

웹 기반의 데이터 마이닝 솔루션 개발에 대하여

구 자 용*, 박 현 진**, 최 대 우***

The Development of Data Mining Solution based on Web

Ja-Yong Koo, Heon Jin Park, Daewoo Choi

요 약

최근 데이터 웨어하우징의 활발한 구축과 우수고객 확보를 위한 치열한 경쟁으로 데이터 마이닝은 많은 업체의 큰 관심을 끌고있다. 본 연구는 풍부한 알고리즘과 과학적 그래프를 제공하여 사용자로 하여금 최상의 데이터 마이닝 효과를 거둘 수 있도록 StatServer를 핵심 엔진으로 사용한 인터넷 기반의 데이터 마이닝 솔루션 개발에 관한 것이다.

Key words: trec model, split-rule, one-sided purity

1. 서론

Mathsoft사의 StatServer는 통계 전문가 분석 툴인 S-PLUS의 server version으로 S-PLUS, Web, Excel, Visual Basic등 다양한 client 환경을 제공한다. 이미 알려진 바와 같이 S-PLUS는 Lucent Technology의 Bell Lab.에서 개발한 S-language를 기반으로 하여 2,000여 개의 통계 함수를 내장하고 있을 뿐 아니라 데이터 마이닝에서 유용하게 쓰일 수 있는 Trellis와 같은 첨단 그래픽 library와 projection pursuit regression, tree model, neural network 및 Polyclass등 다양한 알고리즘을 제공하고 있다. 그럼에도 불구하고 전문가 중심의 사용자 interface만을 제공하고 있어 학교나 연구기관 외의 산업현장에서는 거의 쓰여지지 않고있다. 그러나 StatServer에서 제공하는 다양한 client 환경을 이용하여 S-PLUS를 모르는 end-user도 쉽게 사용할 수 있는 데이터 마이닝 솔루션을 개발할 수 있었던 것이다. 본 연구의 결과물은 Statistical Design, Exploration, Layout, Process and Analysis의 첫 글자로 구성된 S-Delpa라고 명하고 있다.

본 논문의 구성을 살펴보면 다음과 같다. 제 2 절에서는 데이터 마이닝 분석 프로세스인 DELPA 과정에 대하여 논하겠다. 제 3절에서는 StatServer와

S-Delpa의 구조를 살펴보고, 제 4절에서는 S-Delpa에 구현된 각종 그래프 및 모델링 기법에 대하여 소개하겠다.

2. 데이터 마이닝 프로세스

데이터 마이닝의 분석과정은 분석자료의 설계(Design), 자료에 대한 탐색(Exploration), 결과의 정리(Layout), 필요 자료의 재추출 및 변형 등의 처리(Process), 그리고 모형 도출 및 최종 모형 선정을 위한 분석(Analysis)의 DELPA라는 5단계로 표현할 수 있다. 물론 분석 과정에 대한 다른 소개들도 있다. 그 예로 모형 구축위주의 과정을 표현한 SEMMA(Sampling, Exploration, Modifying, Modeling and Assessment)라는 방법론도 있으나 대용량 고객 정보로부터 분석을 위한 자료 추출이 중요시되는 데이터 마이닝에 있어서는 편협한 접근 방법이라고 할 수 있다.

DELPA 분석단계에 대하여 자세히 살펴보면 다음과 같다.

① Design

* 한림대학교 정보통계학과 교수

** 인하대학교 수학, 통계학부 부교수

*** 한국의국어대학교 정보통계학과 조교수

: 데이터 마이닝에 있어 가장 중요한 과정으로써 여러 형태로 존재하는 고객에 관한 정보 및 고객의 서비스 사용과 관계된 데이터베이스를 데이터 추출하여 분석이 가능한 형태로 재구성하게 되는데 이것을 데이터 마트(data mart)라고 부른다. 이 과정에서는 사용 가능한 정보가 어떠한 것이 있는지 전체 사용정보를 파악하고 주제에 맞는 데이터 마트의 형태를 디자인하는 것이다. 아울러 마트로부터 적절한 표본추출(sampling) 방법을 사용하여 분석자료를 마련하는 것도 이 과정에서 해야 할 작업이다.

