주요용어 : 간호요구, 데이터 마이닝

# 데이터 마이닝을 이용한 입원 암 환자 간호 중증도 예측모델 구축*

박 선 아**

## Ⅰ. 서 론

우리나라 암 사망자 수는 2002년 우리나라 총 사망자 246,515명 가운데 25.5%(남자 사망자의 29.6%, 여자 사망자의 20.5%)인 62,887명으로 암으로 인한 사망이 사망원인의 1위를 차지하고 있다(국립암센터, 2004).

2000년도 건강보험통계자료에 의하면 건강보험에서 암치료를 받은 사람은 218,735명으로 이 가운데 입원치료를 받은 사람만 157,440명이었으며 2000년도에 새롭게 입원한 환자는 101,781명이었다(박찬형, 2001).

전국의 암 환자 이용 진료건수가 년간 2,400,000건 정도이고, 입원의료이용건수는 437,000건으로 전체 입원의료건수의 10%를 차지하고 있어 진료건수와 진료비가 지속적으로 증가하고 있으며, 이러한 증가경향은 더욱 늘어날 것으로 예상되고 있다. 따라서 암질환에 대한 포괄적인 관리가 요구된다(박찬형, 2001).

암 환자를 간호하는 인력은 2003년 10월 전문간호사 제도가 확대되면서 중환자간호에 암 환자 간호를 포함시킴으로써 그 전문성을 인정받았다.

현재까지 우리나라에서 암 환자의 연구에 사용된 분류는 병태 생리학적인 것이었는데, 간호중증도 분류를 입원 암 환자에게 적용함으로써 환자의 상태를 간호요구도와 간호제공에 필요한 시간에 따라 일정한 기준으로 분류(박정호, 송미숙, 1990)하여 암 환자 간호업무량을 예측하고자 하였다.

임상의 간호자료는 증가하고 있는데에 반해, 연구에 간호자료를 이용하는 것은 시작에 불과 하다. 간호연구자들은 서서히 간호의 질과 효율성, 효과에 대규모 자료를 이용하고 있다(Mass, Meridean, Delaney, 2004).

데이터 마이닝은 임상의 데이터로부터 의사결정을 지원하며 근거에 기초한 간호를 제공하는데 자료를 제시하는 유용한 방법이다. 데이터 마이닝을 간호연구에 활용함으로써 간호 자료의 질을 향상시키고, 간호이론을 보다 확실히 적립할 수 있다(Goodwin, Van, Lin, Talbert, 2003).

데이터 마이닝을 활용한 연구들은 주로 간호의 인력산정과 조직, 생산력 정도를 연구하는 데 초점이 맞추어져 있다(Mass Meridean, Delaney, Connie, 2004).

이에 본 연구에서는 로지스틱 회귀분석, 의사결정나무, 그리고 신경망 분석의 3가지 기법을 이용해 모델을 구축한 후 각 모델간의 예측력을 비교하여 가장 우수한 모델을 선정하도록 하였다.

데이터 마이닝을 적용하여 입원 암 환자의 간호중증도에 영향할 수 있는 여러 요인들을 알아내고 이를 고려한 적정한 암 환자 간호인력을 예측하는 기초자료를 만들고자 수행하였다.

## Ⅱ. 이론적 배경

### 1. 환자간호중증도 분류

환자중증도 분류는 환자의 간호요구사정(Patient Care Needs Asseaament)을 기초로 하여 발전하였다. 환자분류는

* 본 논문은 2004년도 고려대학교 의과대학 보건학과 석사학위 논문임
** 국립암센터 QI실

간호의 질을 파악하여 간호업무량을 예측하도록 해주는 지속적이며, 객관적인 방법으로 환자를 구분하는 방법이다(박정호와 송미숙, 1990).

우리나라에서는 적정간호 인력의 예측을 위하여 환자 분류제도에 대한 연구가 시도되었는데, 박정호 등(1990)의 종합병원에 입원한 환자의 간호원가 산정에 관한 연구에서 환자 중증도 결정을 위해 개발된 도구로 환자를 내·외과별로 구분하여 특정시간에 영양, 위생, 운동, 투약, 검사, 처치, 관찰 및 특정 8개 영역에 대한 환자 간호요구도에 따라 점수를 부과하고 이를 Ⅰ군 (경환자), Ⅱ군 (중등환자), Ⅲ군(중환자), Ⅳ군(위독환자)로 분류하였다.

이를 1992년 임상간호사회에서 채택하여 사용하고 있으며, 본 연구에서 사용하였다.

## 2. 데이터 마이닝

1) 데이터 마이닝의 개념

데이터베이스로부터 과거에는 알지 못했지만 데이터 속에서 유도된 새로운 데이터 모델을 발견하여 미래에 실행 가능한 정보를 추출해 내고 의사 결정에 이용하는 과정을 말한다.

