응용통계연구
제 4 권 제 1 호, 1991년

통계전문가시스템의 특성과 개발도구의 선택*

허 문 열**

〈요 약〉

본 논문에서는 마이크로 컴퓨터를 환경으로 하여 개발되고 있는 통계 소프트웨어의 경향과 문제점을 기술하고 이를 보완하기 위한 하나의 방안으로 통계전문가시스템을 제안하며, 이 시스템의 구성 방안과 이를 구현하기 위한 도구 선택의 전략에 대해 논하고자 한다.

1. 통계소프트웨어의 발전 경향

1975년 Intel 8080 프로세서에 256 바이트의 기억용량을 갖춘 Altair 8800이 발표된지 15년이 지난 지금 우리는 마이크로 컴퓨터분야에서 엄청난 변화를 목격하고 있다. 이러한 발전은 통계학에도 지대한 영향을 미치고 있으며 통계인들이 많이 사용하던 메인 컴퓨터용 팩키지들이 이제는 PC에서도 가능하게 되었다. 또한 최근에는 PC만을 환경으로 하는 팩키지들이 개발되고 있으며 종래에는 불가능하던 여러가지 기능이 PC용 팩키지에서 가능해지고 있다. 이들의 경향은 다음의 두 가지로 집약될 수 있다.

(1) 통계팩키지의 대중화

통계팩키지에 전혀 경험이 없는 사람이라고 하더라도 한 두시간의 노력으로 이를 이용할 수 있도록 설계되고 있다. 특히 메뉴판을 이용한 통계모형의 선택과 그래픽기능을 이용한 도형처리 기능이 추가됨으로써 사용자에게 더욱 편리하고 친밀감을 주는 팩키지로 발전하게 되었다. 즉, 컴퓨터의 대중화와 더불어 통계팩키지의 대중화가 이루어지고 있다.

(2) 팩키지의 대형화와 기능의 직접화

자료의 편집, 그래픽 기능에 의한 도형처리, 수치계산, 팩키지 사용안내 및 도움말 기능, 그리고 제반 통계계산 모듈들이 한 팩키지에 모두 포함되어 있다. 따라서 팩키지 사이즈가 필연적으로 대형화되어가고 있다.

이러한 경향은 다음과 같은 이유로 통계인들에게 많은 우려가 되고 있다.

* 이 논문은 1990년도 문교부 지원 한국학술진흥재단의 자유공모과제 학술연구조성비에 의하여 연구되었음.

** (110-745) 서울시 종로구 명륜동 성균관대학교 경상대학 통계학과

1. 팩키지가 직접화, 대형화로 변화하면서 팩키지의 값이 비싸지며, 팩키지를 수용할 수 있는 컴퓨터 하드웨어의 값이 비싸진다. 또한 여러가지 기능을 동시에 효율적으로 운용하도록 시스템을 구성하여야 하기 때문에 시스템의 효율이 떨어진다. 사용자가 실제로 사용하는 부분은 매우 한정되어 있다. 따라서 거대한 팩키지를 구입하여 설치한 후 한번도 테스트조차 하지 않은 부분이 많다.

2. 팩키지가 대중화되고 사용이 편리해짐에 따라 팩키지에 제공되는 데이타가 어떤 형태이든 일단 입력이 되면 아름다운 결과가 보장된다. 특히 그래픽 기능 등에 의해 분석결과가 아름답게 포장될 때 통계분석 과정의 오류들이 덮어져버릴 수 있다.

그러나 최근에는 더욱 성능이 좋은 하드웨어와 스프트웨어들이 저렴한 값으로 소비자에게 제공되고 있기 때문에 첫번째 우려는 없어질 전망이다. 예를 들면 S, SAS, SPSS 등의 대형 통계팩키지들을 PC에 설치하여 사용하는 데 필요한 하드웨어가 이제는 대중화되고 있으며 앞으로는 이러한 추이가 더욱 가속화될 전망이다. 더욱이 하드웨어의 대중화와 더불어 팩키지의 대중화가 이루어지면 팩키지 가격이 저렴해지고 이러한 영향은 더욱더 팩키지의 대형화와 대중화에 기여하게 될 것이다. 이러한 경향이 가속화 될 수록 통계인이 갖는 두번째 우려는 더욱 심각해지게 된다. 다행히 최근에 활발히 연구되고 있는 전문가적 접근법을 적용하면 이러한 문제가 어느정도 해결될 수 있다.