② Exploration

: 분석을 위한 데이터 마트 구축이 완료되면 각 변수에 대해 분포, 데이터 마이닝 과제에서 설정한 주제와의 관계 및 변수간의 선형, 비선형 관계 등을 관찰하게 된다. 이 과정에서는 분산분석(ANOVA)나 cross-tabulation 등의 통계분석 방법이 사용되기도 하지만 각종 그래프를 통한 시각적 관찰이 유용하게 쓰인다. 특히 시각적 관찰을 위한 방법(data visualization tools)들이 다양하고 첨단화 알고리즘에 의해 구현되어 있어 최근 들어 자료분석에 있어 큰 비중을 차지하고 있다.

③ Layout

: 자료 탐색 과정에서 발견된 여러 가지 결과들을 조합하는 과정이다. 특히 각 변수에 대하여 누락된 경우가 있는지 살피고 그 원인을 파악하는 것은 매우 중요한 일이다. 누락된 경우가 지나치게 많은 변수는 일반적으로 분석에서 제외시키기도 하지만, 중요 변수가 그러한 경우 데이터 마트 디자인 과정부터 다시 시작하여야 한다.

마이닝 분석에 사용할 수 있는 고객 중심의 자료형태로 추출(extraction)하는 과정을 말한다. 즉, 각종 데이터베이스로부터 분석용 자료를

④ Process

: 자료의 탐색과 그 결과의 정리, 분석이 끝난 후, 데이터 마트 재구성이 필요한 경우 설계단계로 돌아가고 그렇지 않으면 기존의 변수를 변환하거나 변수들간의 결합을 통하여 분석에 새로운 변수를 생성한다. 아울러 분석에 사용될 수 없거나 상관도가 높은 변수들은 데이터 마트에서 제외하는 작업도 이 과정 중에 해야 할 일이다. 그리고 변환되거나 결합되어 새롭게 생성된 변수는 탐색과정을 통해 다시 관찰되어야 한다.

⑤ Analysis

: 모형도출에 필요한 자료들이 준비되면 적절한 모형들을 선택하여 모형 적합을 시도한다. 데이터 마이닝에서는 대부분 비모수적인 모형을 사용한다. 모수적 모형은 여러 가정 하에서 모형 적합이 이루어지기 때문에 앞으로 일어날 오류를 측정할 수 있지만 가정이 어긋나는 경우 모형의 정확도가 크게 떨어진다. 비모수적 모형은 이와 반대이다. 즉, 앞으로의 자료에 대한 예측력이나 분류에 대한 정확성을 모형 도출시 측정할 수 없다는 단점이 있다. 이와 같은 단점을 극복하기 위해 분석에 사용할 전체 자료 중 70%만 모형 적합에 사용하고 나머지는 구하여진 모형의 정확성을 측정하기 위한 확인 자료(validation data)로 사용한다. 여러 모형들에 대한 평가는 동일한 확인 자료에 의한 오류에 의해 이루어지고 그 중 가장 정확하고 안정적이며 목적에 맞는 모형이 선택되어 진다.

