S-PLUS와 StatServer를 이용한 Data Mining 도구 개발

정인석* · 이재준**

Development of Data Mining Tool Using S-PLUS and StatServer

In-seok Jeong* · Jae-June Lee**

요 약

통계 software에는 data mining에 필요한 다양한 모형과 함수들이 제공되고 있어 이를 이용한 data mining 도구가 소개되고 있다. 본 논문에서는 data mining을 수행하는데 효과적인 환경을 제공하는 S-Plus로 data mining 기법들을 구현하거나 재 구성하였으며, StatServer를 이용하여 대용량의 data base를 직접 관리할 수 있게 하고, S-PLUS의 분석기능을 Internet을 통하여 사용할 수 있게 하여 원거리에서 data mining작업을 수행될 수 있도록 구성하였다. 또한 분석자는 찾아낸 모형을 복잡한 프로그래밍 작업 없이 새로운 웹 페이지를 만들 수 있으며, 이를 통해 운영계의 사용자가 최적 모형이 제시하는 결과를 실제 업무에 즉시 이용할 수 있도록 하였다.

Keywords : data mining, S-PLUS, StatServer, logistic regression, tree.

* 인하대학교 통계학과 석사과정
** 인하대학교 통계학과 부교수

1. 서 론

정보 관련 기술의 급격한 발달로 인해 대용량의 data base들이 구축되어 왔으며, 이렇게 구축된 대용량의 data base로부터 유용한 정보를 얻어내어 활용하려는 연구가 활발히 이루어져 왔다. data mining 또는 Knowledge Discovery in Data Base(KDD) 는 넓은 의미에서 대용량 data에 내재된 data간의 관계, 패턴, 규칙 등을 찾아내고 모형화 하여 유용한 정보로 변환하는 일련의 과정(process)을 의미한다. 통계학은 비록 data miner에게 모든 해답을 주지는 못하지만 유용하고 실제적인 프래임워크를 제공하며(Glymour et al, 1997), 통계학적 입장에서 data mining은 거대한 크기의 복잡한 자료 집합에 대해 컴퓨터로 수행되는 자동화된 탐색적 자료 분석(Exploratory Data Analysis)이라고 정의된 바 있다(Friedman, 1997).

data mining의 효용성은 data mining전문가의 지식이 가장 중요하지만, 그에 못지않게 보다 많은 모형을 지원하고, 모형의 적합결과를 비교하여 최적 모형을 선택하며, 선택된 모형을 업무에 쉽게 적용할 수 있게 하는 data mining도구의 역할도 중요하다고 할 수 있다. 통계 software에는 data mining에 흔히 사용되는 탐색적 자료분석(EDA; Exploratory Data Analysis)에 관련된 통계량 및 도표, 안정된 알고리즘으로 작성된 판별분석(discriminant analysis)이나 군집분석(cluster analysis)과 같은 다변량 분석(multivariate analysis) 기법들이 제공되고 있고, 또한 data visualization에 필수적인 graphic 기능이 제공되고 있다. 따라서 최근에 이러한 통계 software의 장점을 활용한 data mining 도구가 소개되고 있는데, SAS의 Enterprise Miner와 SPSS의 관련 module 등을 들 수 있다.

지금까지 소개된 통계 software를 이용한 data mining도구는 특정 목적을 위한 도구라기 보다는 다양한 분야의 문제에 적용되도록 개발된 일반화된 도구들로서, 급속히 발전하고 있는 data mining분야의 최첨단 알고리즘을 첨가시키는 것이 쉽지 않고, C나 Fortran등으로 code화 된 프로그램을 직접 접목하여 활용하는 것이 쉽지 않은 단점이 있다.

이러한 배경에서, 본 논문에서는 data mining을 수행하기에 편리한 여러 가지 장점을 가지는 S-PLUS란 통계/수학 software를 이용하여 data mining에 관련된 일부 기법들을 S-PLUS로 구현하거나, 이미 S-PLUS에 내장되어 있는 함수들을 활용하여 data mining을 수행 하는데 편리하도록 재구성하였다. 또한, StatServer를 이용하여 대용량의 data base를 직접 관리할 수 있게 하고, S-PLUS의 분석기능을 internet을 통하여 사용할 수 있게 하여 원거리에서 data mining작업을 수행할 수 있도록 구성하였다. 이에 더하여 홈페이지를 재구성하여 data mining 결과로 얻어진 모형을 internet을 통해 실제 업무에 즉시 활용할 수 있게 하였다. 본 논문에서 개발한 data mining도구는 Miner S라고 부르기로 한다.