통계전문가시스템이라고 하면 사용자가 통계적 문제를 해결하고자 할 때 통계전문가가 이를 해결하는 것과 유사한 과정을 컴퓨터가 수행하도록 만들어 놓은 소프트웨어를 의미한다. 이러한 소프트웨어는 팩키지 사용자를 교육시킬 수도 있고 사용자가 팩키지를 잘못 이용하는 경우 이를 시정할 수도 있다. 따라서 이상적인 통계전문가시스템이 만들어진다면 팩키지의 대중화로 인한 여러가지 문제점이 해결될 수 있다.

2. 통계전문가시스템의 현황

사용자가 통계전문가에게 상담을 구하는 문제는 구조적(structured)인 문제로부터 비구조적(ill-structured)인 문제에 이르는 스펙트럼을 갖고 있다. 문제가 잘 정의되어 있는 경우를 구조적이라고 한다. 예를 들어 정해진 실험계획에 의해 실험을 수행하고 여기서 획득된 데이타를 분산분석을 통해 분석하고자 하는 경우가 여기에 속한다. 이에 반해 비구조적인 문제라고하면 문제 자체가 명확히 정의되어 있지 않은 경우를 의미한다. 여론조사를 위한 표본설계를 작성하는 문제, 또는 실험을 통해 획득된 데이타에 적절한 통계적 방법을 찾아내는 문제 등이 여기에 속한다. 또한 문제 자체는 명확히 정의되어 있다고 하더라도 이를 구조화된 규칙으로 표현하기가 어려운 경우가 있다. 현재까지 개발된 대부분의 통계전문가시스템은 비교적 구조적인 문제를 해결하는데 목표를 두고 있다. 이들 중 많이 알려진 것들은 다음과 같다.

REX(Gale and Pregibon, 1986)

회귀분석을 위한 전문가시스템으로서 다음과 같은 기능이 있다.

· 입력 데이타가 회귀분석을 위한 가정에 적합한지를 검사한다.
· 데이타가 가정에 맞지 않을 때 적절한 변환을 제시한다.
· 사용자가 요구할 때 시스템이 수행한 절차에 대해 설명을 해준다.

UNIX 에서 LISP 언어에 의해 작성되었으며 통계계산은 통계팩키지 S 를 이용한다. 다양한 그래픽기능을 이용하고 있다.

NAXPERT(Schulze and Cryer, 1988)

FORTRAN 으로 만들어진 50 여개의 수학 루틴들 중에서 사용자가 자기 목적에 맞는 것을 골라내는 데 도움을 준다. IBM PC 에서 PROLOG 언어로 작성되었으며 그래픽기능이 없다.

KENS(Hand, 1987)

사용자에게 약 500 여가지의 비모수적 방법드레 대한 지식을 보강하여 줌으로서 관심있는 내용에 대한 적절한 비모수적 방법을 사용자가 선택하는 데 도움을 제공한다. IBM PC 에서 C 언어로 작성하였다. 이 시스템에는 지식 베이스만 포함되어 있으며 통계계산부분은 없다. 또한 그래픽 기능도 없다.

FORECAST PRO(Goodrich and Stellwagen, 1988)

시계열자료의 예측을 위한 상업적인 전문가시스템으로서 처리과정은 REX 와 비슷하다. IBM PC 에서 수행되고 메뉴판 등의 그래픽 기능이 있다.

FMSCS(Kwong and Cheng, 1988)

여러가지 예측 모형 중에서 가장 적절한 모형을 선택하는 데 도움을 준다. FORECAST PRO 와는 달리 실제 예측을 수행해 주는 통계계산 부분은 포함되어 있지 않다. IBM PC에서 PASCAL 로 작성되었다.