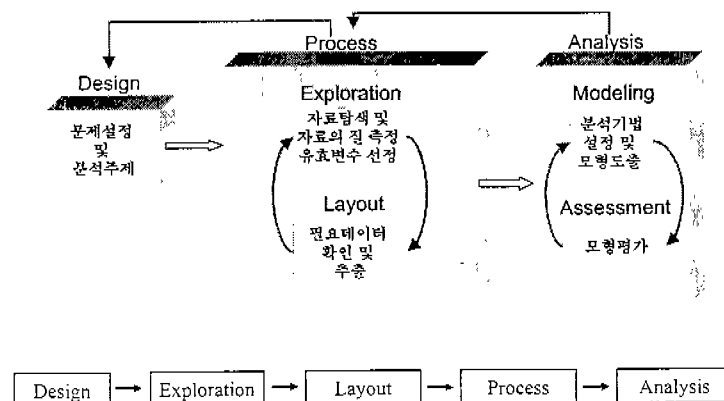


그림-1 DELPA procedure

3. S-Delpa의 구조

S-Delpa는 StatServer를 기본 엔진으로 사용하고 있다. 이 절에서는 우선 StatServer의 구조 및 그 특징을 살펴본 후 S-Delpa의 설계 및 운영방식에 대하여 소개하겠다.

3.1 StatServer

S-PLUS의 server 및 enterprise version인 StatServer는 Web, Excel 및 마이크로소프트사의 Visual Studio를 이용하여 제작된 application 등 다양한 client 및 개발 환경을 제공하고 있다. 현재 S-Delpa에서 사용한 StatServer의 version은 2.2로 다음과 같은 다양한 기능을 제공하고 있다.

- 웹, Excel, S-PLUS 등 다양한 client 환경을 제공한다.
- OLE Automation으로 마이크로소프트사의 각종 프로그램과의 연동이 가능하다.
- Sybase 기반의 데이터 베이스가 내장되어 StatServer 전반의 작업을 관리한다.
- Microsoft사의 SQL, IBM의 DB2, Informix, Sybase, Oracle 등 각종 database와의 ODBC가 가능하다.
- 그리고 분석 결과물을 S-PLUS는 물론 excel, SAS, SPSS, FoxPRO 등 다양한 형태로 출력할 수 있다.

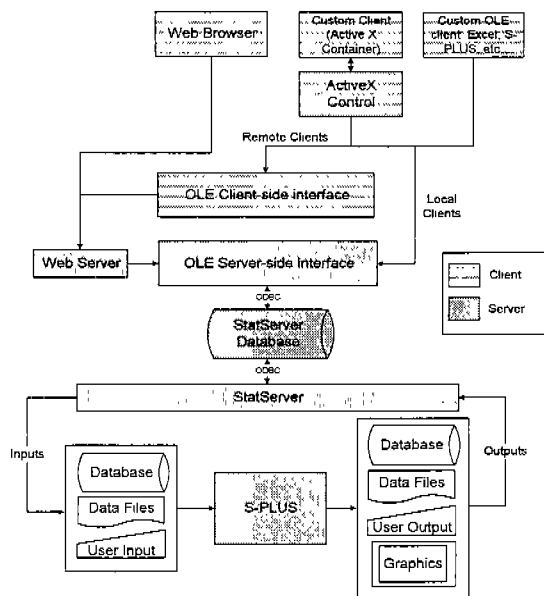


그림-2 StatServer의 구조

- StatServer의 실행 단위인 각 analytic을 스케줄링에 의해 원하는 시간에 작동할 수 있다.

이 외에 다양한 기능이 제공되고 있고 위에서 언급한 사항들은 특히 S-Delpa의 구현에서 활용된 대표적인 사항들이다.