2. 통계 software를 이용한 data mining 도구

data mining(DM)은 최근에 소개된 분석방법으로서, 대용량 data에 내재하는 data 사이의 관계, 패턴, 규칙을 찾아내고, 모형화하여 유용한 정보로 변환하는 일련의 과정(process)을 나타낸다. DM은 데이터베이스 운용(data base management), 인공지능(artificial intelligence), machine learning,

pattern recognition, data visualization등의 공통된 학문영역에 자리잡고 있는 새로운 분야라 할 수 있다.(Adriaans and Zantinge, 1997)

2.1. 통계software의 장점

DM의 관점에서 DM 전문가의 지식이 가장 중요하겠지만, 그에 못지않게 DM도구의 역할도 중요하다고 할 수 있다. DM 도구는 보다 많은 모형을 지원하며, 지원되는 대상의 모형을 적합(fitting)하고 그 결과를 비교하여 최선의 모형을 선택(model selection)하며, 선택된 모형이 업무에 쉽게 적용될 수 있게 하는 기능이 요구된다. 통계 software는 DM에서 흔히 사용되는 다양한 통계 기법들과 안정된 알고리즘으로 작성된 최신의 DM분석 모형들을 제공하기 때문에, 통계 software를 이용한 DM도구는 다음과 같은 장점이 있다고 할 수 있다.

첫째. DM에서 흔히 사용되는 탐색적 자료분석(EDA; Exploratory Data Analysis)에 관련된 통계량(평균, 표준편차, 상관계수, 중위수, 백분위수 등) 과 다양한 도표(table, graph) 등이 제공되고 있다.

둘째. 판별분석(discriminant analysis)이나 군집분석(cluster analysis)등 DM에 사용되는 다양한 다변량 분석 기법들이 내재되어 있다.

셋째. DM에 필수적인 data visualization을 용이하게 하는 graphic기능이 제공되고 있다. 특히 자료의 경향이나 특성을 쉽게 파악할 수 있도록, box plot, parewise plot 등 통계학에서 제시하는 여러가지 그래프들이 제공되고 있다.

넷째. CART(Classification And Regression Tree), logistic regression등 통계모형과 neural network 등 일부의 DM기법들이 제공되고 있다.

이에 더하여 여러가지 첨단 DM기법들이 개발되어 소개되고 있는데, 이러한 새로운 기법들을 지원하기 위해 프로그램 개발에 많은 시간과 인력이 필요하게 된다. 반면에 통계 software를 이용한 DM 도구는 직접 개발된 프로그램에 비해 안정적이고 신뢰할 수 있으며, 새로운 기법을 통계 software에서 지원하기만 한다면 DM에 쉽게 추가하여 적용할 수 있다.

2.2. Data mining의 일반적 절차.

많은 DM 도구들은 서로 다른 방법론을 제시하고 있지만, 일반적으로 자료준비, 자료 변형과 축소, 모형화, 평가의 단계로 DM 작업이 수행된다.

① 자료 준비 (data preparation)

이 단계는 DM에 사용할 자료를 여러 다른 data base나 출처로부터 수집하는 단계이다. 전체 data를 사용하기에 용량이 너무 크다면, 분석시간과 비용이 필요이상으로 소비되므로, 일부 자료를 표본으로 추출(sampling)하여 사용하기도 한다. 또한 추출된 자료를 분할하여, 일부는 모형을 적합 하는데, 일부는 모형을 선택 하는데, 그 나머지는 적합된 여러 가지 모형을 평가하는 자료로 사용된다.

② 자료의 변환과 축소(data reduction)

이 단계는 준비된 자료를 탐색적 분석을 통해 대략적인 형태를 파악하고, 이상치(outliers)나 결측치(missing data), 혹은 오류가 있는 데이터를 정제하는 과정이다. 또한, 모형화의 조건에 부적합한 자료는 여러 가지 변환(transformation)방법을 통해 모형에 적합한 자료형태로 변환되기도 한다.

③ 모형화 (modeling)

최종적으로 준비된 자료를 이용해서 여러 가지 모형을 적합하는 단계로서, 이때 사용되는 기법은 case-based learning (k-nearest neighbor), decision trees, association rules, neural networks, genetic algorithms등이 있다.