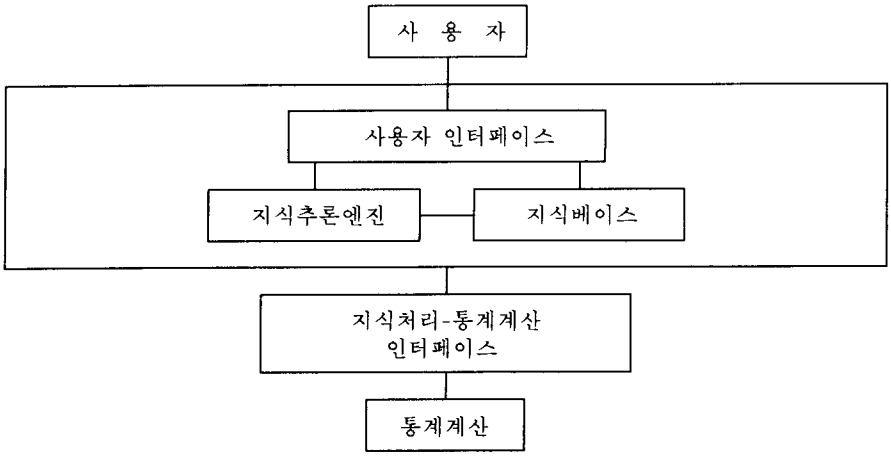
이 외에도 Rodel 과 Wilke(1990)에 의한 시스템으로 2 변수 분포에서 독립성 검정을 위한 적절한 방법을 찾아내고 이 방법에 따라 검정을 실시해주는 지식베이스시스템이 있다. 이 시스템도 IBM PC 에서 운영되고 있으며, 지식처리는 PROLOG, 통계계산은 PASCAL 을 이용하고 있다.

3. 통계전문가시스템의 구성

통계전문가시스템은 이 시스템을 사용하는 사용자, 통계전문가의 지식을 저장, 관리, 처리하는 지식처리부분, 통계적 분석에 필요한 여러가지 알고리즘을 처리해주는 프로그램의 집합으로 이루어진 통계계산부분, 그리고 지식처리부분과 통계계산부분을 연결시켜주는 인터페이스부분으로 이루어져 있다.

사용자는 통계적 지식이 부족한 초보자로부터 통계전문인에 이르는 사람을 다 포함하고있다. 통계계산부분은 기존 통계팩키지 안에서 처리될 수 있는 것들이다. 지식처리부분은 해당분야의 전문지식들을 모아놓은 지식베이스(knowledge base), 지식베이스를 통제, 관리하고

추론을 통해 주어진 문제를 해결하는 추론엔진(inference engine), 그리고 사용자와 시스템과의 대화를 처리해 주는 사용자인터페이스(user interface)로 구성되어 있다. 여기서는 지식처리부분과 인터페이스 부분에 대해서만 더 자세히 설명하기로 한다.



〈그림 1〉 통계전문가시스템 구성도

3-1. 통계전문가시스템의 지식처리부분

예를 들어 데이타분석을 하기위한 첫번째 과정으로 많이 이용되는 데이타 변환에 대해 다음과 같은 규칙이 있다고 하자.

규칙 A :
IF 데이타가 모두 양이고 분포의 형태는 좌측으로 기울어져 있다.
THEN 바람직한 변환은 LOG 변환이다.

규칙 A-1 :
IF 데이타의 최소값이 0보다 크다.
THEN 데이타는 모두 양이다.

규칙 A-2 :
IF 외도가 0보다 작다.
THEN 분포의 형태는 좌측으로 기울어져 있다.

지식베이스에는 위와 같은 규칙들의 집합이 들어 있다. 좋은 팩키지라고 하면 효율적인알고리즘을 많이 갖고 있어야 하는 것과 같이, 좋은 전문가시스템이라고 하면 고질의 지식을 많이 갖고 있어야 한다. 따라서 전문가시스템을 지식베이스시스템(knowledge-based system)

이라고 부른다.

