine	E-Miner	I-Miner
일반적 연관성분석	O	O	O
시차 연관성분석	×	O	O

4.2 군집분석

다음 <표 4-2>에서 볼 수 있는 바와 같이 군집분석의 경우에는 Clementine이 다른 소프트웨어보다 다양한 알고리즘을 제공하고 있는 것을 알 수 있으며, I-Miner는 군집분석을 위해 가장 일반적으로 사용되고 있는 k-Means 알고리즘을 포함하고 있지 않다. 이 표에서 TwoStep 군집분석과 데모그래픽 군집분석은 각각 SPSS와 IBM에서 개발한 군집분석 알고리즘이다.

<표 4-2> 군집분석 알고리즘

알고리즘	소프트웨어		
	Clementine	E-Miner	I-Miner
k-Means	O	O	×
Kohonen Map	O	O	O
TwoStep	O	×	×
Demographic	×	×	O

4.3 의사결정나무분석

다음 <표 4-3>에서 볼 수 있는 바와 같이 Clementine은 C5.0과 CART 알고리즘을 포함하고 있으며, E-Miner는 CHAID와 CART 그리고 C4.5 알고리즘을 복합적으로 포함하고, I-Miner는 SPRINT 알고리즘을 포함하고 있다는 것을 알 수 있다. 따라서 의사결정나무분석의 경우 E-Miner가 가장 많은 알고리즘을 포함하고 있는 것을 알 수 있다.

<표 4-3> 의사결정나무분석 알고리즘

알고리즘	소프트웨어		
	Clementine	E-Miner	I-Miner
CHAID	×	○	×
CART	○	○	×
C4.5 or C5.0	○	○	×
SPRINT	×	×	○

4.4 신경망분석

다음 <표 4-4>에서 볼 수 있는 바와 같이 3가지 소프트웨어 모두 MLP 알고리즘과 RBF 알고리즘을 포함하고 있다는 것을 알 수 있다.

<표 4-4> 신경망분석 알고리즘

알고리즘	소프트웨어		
	Clementine	E-Miner	I-Miner
MLP	○	○	○
RBF	○	○	○

4.5 회귀분석

다음 <표 4-5>는 3가지 소프트웨어에 대한 회귀분석 알고리즘의 포함여부를 나타낸 것으로 이를 살펴보면, Clementine과 E-Miner는 선형 회귀분석과 로지스틱 회귀분석을 모두 포함하고 있으나 I-Miner는 로지스틱 회귀분석을 포함하고 있지 않음을 알 수 있다.

<표 4-5> 회귀분석 알고리즘

알고리즘	소프트웨어		
	Clementine	E-Miner	I-Miner
Linear Regression	○	○	○
Logistic Regression	○	○	×

5. 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 세금우대상품에 추가가입 했는지의 여부이며, 입력변수(input variable)로는 고객의 인구사회적 속성과 금융거래 속성 등 153개의 변수가 사용되었다.

본 사례분석에서는 로지스틱 회귀분석, 의사결정나무분석, 신경망분석 등 3가지 모델링 기법을 각각 수행하여 그 결과를 비교하였다. 이 때 3가지 소프트웨어 모두 의사결정나무분석 알고리즘으로 CART 알고리즘을 사용하였고, 신경망분석은 모두 MLP 신경망을 이용하여 분석하였다.

수행결과를 비교하기 위하여 데이터의 60%인 30,000개를 분석용으로 사용하였고 나머지 40%인 20,000개를 검증용으로 사용하였으며, 다음 <표 5-1>에는 정분류율(accurate rate), 민감도(sensitivity), 특이도(specificity) 등 모형정확도의 3가지 측면과 수행시간을 제시한 것이다(정분류율, 민감도, 특이도 등에 대해서는 강현철 · 한상태 외(2001) 및 최종후 · 한상태 외(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 모델별로 큰 차이를 보이고 있는데, 로지스틱 회귀분석과 의사결정나무분석에 있어서는 E-Miner가 가장 적은 수행시간이 소요되었고, 신경망분석의 경우에는 Clementine이 가장 적은 수행시간이 소요되었다. 특히 Clementine의 로지스틱 회귀분석과 E-Miner의 신경망분석의 경우 매우 많은 수행시간이 소요된 것을 알 수 있다. 참고로 본 사례분석을 수행한 컴퓨팅 환경은 CPU는 펜티엄 800 듀얼, RAM은 516M, 그리고 OS는 Windows 2000이 사용되었다.

<표 5-1> 사례분석의 결과비교

			Clementine	E-Miner	I-Miner
분석용 데이터	로지스틱 회귀분석	정분류율	73.87%	73.89%	없음
		민감도	45.32%	44.83%	없음
		특이도	87.64%	87.91%	없음
		수행시간	6시간 16분	25분	없음
	의사결정나무 분석	정분류율	78.32%	79.22%	85.02%
		민감도	51.27%	53.94%	72.63%
		특이도	91.37%	91.41%	91.00%
		수행시간	6분 48초	1분	20분 48초
	신경망분석	정분류율	75.38%	72.75%	74.33%
		민감도	52.15%	52.94%	67.76%
		특이도	86.59%	82.23%	77.49%
		수행시간	20분 49초	3시간 10분	1시간 19분 6초
검증용 데이터	로지스틱 회귀분석	정분류율	73.95%	73.35%	없음
		민감도	63.52%	43.84%	없음
		특이도	87.40%	88.11%	없음
	의사결정나무 분석	정분류율	77.71%	79.41%	78.89%
		민감도	63.52%	54.69%	63.88%
		특이도	83.79%	91.31%	86.11%
	신경망분석	정분류율	73.80%	73.77%	75.86%
		민감도	50.47%	50.42%	52.72%
		특이도	85.03%	85.00%	87.00%

참고문헌

[1] 강현철 · 한상태 · 최종후 · 김은석 · 김미경 (2001). 「SAS Enterprise Miner를 이용한 데이터마이닝 -방법론 및 활용-」, 서울 : 자유아카데미.

[2] 최종후 · 한상태 · 강현철 · 김은석 · 김미경 · 이성건 (2001). 「SAS Enterprise Miner를 이용한 데이터마이닝 -기능과사용법-」, 서울 : 자유아카데미.

[3] IBM Institute (1999) *IBM DB2 Intelligent Miner for Data Using the Intelligent Miner for Data, Version6 Release1*, IBM Inc

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

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