즉 데이터에 숨겨진 패턴과 관계를 찾아내어 광맥을 찾아내듯이 정보를 발견해 내는 것이다. 여기에서 정보 발견이란 데이터에 고급 통계 분석과 모델링 기법을 적용하여 유용한 패턴과 관계를 찾아내는 과정이다.

2) 간호분야에서 데이터 마이닝 적용

간호분야에서는 데이터 마이닝을 이용한 연구는 시작 단계이나, 최근 대규모 자료들이 많이 지면서 연구가 증가 하고 있는 추세이다.

병원 성과에 대한 연구(Aiken, Sloane, 1998), 간호인력과 환자성과에 대한 연구(Blegen, Goode, Reed, 1998), 조산아의 기관튜브 발관 결과 예측연구(Mueller, Wagner, 2004)에 활용되었다.

데이터 마이닝은 대규모 간호자료에서 의미있는 규칙들을 발견해 냄으로써 간호사들의 의사결정의 우선순위 결정에 도움을 주며, 근거에 기초한 간호에 자료를 제공해준다. 데이터 마이닝을 이용한 연구를 통해 간호의 질과 효율성 향상을 기대할 수 있다(Mass et al., 2004).

# Ⅲ. 연구대상 및 방법

## 1. 연구대상

본 연구에서 사용한 데이터는 국내 암전문병원의 2003.1.1-2003.12.31 사이의 병원입원환자 2,228명의 환자 중증도 분류 데이터베이스 165,073건을 사용하였다.

## 2. 자료수집절차

1) 자료 정화(data cleaning)

데이터베이스의 테이블명은 등록된 진단내역(3,025건), 분류상세(138,141건), 분류등급(16,962건), 암등록내역(4,717건), 환자사항(2,228건)에 대한 자료(165,073건)을 등록번호(ID)를 중심으로 기본키로 생성하고 자료를 정렬한 후 결측치와 중복자료를 보정하고 자료정화과정을 거쳤다.

자료정화 결과 총 7,070명의 환자자료가 ID 중복, 진단명 중복 등의 중복요소 보정 및 결측치 삭제 등의 작업을 통해 실제 환자자료는 2,228명으로 나타났으며, 원시 자료(raw data) 자체의 데이터베이스 구조 자체의 정의가 불분명하고, 관계형 데이터베이스 형태를 갖추고 있지 않아 데이터베이스 테이블의 데이터 세트를 데이터모델링 작업을 통해 관계형 데이터베이스로 전환 한 후 분석작업을 실시하였다.

2) 변수의 선택

E-Miner에서 제공하는 변수 선정방법으로는 입력변수와 목표변수간의 결정계수($R^2$)를 이용하는 방법과 카이제곱통계량을 이용하는 방법이 있는데, 두가지 방법을 모두 적용해 본 결과 결정계수를 이용하여 입력변수를 선정하였을 때 모델의 예측력이 더 정확하게 나타나므로 본 연구에서는 결정계수를 이용한 변수선정방법을 선택하였다.

〈표 1〉 최종선정 된 입력변수

| 순위 | 변수명 | 변수내용 |
|---|---|---|
| 1 | section | 진료과 |
| 2 | month | 입원월 |
| 3 | dx | 진단명 |
| 4 | place | 장소 |
| 5 | ray | 방사선치료여부 |
| 6 | day | 재원일수 |
| 7 | sex | 성별 |
| 8 | age | 연령 |
| 9 | meta | 전이여부 |

## 3. 연구방법

본 연구에서는 SAS v.8.1을 사용하였다. 일반적 특성 및 데이터 분포를 확인하기 위하여 기술통계 및 빈도분석을 하였으며, 중증도 등급과 변수와의 차이를 비교하여 반응변수를 추출하기 위하여 교차분석(상관관계분석포함)을 이용한 통계방법을 시도하였다.

데이터 마이닝 분석은 SAS사의 Enterprise miner v.4.0을 이용하여 분석하였다.

로지스틱 회귀분석에서는 변수선택방법으로 단계적 방법(Stepwise method)을 선택하였고, 모델선택의 기준으로는 타당도 오류율(Validation error)을 선택하게 되면 평가용 데이터세트에서 가장 작은 오차율을 갖는 모델을 선택하게 된다. 로지스틱 회귀분석에서 오차율은 우도에서 -1을 곱한 값이 사용되었다(안지현, 2002).

의사결정나무 모델을 구축하기 위한 분리기준으로는 카이제곱통계량을 이용한 모델로 선정하였다.