이제 "바람직한 변환이 무엇인가?"와 같은 문제가 주어지면 여기에 적절한 답을 획득하기 위해 지식베이스를 탐색한다. 탐색하는 방법은 여러가지가 있으나 역방향 사슬(backward chaining)에 의한 추론방법이 가장 널리 이용되고 있다. 바람직한 변환을 찾아내기 위한 역방향사슬의 추론절차는 다음과 같다. 우선 규칙 A의 전제조건인 "데이타가 모두 양이고, 분포가 좌측으로 기울어져 있다"의 진위를 조사한다. 그러나 규칙 A-1에 의하면 외도가 0보다 작을 때 분포가 좌측으로 기울어져 있다고 하였다. 이 두가지 규칙의 전제들에 대한 진위를 결정하는 것이 다음 작업 과정이다. 이 과정을 처리하기 위해서는 데이타로부터 최소값과 외도를 계산하는 방법이 있고, 사용자와 대화를 통해 진위를 결정하는 방법이 있다. 각 방법에는 장·단점이 있으며 여기에 대해서는 다음에 논하기로 한다. 다만 여기서 언급할 내용은 추론 엔진에서는 이와 같이 지식베이스에 있는 규칙들을 관리, 탐색하고 추론 과정을 결정하여 주어진 문제를 해결해 나가는 기능을 수행한다는 것이다.

사용자와 시스템과의 대화는 사용자 인터페이스에서 이루어진다. 문제에 따라서는 사용자 인터페이스가 매우 중요한 역할을 수행하게 된다. 예를 들어 데이타가 모두 양인가를 조사하기위한 규칙 A-1대신 시스템이 사용자에게

"모든 데이타가 양인가?"

라는 질문을 함으로써 문제를 쉽게 해결할 수 있다. 또 분포의 형태가 좌측으로 기울어져 있는가를 알기 위해 시스템이 데이타를 이용하여 히스토그램을 그리고 사용자에게 이를 제시하여 사용자로 하여금 이를 판단하게 한다. 이러한 경우 이외에도 지식을 규칙으로 형식화하기가 어려운 경우 사용자와의 대화를 통해 간단히 해결할 수 있다. 예를 들어 관심이 있는 변수가 순위변수인지 명목변수인지 알고자하는 경우 사용자와 대화를 통해 간단히 해결할 수 있으나 이를 전문가적 지식에 의한 규칙으로 만들고자 하면 문제가 복잡해진다.

사용자 인터페이스에서 수행하는 작업중에 또 다른 중요한 부분은 작업을 수행하는 도중 어느 때에라도 사용자의 요구가 있다면 사용자에게 전문지식, 또는 시스템을 운용하는 지식에 대해 설명을 하여주고, 현재까지 진행해온 제반 과정에 대해서도 사용자의 요청이 있을 때 설명하여주는 것이다. 최근에 발표되고 있는 팩키지들 중에도 "user-friendly"의 개념을 도입한 팩키지는 메뉴판 등을 이용하여 사용자와 팩키지 간에 대화를 수행하는 기능이 있으나 이는 해당 통계계산 모듈에 대한 정적인 설명과 데이타의 입력 형식에 대한 검사 또는 도움말 등이 이루어질 뿐이다.