④ 모형 평가와 비교(model assessment and solution analysis)

이 단계에서는 자료 준비단계에서 분할된 test 자료나 새로운 입력자료로 분류, 예측, 정확도, 비용 등을 고려한 통계치를 계산하여 모형을 비교, 평가하여 최적 모형을 선택한다.

2.3. 통계 software를 이용한 data mining 도구 비교

통계 software를 이용한 DM 도구로는, SAS의 Enterprise Miner와 SPSS의 관련 module등을 들 수 있다. SAS 사의 Enterprise Miner는 DM 프로젝트를 계획, 관리 실행하기 위한 방법으로 SEMMA(Sample, Explore, Modify, Model, Assess)라는 개념을 사용하고 있다. Enterprise Miner는 SAS가 제공하는 통계기법에 더하여 neural network, decision tree 및 association기능을 제공하고 있으며, user defined model의 기능이 활용될 수 있다.

SPSS는 정형화되고 통합된 DM 도구로 구성되어 있지는 않지만, Neural Connection, Answer Tree, Diamond 등의 개별적인 제품을 통한 DM 환경을 제공하고 있다. SPSS의 DM 방법론은 The 5As(Assess, Access, Analyze, Act, Automate)라는 개념을 도입하고 있다. 이 software들의 DM방법론을 요약하면 <표 1>과 같다.(본 연구에서 개발한 Miner S의 방법론은 4절의 개발 내용부분에 설명되어 있다.)

SAS의 Enterprise Miner와 SPSS의 DM환경은 2.1절에서 제시된 통계 software의 장점을 활용한 DM도구라고 할 수 있지만, 많은 DM기법들이 새로이 개발되어 소개되는 현 시점에서 볼 때 다음과 같은 한계와 문제점이 있다고 할 수 있다.

첫째. 새로운 DM기법이 첨가될 수 있는 유연성(flexibility)이 부족하다. SAS Enterprise Miner의 경우 user defined model 기능이 있지만 사용하기 쉽지 않은 면이 있으며, SPSS의 경우 새로운 model의 추가기능은 매우 미약하여 projection pursuit regression, polychofomous regression 등 새로이 개발된 최첨단 DM기법의 활용이 어렵다.

둘째. DM 분석 목적에 따른 사용자에 적합한 수준의 DM도구를 제공하는 것이 어렵다. 기존의 통계 software를 이용한 DM도구는 일반화된

〈표 1〉 data mining 방법론 비교

일반적 단계	SEMMA(SAS)	5As(SPSS)	Miner S
요구사항 파악	Access and Transform, Sampling	Assess	Loading Data from DB
자료준비		Access	
자료 변환과 축소	Explore/Modify	Analyze	Exploring
모형화	Modeling		Modeling
평가와 비교	Assess		Assessing
보고	Result Presentation	Act	
System에 적용		Automate	Publish

system으로서 특정 문제 해결을 위한 기능 위주의 주문형(customizable) DM도구에 비해 불필요한 기능들까지 포함할 가능성이 있다.

셋째. C나 Fortran으로 code화된 프로그램을 통계 software를 이용한 DM도구에 접목시키는데 어려운 면이 있다. 이미 code가 공개되어 있는 DM기법을 사용하거나, 계산 속도의 향상을 위해 C나 Fortran을 직접 이용할 필요성이 있는데, SAS의 경우 다소 사용하기 불편한 점이 있고 SPSS는 기능이 미약한 편이다.

S-PLUS는 통계 software로서, 앞에서 지적한 DM에서 통계 software의 장점을 모두 만족할 뿐 아니라, 위에서 지적된 기존 통계software로 개발된 DM도구의 문제점에 대처할 수 있는 몇 가지 장점을 가지고 있다. 따라서 본 논문에서는 S-PLUS를 이용한 DM도구를 소개하고자 한다.

3. S-PLUS 와 StatServer를 이용한 data mining 도구

2.3절에서 소개된 통계 software를 이용한 DM 도구는 안정된 알고리즘을 바탕으로 특정 목적을 위한 도구라기 보다는 다양한 분야의 문제에 적용되도록 개발된 일반화된 도구라 할 수 있다. 반면에 S-PLUS는 급속히 발전하고 있는 DM분야의 최첨단 알고리즘이 다른 통계software에 비해 빨리 제공되는 편이며, 또한 새로운 알고리즘을 사용자가 쉽게 작성하여 첨가시킬 수 있을 뿐만 아니라, C나 Fortran등으로 code화 된 프로그램을 직접 접목하여 활용하는 것이 용이한 면이 있다. 또한 GIS와 같이 향후 DM의 응용 분야에 확장이 기대되는 문제에서 GIS도구와의 연동이 용이한 장점이 있다.