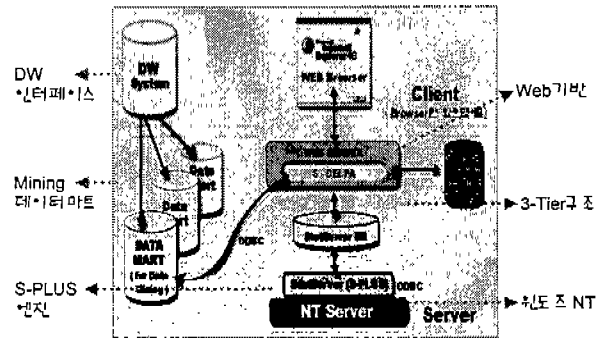


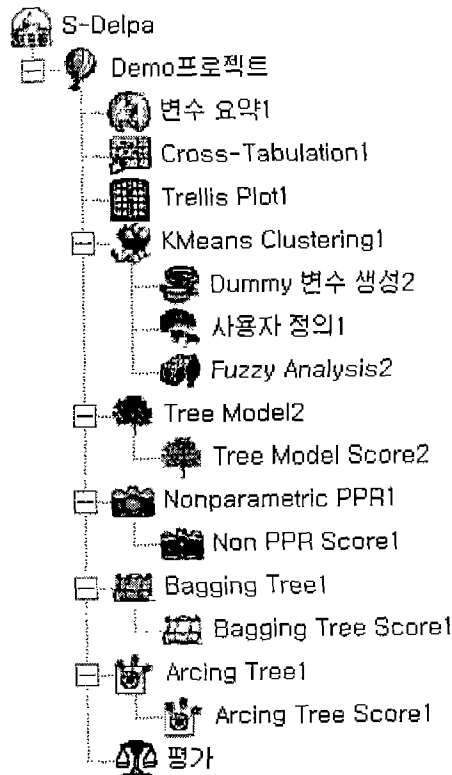
그림-3 S-Delpa의 구조

3.2 S-Delpa의 설계 및 운영방식

인트라넷 기반의 데이터 마이닝 솔루션인 S-Delpa는 3가지의 컴포넌트로 구성되어 있다. 각 컴포넌트의 기능을 살펴 보면 다음과 같다.

- Analytic 부분: 분석, 그래프, 자료변환 등을 실행하기 위한 S-plus의 명령어 이루어진 함수들의 집합.
- 작업관리 database 부분: 마이닝 프로젝트 및 계정 관리를 담당하는 데이터 베이스. 현재 마이크로소프트 SQL 7.0로 구축되어 있으나 어떠한 DB로도 구축이 가능하다.
- ActiveX controller 부분: 마이닝 과정을 연결하고 작업간의 argument들을 전달하는 부분. 아울러 각 ActiveX controller는 노드(node)를 구성하며 각 노드는 analytic과 1대1, 혹은 1대2 대응의 관계를 갖고 있다.

그림4의 각 그림은 노드이며 StatServer의 분석 단위인 analytic과 연동되어 있다. 각 노드간에는 일정한 규칙 하에 정보가 교환되어 진다. 이러한 노드들의 연결을 그림4과 같이 시각적으로 표현함으로써 데이터 마이닝 작업과정의 기록 및 업무와 약에도 도움이 된다.



.그림-4 ActiveX controller로 이루어진 노드를 이용한 데이터 마이닝 분석의 예

4. 구현된 데이터 마이닝 기법

S-Delpa에서는 데이터 마이닝에 필요한 모든 과정들을 노드로 표현하고 있다. 특히 S-Delpa가 제공하는 첨단 그래픽 기능이나 다양한 데이터 마이닝 알고리즘은 타 데이터 마이닝 솔루션과 차별화되는 우수한 기능이다. S-Delpa에 구현된 그래픽 기능과 알고리즘을 살펴보면 다음과 같다.

4.1 Trellis 그래픽스

Trellis 그래픽스란 Lucent Technology 의 Bell Lab.에서 개발한 S-PLUS 용 그래픽 라이브러리이다. 특징을 살펴보면 다음과 같다.

- 조건(conditioning)에 의한 분리된 그래프 출력
- 다양한 그래프들의 overlapping
- 사용자 정의 혹은 자동 scaling

Parallel-coordinate plot(그림-5 참조), density plot, dot plot 등이 이에 속하며 S-Delpa 에는 20 여가

지의 다양한 그래프 기능이 준비되어 있다. 이와 같은 첨단의 기능을 이용함으로써 수 많은 변수로부터 발생하는 다차원의 저주(curse-of-dimensionality)를 극복할 수 있다.