3-2. 지식처리-통계계산 인터페이스

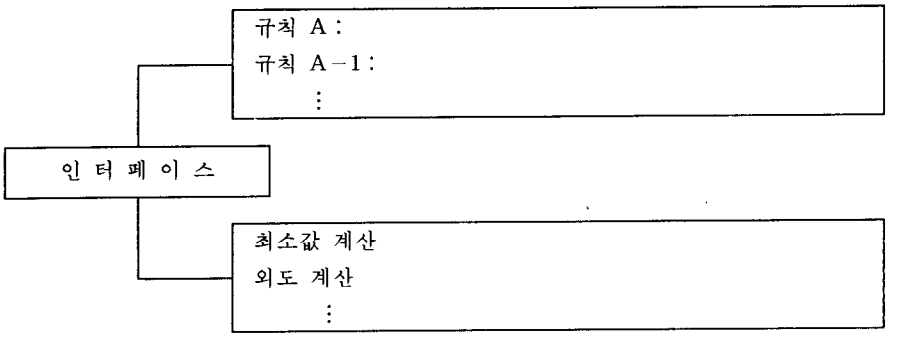
지식처리부분과 통계계산부분은 각각 처리하고자 하는 대상의 성격에 차이가 있다. 전자의 경우 전문가의 지식을 규칙으로 만들어 놓은 지식베이스를 탐색하여 문제에 적절한 결론을 제시하는 것이 주 목적이다. 여기서 전문가의 지식은 논리적인 문장들로 이루어져 있으므로 지식처리는 심볼처리가 대부분을 이루고 있다. 따라서 수학적 계산이 주류를 이루고

있는 통계계산 부분과는 처리하는 대상이 다르다. 이와 같이 상이한 목적을 갖는 두가지 부분을 연결시켜주는 기능을 수행하는 것이 지식처리-통계계산 인터페이스이다.

예를 들어 앞절에 주어진 규칙 A-2의 IF 부분인 '외도가 0보다 크다'를 판단하기 위해 통계계산부분의 '외도 계산' 부분을 찾아가야 한다. 외도의 계산이 끝나면 이 값을 갖고 지식처리부분의 해당규칙(A-2)에 돌아가야 한다.

지식처리를 위해 많이 이용되는 도구는 LISP, PROLOG 등과 같은 AI 언어이다. 이 중에 PROLOG 는 IF-THEN 으로 이루어지는 논리식을 처리하는데 매우 효율적이다. 반면 LISP 는 기호처리를 하는 데 효율적이다. LISP 나 PROLOG 가 지식처리를 위해 개발된 언어라고 하더라도 이를 숙달하여 통계전문가시스템을 만드는 데는 어려움이 많다. 이러한 불편을 덜기 위해 전문가시스템 개발도구가 개발되었으며 OPS5, EMYCIN, PCPLUS 등이 여기에 속한다. Gottinger(1988)도 EMYCIN 을 이용하면 통계전문가시스템을 쉽게 만들 수 있다고 제안하였다. 그러나 현재까지 개발된 대부분 통계전문가시스템들은 이러한 개발도구를 이용하지 않고 있다.

통계계산을 위해서는 이미 많은 팩키지들이 개발되었다. 물론 이들 팩키지들은 각각 장단점을 갖고 있다. 팩키지가 STATGRAPHICS 와 같이 고급화가 될수록 사용하기에는 편리하지만 외부 시스템과 데이타의 호환성이 줄어든다(여기서 "고급"이라고 하면 여러가지 기준을 적용할 수 있겠으나 사용자에게 더욱 친밀한 도구를 의미한다). 반면 GAUSS, IMSL 등과 같은 팩키지의 경우 이를 숙달하기 위해서는 상당한 노력이 필요하나 다른 소프트웨어와 자유로이 데이타를 호환할 수 있는 기능이 있어 유용한 도구로 이용될 수 있다. 물론 S 나 SAS, SPSS 등과 같은 통계팩키지도 외부 시스템과 데이타를 호환할 수 있는 훌륭한 기능을 갖고 있다. 그러나 통계전문가시스템은 본질적으로 지식처리기능이 주류를 차지하고 있기 때문에 지식처리도구가 셸(shell)이 되고 통계계산 부분이 요소가 되어야 한다. 이러한 관점에서 볼 때 S 등과 같은 팩키지는 이 자체가 하나의 시스템이기 때문에 지식처리 도구의 한 요소로 처리하기가 어려워 통계계산도구로 사용하기가 어렵다.