3.1. S-PLUS 의 특징

S-PLUS는 MathSoft사가 제시한 통계 분석 도구로서, 원래는 미국 AT&T사의 Bell Laboratory에서 자체 개발하여 S라는 언어로 Unix상에서 사용되던 것이 발전되어 지금의 형태를 가지고 있다. S-PLUS는 통계적인 자료분석이나 그래픽에 이용되는 언어이며, 동시에 상호 호환적인 프로그래밍 환경(interactive programming environment)이라 할 수 있다(Venables and Ripley, 1996). S-PLUS는 다른 통계분석 도구에 비해 크게 여섯 가지의 특징을 들 수 있다.(Mathsoft, 1998a,b)

첫째, 언어가 객체지향적(object oriented)이며 매우 유연하다. 따라서 비교적 짧은 시간 안에 새로운 알고리즘을 프로그램으로 구현할 수 있다.

둘째, 약 2000여 개에 달하는 통계/수학 분석 함수들이 내장되어 있고, 특히 최근에 개발된 통계 기법들도 상당부분 함수로 제공되고 있으며, 새로운 통계기법을 software에 반영하는 기간이 다른 software에 비해 빠른 편이다.

셋째, 그래프 작성시 아주 세밀한 부분까지 조절이 가능하며, Trellis Graphics와 같은 고급 그래픽 기능이 제공되어 다차원 자료의 경향도 쉽게 눈으로 파악할 수 있게 한다.

넷째, C/Fortran언어와의 연계 기능이 우수하여 이런 언어로 작성된 함수를 S-PLUS에서 직접 이용하는 것이 가능하며, 좀 더 빠른 연산을 수행할 수 있다.

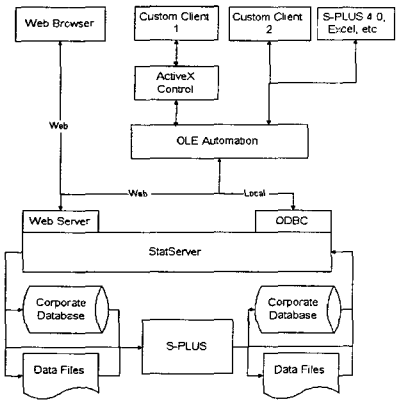
다섯째, OLE(Object Linking and Embedding)기능이 제공되고 있어, 이 기능을 통해 S-PLUS를 계산 엔진으로 새로운 도구를 개발하기 쉽다.

여섯째, ArcView와 같은 GIS 도구와의 연동이 가능하며, S+ SpatialStats와 같은 추가 모듈이 제공되어 공간 data의 분석도 가능하다. 따라서 공간 data에 대한 DM작업도 가능할 수 있다.

이러한 S-PLUS의 여러 가지 장점들은 DM도구 개발에 적합한 환경을 제공한다고 할 수 있다. 그럼에도 불구하고, S-PLUS를 이용한 전문적인 DM 도구가 아직까지 개발되어 있지 않은 실정인데, 본 논문에서는 제 4장에서 S-PLUS를 이용한 DM도구를 개발하여 제시하였다.

3.2. StatServer

StatServer역시 Mathsoft에서 개발한 프로그램으로서, 분석전문가 또는 일반사용자가 인터넷을 통해 S-PLUS의 함수와 그래프 등을 연계된 data와 함께 사용할 수 있게 하는 환경을 제공하는 server system이다. StatServer는 data base system이 구축되어 있는 server system과 연계하여 data base의 자료를 S-PLUS를 통해 분석하고, 그 결과를 넷스케이프나, 인터넷 익스플로러와 같은 web browser를 통해 보여준다. 또한, StatServer는 web뿐만 아니라 S-PLUS나 Excel을 client로 사용할 수도 있다. 이런 일련의 과정은 전문 사용자가 환경을 구축하는 작업과, 일반 사용자가 분석된 결과를 interactive하게 이용하는 두 가지로 분류할 수 있다.