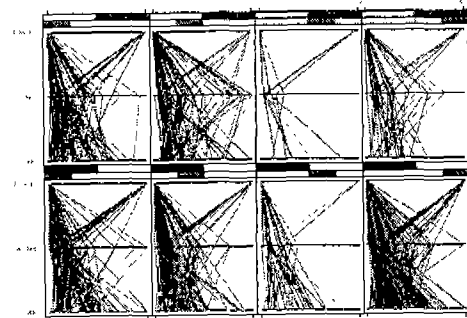


그림-5 Parallel Coordinate Plot

4.2 Combining Learners와 Tensor Product에 기초한 데이터 마이닝 알고리즘들

데이터 마이닝에서 활용되는 알고리즘은 과거의 문자, 음성 인식등에서 활용되었던 분류(classification) 알고리즘들이다. 알고리즘의 성능을 향상시키던 노력은 Breiman (1996)이 제안한 Bagging (Bootstrap Aggregatin)이라는 combining learner 방법이 제안되면서 큰 전기를 마련하였다. 현재 S-Delpa 에도 이와 같은 bagging 뿐 아니라 Breiman(1998)의 arcing 도 구현되어 있다.

한편 통계학 분야에서는 나무 모형(Tree model)과 신경망(neural network)모형의 장점을 결합한 Tensor product approach 분야가 연구되어 왔다. 대표적인 것으로는 Freidman(1991)의 MARS (Multivariate Adaptive Regression Spline)와 Kooperberg, Bose 및 Stone(1997)의 Polyclass 와 PolyMARS 이 있으며 S-Delpa 에 구현되어 있다.

Tensor product 계열의 모델링 방법의 장점은 다음과 같은 것들이 있다.

- 결과 해석이 가능하다.
- 변수 선택 기능이 우수하다.
- 이론상 모든 함수의 추정이 가능하다.
- Over-fitting의 염려가 없다.

그러나 위와 같은 장점에도 불구하고 알고리즘의 이해가 쉽지 않아 대부분의 데이터 마이닝 솔루션에 구현되어 있지 않고 있다.

이와 같이 S-Delpa 에 첨단의 알고리즘이 구현되어 있고 앞으로도 새로운 방법이 계속 갱신될 수 있는 이유는 S-PLUS 가 Fortran 과 C/C++등의 일

반 언어와의 interface 가 용이하고 수 많은 연구 결과가 S-PLUS 를 중심으로 구현되기 때문이다.

4.3 혼성모형 (Hybrid model)

혼성모형은 이질적인 모형들을 결합하여 각 모형이 갖고 있는 장점을 살리는 패러다임이다. 몇 년 전 국내 모 이동통신 회사의 고객 이탈예측 시스템에서 사용된 이 후 많은 관심의 대상이 되기도 한 이 모델링 기법을 S-Delpa에서는 Neural Net-after-CART 등 총 4가지의 방법을 구현하였다.

이 절에서는 국내 이동통신사의 사례를 통해 혼성모형의 출현 배경 및 구현방법, 그리고 앞으로의 연구 과제에 대하여 소개하겠다.

이동통신 해지자 예측분석 당시 분류예측 모형에 대한 정확도 기준으로써 이득률(gain)을 사용하였다. 이득률이란 분류예측 모형이 산출하는 스코어에 근거하여 해지가능성이 높은 고객을 관리하였을 때, 모형을 사용하지 않은 관리 보다 얼마나 많은 이익을 실현하였는가를 나타내는 척도이다. 예를 들어 월 평균 해지율이 1.3%인 경우 스코어에 의해 해지율이 높은 고객 5%만 관리하여 이득률 20을 얻었다는 것은 5% 관리 고객 중 해지율이 $1.3 \times 20 = 26\%$ 임을 의미한다.