〈그림 2〉 지식처리-통계계산 인터페이스의 예

4. 결 론

통계전문가시스템이 실용성을 갖추려면 좁은 분야에서부터 출발하여야 한다. 또한 이 시스템은 고질의 지식을 많이 갖고있어야 한다. 이 지식들은 IF-THEN 의 형식에 의한 논리적인 형태를 갖추어야 하며 각각의 규칙은 다른 규칙들과 독립적으로 만들어져서 그 자체 하나만으로 완전한 의미를 가져야 한다. 이를 위하여 각 분야의 통계전문인들로부터 고질의 지식을 획득하여 이를 규칙으로 체계화시키는 작업이 필요하다.

통계전문가시스템은 다른 전문가시스템과 달리 논리적 지식처리 기능 뿐만 아니라 지식추론과정에서 필요한 여러가지 통계계산 기능을 갖고 있어야 한다. 지식처리를 위해서는 전문가 시스템 개발도구를 이용하는 것이 효율적이고 통계계산을 위한 도구로는 기존 통계팩키지를 이용하는 것이 효율적이다. 그러나 이 두가지 도구가 하나는 기호처리를 위한 것이고 하나는 수식처리를 위한 것이기 때문에 서로 상반된 특성을 갖게 된다. 대부분의 전문가시스템 개발도구는 C 나 FORTRAN 등의 언어와 호환성을 갖도록 만들어져 있으므로 계산과정이 간단한 경우 이들 언어를 이용하여 통계계산을 처리하면 어려움을 피해갈 수 있다. 그러나 통계계산이 복잡해지면 이러한 방법은 한계를 갖게 된다. 이러한 경우에는 통계계산을 위해서는 GAUSS, IMSL 등을 이용하고 지식처리-통계계산 인터페이스는 C 언어를 이용하는 것이 바람직하다. 또 다른 대안으로 S 나 SAS 등의 시스템적 팩키지가 셀이 되는 통계전문가시스템을 생각할 수도 있다. 이 시스템은 통계계산이 주(主)가 되고 지식처리가 부(副)가 되는 것으로서 지식처리를 하는 과정이 어려워진다. 이 경우 이들 팩키지가 갖고 있는 외부 데이타와의 인터페이스 기능을 통해 LISP 나 PROLOG 로 만들어진 지식처리부분을 연결함으로서 시스템의 구성이 가능할 수도 있으나 이를 구현하는 과정은 많은 어려움이 있을 것으로 예상된다.

◇ 참 고 문 헌 ◇

[1] Goodrich, Robert and Eric Stellwagen, *FORECAST PRO OPERATIONS MANUAL*, Buisiness Forecast Systems, Inc. 1988

[2] Gottinger, Hans. W. (1988), "Statistical Expert System", ***Expert Systems***, **5**, No. 3(XSTAT)

[3] Hand, D. J. (1987), *KENS Users Manual.*

[4] Harmon, Paul and David King (1985), *Artificial Intelligence In Buisiness.*

[5] Kwong, K. Kern and Donald Cheng (1988), "A prototype microcomputer forecasting expert system", The Journal of Buisiness Forecasting, Spring.

[6] PCPLUS (1987), *Personal Consultant Plus reference manual,* Texas Instrument Inc.,

[7] Pregibon, D. and W. Gale (1984), "REX : an expert system for regression analysis", *COMPSTAT*, 242 – 248.

[8] Rodel, Egman and Robert Wilke (1990), "A knowledge based system for testing bivariate dependence", Statistical Software Newsletter **16**, No. 1.

[9] Schulze, Klaus and Colin W. Cryer (1988), "NAXPERT : a prototype expert system for numerical software", *SIAM J. Sci. STAT. COMPT.*, **9**, No. 3, May.

On the characteristics of statistical expert system and a strategy to choose development tools *

Huh, Moon Yul**

〈Abstract〉

This paper describes the trend and inherent problems of the current statistical packages, and statistical expert system is suggested as an alternative to the conventional statistical packages. The paper then describes the components and characteristics of statistical expert system, and suggests a strategy to choose development tools to build a system.

* This paper was supported in part by NON DIRECTED RESEARCH FUND, Korea Research Foundation, 1990.

** Department of Statistics, Sung Kyun Kwan University