〔그림 1〕 StatServer의 작동 구조

전문사용자는 일반사용자가 필요로 하는 정보를 미리 파악하고, S-PLUS코드를 작성하여 원하는 결과가 얻어질 수 있도록 프로그램하며, 다음으로 알맞은 형식에 맞추어 internet web에 interface를 작성한다. 이러한 전문사용자의 작업이 끝나면, 일반 사용자는 자신이 원하는 옵션이나 필요로 하는 결과의 내용을 구축된 interface를 통해 StatServer에 전달하며, StatServer는 그 작업을 받아서 S-PLUS로 수행되게 하고, S-PLUS가 작업한 결과를 다시 web을 통해 사용자에게 제공하게 된다. [그림 1]은 StatServer의 작동구조를 도식화한 것이다.

3.3. Miner S의 특징 및 구조

Miner S는 DM을 수행하기에 편리한 여러 가지 장점을 가지는 S-PLUS를 이용하여 DM에 관련된 일부 기법들을 S-PLUS언어로 구현하거나, 이미 S-PLUS에 내장되어 있는 함수들을 활용하여 DM을 수행 하는데 편리하도록 재구성하였다. 또한, 대용량의 data base를 직접 관리할 수 있고, S-PLUS의 분석기능을 internet을 통하여 사용할 수 있게 하는 StatServer의 기능을 이용하여 원거리에서 DM작업을 수행할 수 있도록 구성하였다. 사용자 interface를 보기 좋고 사용하기 편리하도록 internet web 기술인 ASC(Active Server Page)와 Java Script를 사용하였다.

이런 이유로 Miner S는 S-PLUS의 많은 통계모형을 쉽게 추가할 수 있으며, C나 Fortran언어로 작성된 프로그램으로의 확장성 등 3.1절에서 설명한 S-PLUS의 장점을 충분히 활용한 DM도구라 할 수 있다. 또한 Miner S는 어떤 server system에도 customizable하게 적용이 가능하고 원격지에서 internet을 통해 DM 작업을 수행할 수 있는 StatServer의 장점을 함께 지닌 DM도구라

고 할 수 있다. 이에 더하여 Miner S에서는 홈 페이지를 재구성하여 DM 결과로 얻어진 모형을 internet을 통해 실제 업무에 즉시 활용할 수 있게 하였다. 현재 Miner S는 초기 단계의 DM도구로서 많은 모형을 지원하고 있지는 않으나 기존의 통계적 모형 뿐만 아니라 앞으로 개발될 새로운 모형들을 지원할 수 있도록 쉽게 확장할 수 있는 장점이 있다.

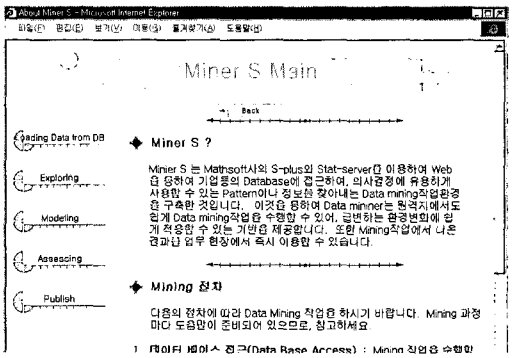
4. 개발 내용 및 사용 예

Miner S의 기능과 사용 방법을 설명하기 위하여 분석 대상 data로 German Credit Data[1]를 이용하였다. 이 data는 각 개인의 배경 요인들에 해당되는 21개의 독립(설명)변수와 신용의 등급이 좋고 나쁨을 평가하는 1개의 종속변수로 구성되어 있다. StatServer는 ODBC를 지원하는 어떤 형태의 data base와 연결이 가능하다. 본 예에서는 운영계의 data base로부터 data를 읽어 들이는 기능을 보이기 위해, ASC II 형태의 원 자료를 Microsoft 의 Access형식의 data base로 변환한 후, 이것을 ODBC를 이용하여 StatServer와 미리 연결해 사용하였다.

4.1. Main 페이지

Miner S는 DM의 일반적 단계를 그대로 메뉴 형식으로 반영하였다. 메인 페이지는 [그림 2]에서와 같이 Loading data base from DB, Exploring, Modeling, Assessing과 Publishing의 다섯 가지 단계로 구성되어 있다.