해지예방 캠페인에 대한 효율을 높이기 위해서는 당연히 분류예측의 정확도가 높은 모형을 사용하여야 한다. 일반적으로 예측 정확도가 높은 모형은 신경망, projection pursuit regression 등과 같은 복잡한 비모수적 이론에 근거하여 결과의 해석이 불가능하다. 반면 나무모형(Tree model)은 결과 해석이 명쾌하고 목표 변수의 수준별 고객군의 발굴이 가능하여 마케팅과 관련된 전략도출에 활용될 수 있으나 정확도 측면에서는 그리 매력적이지 못하다.

이와 같이 사용 모형에 분명한 특징이 있으므로 데이터 마이닝 분석 시 사용할 모형의 선택은 분석 결과를 어떠한 부서에서 어떤 목적으로 사용하는가에 따라 달라져야 한다. 예를 들어 해지의 가능성이 높은 고객에 대한 특성을 파악하여 캠페인 전략을 세워야하는 마케팅 부서에서는 나무모형을 선호하고 해지예측에 대한 정확성이 업무 효율에 직접적인 영향을 미치는 콜 센터에서는 신경망과 같은 모형을 필요로 할 것이다. 즉, 분석 결과를 어느 부서에서 사용하는가에 따라 선호하는 모형이 다르고 결국 부서간 불협화음의 원인이 되기도 한다.

본 연구에서 소개하는 이동통신사의 해지자 예측의 경우 실제로 마케팅 부서와 해지 예상 고객 리스트를 실제로 작성하여 콜 센터에 제공하여야하는 CRM 팀에서의 스코어 사용 목적이 달라 나무모형이든 신경망이든 어떠한 모형을 선택해도 요구 사항을 만족시킬 수 없었다. 그래서 그와 같은 한계를 극복하기 위하여 사용되었던 모형이 혼성모형(hybrid model)이었다. 혼성모형은

해석이 용이한 나무모형과 정확도가 높은 신경망 혹은 projection pursuit regression과 같은 모형을 결합하여 각 모형의 장점을 최대한 활용할 수 있는 방안으로 미국의 데이터 마이닝 솔루션 제조업체인 Salford system에 의해 처음 소개되었다.

혼성모형의 기본 골격은 우선 나무모형을 이용하여 분류모형을 형성 한 후, 나무모형의 종료 노드 번호를 가변수로 형성한 후 신경망과 같이 모형 해석은 불가능하나 정확성이 높은 모형의 설명변수로 사용하는 것이다. 이 과정에서 나무모형의 결과는 모형의 해석이 필요한 마케팅 부서에 제공되고 나무모형 결과를 신경망 등에 적용하여 정확도가 향상된 해지 예상자 리스트를 콜 센터에 전달되어 각 부서의 요구 사항을 모두 만족하자는 것이다. 혼성모형을 사용하여 실제로 나무모형만을 사용하는 것 보다 이득률이 30% 이상 향상되었고 이를 통해 효율적 해지 이탈방지가 가능하였다.

혼성모형은 많은 현실적인 문제를 해결할 수 있으나 모형을 적용하는데 있어 다음과 같은 사항들을 고려하여야 한다.

① 혼성모형 중 첫 단계로서 나무모형을 적용할 때 설명변수로써 어떠한 것을 사용하는 가를 결정하여야 한다. 즉, 모든 설명변수를 사용한 후 나무모형에 의해 제공되는 노드 정보를 가변수화 할 것이지, 아니면 나무모형 분석 후 사용될 신경망은 원칙적으로 연속형 변수만을 입력변수로 사용하기 때문에 나무모형에서는 이산변수만을 설명변수로만 사용할 것인가를 분석가가 설정하여야 한다. 그러나 이러한 문제가 체계적으로 연구된 바 없고 이론적으로 설명되기 어렵기 때문에 분석가의 경험과 여러 차례의 모형선택 과정에 의해 결정될 수밖에 없다.