1 출처 : ftp://ftp.ics.uci.edu/pub/machine-learning-databases/statlog/german/

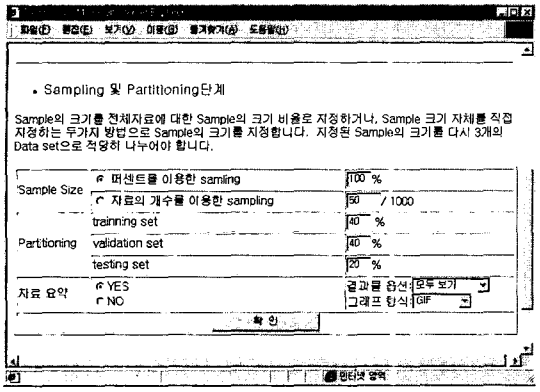


〔그림 2〕 Miner S 메인 페이지.

4.2. Loading Data Base 및 sampling과 partitioning

DM작업의 시작은 data base를 읽어 들여서 S-PLUS가 작업할 수 있도록 객체로 만들어야 하는데, 이 작업을 Loading Data from DB에서 수행한다. Loading Data from DB 화면에서는, StatServer와 연계 되어있는 data base들 중 하나를 선택하고, 선택된 data base안의 data base table을 선택한다. 인터넷을 통해 작업이 이루어지므로, data base의 보안 설정에 따라 사용자 계정과 암호를 물어올 수 있는데, 이때는 data base관리자 수준의 권한을 가진 계정이 필요하다. 필요한 정보를 입력하고 나면 나타나는 data base table 정보에서는 선택한 table안의 변수이름을 selection box안에 나타내어 modeling에서 종속변수와 독립변수로 사용하게 될 변수를 지정할 수 있다. "계속" 버튼을 누르면, 선택된 변수의 정보를 보여주고 바로 sampling과 partitioning 작업을 수행할 수 있는 화면이 나타난다.

[그림 3]의 sampling단계에서는 비율을 지정하는 옵션과, 직접 Case의 수를 지정하는 두 가지 sampling 옵션이 있다. 원하는 옵션의 라디오 박스를 클릭하고, 옵션 우측의 텍스트 박스에 적당

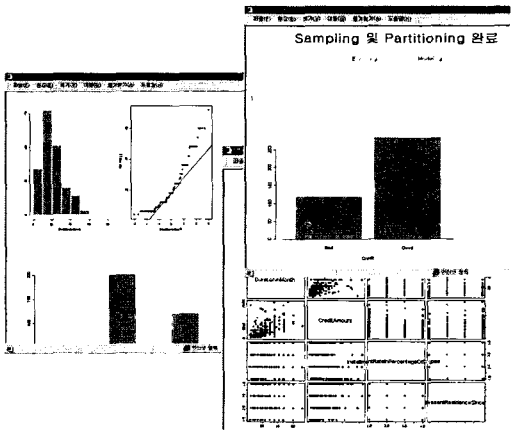


〔그림 3〕 sampling 및 partitioning 화면

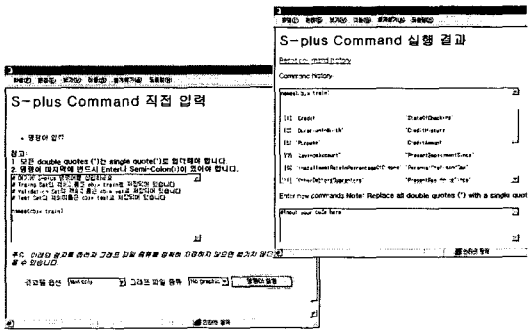
한 숫자를 입력한다. 현재 sampling 방법은 단순 임의 추출 방법만 지원하고 있다. partitioning옵션에서는 sampling된 자료를 다시 세 부분으로 나누는 과정이다. 모형을 적합 하는데 사용하는 training set과 over-fitting을 방지하기 위한 validation set, 그리고 적합된 모형들을 비교하는데 사용할 test set으로 나누게 되며, 각각의 비율의 합이 100이 되도록 지정하면 된다. 예에서는 자료의 개수가 1000개로 작기 때문에, 100%를 sampling하고, training, validation, test set을 각각 40%, 40%, 20%로 지정하였다. sampling 옵션이 지정된 후 다음 단계로 진행하면, [그림 4]와 같이 범주형 변수는 히스토그램이, 연속형 변수인 경우는 normal QQ plot, 히스토그램 그리고, pairwise plot이 제시된다. 그 아래에는 변수의 종류에 따라 연속형은 평균, 중위수, 4분위수 등 기초 통계량을, 범주형 변수는 각 범주의 돗수들을 화면에 표로 나타낸다.

4.3. Exploring

이 단계는 선택된 자료를 탐색하고 변환하는 단계로서, "Graph Menu" 에서는 training Set의 각 변수에 대하여 S-PLUS가 제공하는 그래프를 그릴 수 있으며, "S-PLUS Command"에서는 S-PLUS 명령어를 입력하여 변수를 변환하거나, 결측치 혹은 이상치 등을 처리하는 것이 가능하다. 이 단계는 상황에 따라 생략되어 modeling 작업으로 직접 진행하는 것도 가능하다.[그림 5]



〔그림 4〕 sampling과 partitioning의 결과

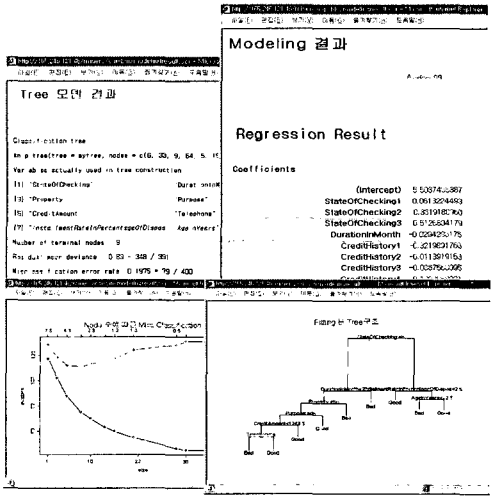


〔그림 5〕 S-PLUS 명령어 입력 및 그래프 메뉴

4.4. modeling

현 버전의 Miner S에서는 logistic regression 모델과 tree 모델의 두 가지를 지원하고 있다. modeling화면에서 원하는 모형에 "O" 마크를 선

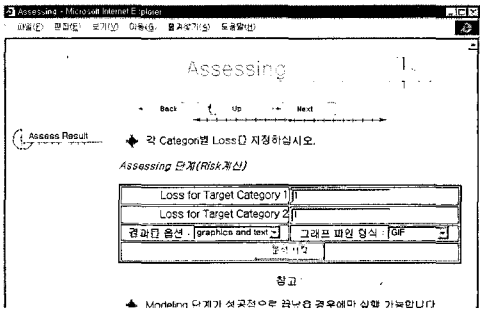
택하고 선택된 모형에 대한 추가 옵션을 지정하면, modeling의 수행결과는 [그림 6]과 같이 출력된다. logistic regression의 경우 적합된 모형의 coefficient들이 각 변수별로 출력되는데, 이때 범주형 변수인 경우 자동으로 dummy 변수가 생성되어 변수의 개수가 증가하게 된다. tree 모형의 경우 노드의 수가 증가함에 따라 misclassification의 수가 변해가는 추세를 training Set 과 validation Set에 대하여 하나의 그래프에 나타나도록 작성되었다. training Set의 misclassification의 수는 일반적으로 노드가 증가함에 따라 계속 줄어드는 경향이 있지만, validation Set의 misclassification 수는 처음에는 줄어들다가, 노드 수가 어느 정도 이상 늘어난 이후에는 증가하는 경향을 보인다. 따라서 validation Set의 misclassification 수가 감소하다가 증가하기 시작하는 때의 노드의 수를 최적으로 간주하고, tree model을 적합하게 된다. [그림 6]의 좌 하단의 그래프는 두 data Set의 misclassification의 추이를 나타내며, 우 하단의 그래프는 최종적으로 적합된 tree 모형의 결과를 그림으로 나타낸 것이다.



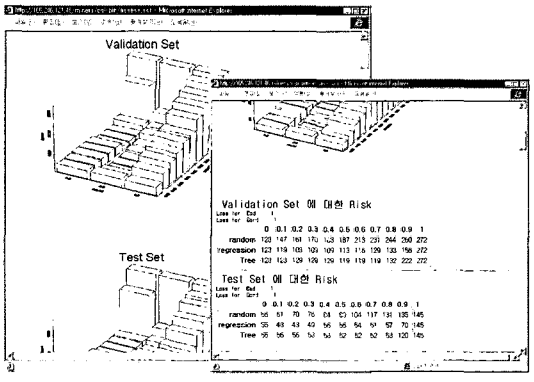
〔그림 6〕 modeling 결과

4.5. assessing

적합된 모형들 중 최적의 모형을 찾는 단계로서, Miner S에서는 모형의 misclassification 수와 함께, 종속변수의 각 범주에서 오판시 발생하는 손실(loss)을 지정할 수 있도록 작성 하였다(그림7 참조). 지정된 손실을 통해 각 범주의 misclassification 수를 손실과 곱한 합을 구하여 예상되는 위험(risk)값을 구하고, 이 위험값 들을 비교하여 가장 적은 값의 모형을 선택할 수 있도록 하였다.