② 혼성모형에 있어 나무모형의 꺾적의 크기, 즉 종료 노드의 가장 적절한 개수를 정하여야 한다. 나무모형은 가지치기 과정을 통해 최적의 나무 크기를 결정하지만 제공되는 나무의 크기가 너무 커서 가변수 증가의 원인이 되는 경우, 오히려 혼합모형이 나무모형만을 사용하는 것 보다 못한 정확도를 제공할 수도 있는 것이다.

③ 나무모형 분석 후 도출된 결과에서 이탈고객의 해지 예상정도를 군의 형태로 파악할 수 있다. 형성된 군, 즉 종료노드를 가변수화 하여 신경망에 적용한 후 산출되는 해지 예상 스코어가 나무모형의 해지 예상 순위와 다를 수 있다는 문제점이 있다. 예를 들어 A고객이 B고객 보다 나무모형에 의하면 해지 가능성이 높다고 예상되었으나 신경망에 적용 후 B고객의 해지 가능성 스코어가 A고객 보다 높을 수 있다는 것이다. 이와 같은 순위의 변동을 통해 혼합모형의 정확도가 나무모형만의 사용 보다 향상되는 것이기도 하지만 너무 심한 순위의 변동은 스코어 사용처, 즉 마케팅 부서 혹은 콜 센터가 서로

판이한 결과를 사용한다는 문제점이 있는 것이다.

Location	Need	Model	Solution
Back-CRM (marketing)	Interpretation	Tree Model	
		NN-after-CART	
Front-CRM (call-center)	Accuracy	Neural Network	

그림-6 혼성모형의 포지션

5. S-Delpa의 화면구성

S-Delpa의 화면은 그림-7과 같이 총 4개의 프레임으로 구성되어 있다. 각 프레임별 기능을 소개하면 다음과 같다.

- 사용자관리 프레임-분석할 데이터 마트를 등록하고 출력할 스코어의 형식을 결정하는 작업 영역으로 S-Delpa 관리자에게는 사용자의 계정관리, 새로운 분석 알고리즘이 구현이 가능하도록 한다.
- 기본 틀 프레임-생성된 모형, 스코어 등을 삭제, 내용보기 등과 같이 분석에 직접 관련은 없으나 작업관리 측면에 필요한 기능을 제공한다.
- 작업 프레임-분석가가 노트를 생성하며 분석하는 장소로 과거의 분석 단계를 알 수 있다.
- S-Delpa 주 프레임-각 노드에 해당되는 파라미터 값들이나 분석할 변수의 선택을 하는 장소

6. 결론

S-Delpa는 인트라넷을 기초로 하여 사용자의 편리성과 정보의 공유성을 극대화하면서 효율적인 계정(account)과 작업관리로 보안을 유지할 수 있다. 결국 다양한 기능의 그래프와 데이터 마이닝 알고리즘과 함께 S-Delpa는 최첨단의 솔루션을 제공하고 있는 것이다.

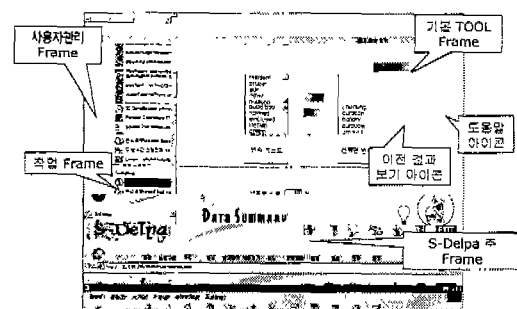


그림-7 S-Delpa 화면구성

참고문헌

- [1] Breiman, L.(1996) Bagging predictors, *Machine Learning*, 26, 123-140
- [2] Breiman, L.(1998). Arcing classifiers (with discussion), *Annals of Statistics*, 26, 801-824
- [3] Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion), *Annals of Statistics*, 19, 1-141
- [4] Kooperberg, C., Bose, S., and Stone, C. J. (1997). *J. Amer. Statist. Assoc.*, 92, 117-127