〔그림 7〕 assessing 페이지



〔그림 8〕 assessing 결과

[그림 8] 상단의 그래프는 각 모델별로 계산된 risk를 3차원 막대그래프로 나타낸 것이다.

그래프의 좌측 하단의 y축은 모델을 나타내는데, 가장 좌측으로부터 logistic regression, tree, random 모델의 순서로 출력된다. 우측 하단의 x축은 일종의 threshold로서 기준 범주일 확률 값을 threshold값과 비교하여 그 범주를 판단하게 된다.

4.6. Publishing

이단계는 적합된 모형을 실제 업무에 이용할 수 있도록 새로운 홈페이지를 개설하는 메뉴이다. 이 부분은 StatServer의 analytic object client wizard의 기능을 그대로 이용할 수 있도록 연동시켰다. 분석자는 홈페이지를 꾸미는 등의 별도 작업 없이 새로운 페이지를 만들 수 있으며, 이를 통해 운영계의 사용자는 최적 모형이 제시하는 결과를 실제 업무에 이용할 수 있게 된다.

5. 결론 및 토의

Miner S는 S-PLUS와 StatServer를 이용하여 web상에서 분석작업이 수행 될 수 있도록 개발된 DM 도구로서, 본 논문에서는 German Credit data를 data base형태로 연결하여, sampling, partitioning, modeling, assessing등 DM의 모든 과정을 단계별로 수행한 결과를 제시하였다. web상에서 S-PLUS로 이런 도구를 개발 할 수 있다는 것은 web을 통해 DM을 수행할 수 있다는 장점 뿐만 아니라, S-PLUS를 사용하는 모든 system환경에서 다양한 환경의 프로그래밍 도구를 사용하여 data base와 연계한 의사결정 시스템이나 DM도구 등을 개발할 수 있다는 가능성을 제시하는데 의의가 있다고 할 수 있다.

본 논문에서 소개된 Miner S는 S-PLUS를 이용한 초기단계의 mining도구로서, 여러 가지 기능이 추가될 수 있다. 현재 제공되는 simple random sampling뿐만 아니라 stratified random sampling, cluster sampling등 복잡하고 다양한 sampling방법이 추가 지원되면, 실제 모형에 좀 더 근접한 모형을 구축하는데 도움이 될 것이다. 또한 neural network 모형 등 다른 DM 도구들이 지원하고 있는 다양한 모형들을 쉽게 추가할 수 있을 뿐만 아니라 S-PLUS 특유의 기법인 GAM(Generalized Additive Model)이나 Projection Pursuit등 보다 많은 통계 모형도 추가될 수 있을 것이다.

참 고 문 헌

[1] Adriaans, P. and D. Zantinge (1997), *Data Mining*, Addison Wesley Longman.

[2] Fayyad, U. M., G. Piatetsky-Shapiro and P. Smyth (1996), "From Data Mining to Knowledge Discovery: An Overview", *Advances in Knowledge Discovery and Data Mining*, AAAI Press, 1-32

[3] Friedman, J. H. (1997), "Data Mining And Statistics: What's the Connection?", Technical Report

[4] Glymour, C., D. Madigan, D. Pregibon and P. Smyth (1997), "Statistical Themes and Lessons for Data Mining", *Data Mining and Knowledge Discovery 1*, Kluwer Academic Publishers, 11-28

[5] Hand, D. J.(1998), "Data Mining: Statistics and More?", *The American Statistician*, Vol52, No.

2, 112-118

[6] MathSoft (1998a), *S-PLUS4 Guide to Statistics*, Data Analysis Products Division, MathSoft, Seattle, WA.

[7] MathSoft (1998b), *S-PLUS4 Programmers Guide*, Data Analysis Products Division, MathSoft, Seattle, WA.

[8] Uthurusamy, R. (1996), "From Data Mining to Knowledge Discovery: Current Challenges and Future Directions", *Advances in Knowledge Discovery and Data Mining*, AAAI Press, 560-569.

[9] Venables, W.N. and B.D. Ripley (1996), *Modern Applied Statistics with S-PLUS*, Springer