신경망 모델을 구축하기 위해 가장 널리 사용되는 다중퍼셉트론(Multilayer perceptron)신경망을 사용하였으며 학습방법 역시 일반적인 감독 학습 알고리즘인 역전파 알고리즘(Back propagation)을 이용하였다. 모델선택기준으로는 평가용 데이터 세트에 대해 오분류가 가장 작은 모델을 선택하도록 설정하였다. E-Miner에서는 기본값으로 평가용 자료의 오차함수의 최소가 되도록 반복해서 추정치를 선택하였다.

## Ⅳ. 연구 결과

### 1. 일반적 특성 및 간호중증도에 영향을 주는 요인

환자 2,228명중 성별은 남자가 1,197명, 여자가 1,031명이였으며, 나이는 60대가 614명으로 가장 많았다. 재원일수로 분석하여 본 결과 7일이하의 단기 입원환자가 1,050명으로 가장 많았다.

간호중증도에 영향을 주는 요인으로 성별, 나이, 재원일수, 방사선치료여부, 항암제 투여여부, 수술여부, 전이여부, 진단명, 진료과, 입원한 월과의 상관관계를 분석하였다.

통계적으로 유의한 상관관계에 있는 변수로는 성별(p=.0204), 나이(p=.0017), 방사선치료여부(p=.0007)등이였다.

### 2. 로지스틱 회귀분석

본 연구에서는 단계적 방법을 적용하였다.

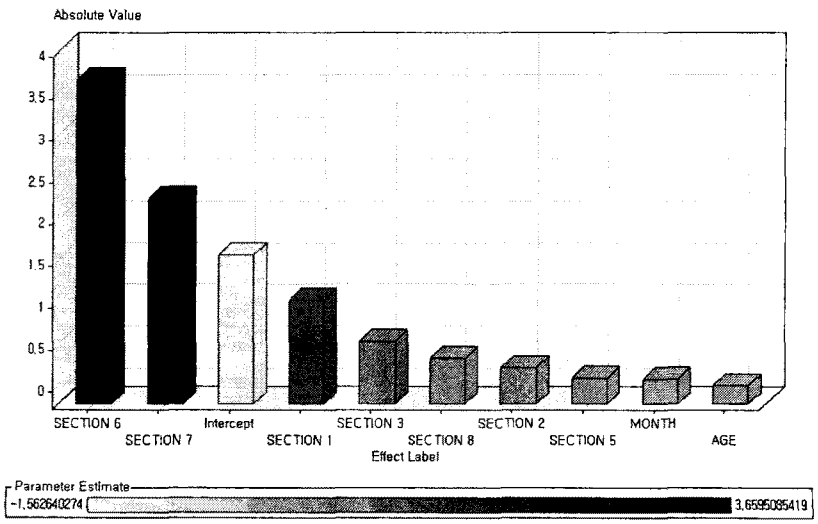
단계적 변수선택은 3단계에 거쳐 종료되었으며, 변수사이에 교호작용은 없는 것으로 나타났다.

〈표 2〉 단계적 변수선택

| 단계 | 변수 | 자유도 | 단계 | P값 |
|---|---|---|---|---|
| 1 | 진료과 | 7 | 1 | <.0001 |
| 2 | 입원월 | 1 | 2 | 0.001 |
| 3 | 나이 | 1 | 3 | 0.1574 |

〈그림 1〉은 로지스틱 회귀분석 모델의 결과로 각 변수의 T-score를 그래프적으로 표현해준다. T-Score는 Wald-카이제곱(chi-square)통계량의 절대값 제곱근을 취한 것으로 각 변수의 중요도를 나타내는 것이라고 할 수 있다. 상대적으로 큰 양의 계수는 진한색으로 표시되고 상대적으로 큰 음의 계수는 흐린색으로 표시되어있다.

회귀분석에서는 회귀계수의 부호가 교차비의 크기는 모델을




〈그림 1〉 효과 Label(Effect label)

해석하는데 유용한 정보를 준다. 즉, 회귀계수가 음의 값을 갖는다(교차비가 1보다 작다)는 것은 그 입력변수가 목표변수의 감소방향으로 영향을 준다는 것을 의미하고, 반대로 회귀계수가 양의 값을 갖는다(교차비가 1보다 크다)는 것은 증가방향으로 영향을 준다는 것을 의미한다. 따라서 위의 결과로부터 진료과, 나이, 입원월이 암 환자 중증도 분류에 영향을 미칠 가능성이 높게 추정된다는 것을 알 수 있다(표 3).

〈표 3〉 교차비(Odds ratio)

| 변수 | 교차비 |
|---|---|
| 나이 | 0.990 |
| 입원월 | 0.914 |
| 진료과 | 0.999 |

## 3. 의사결정나무

본 연구에서는 카이제곱통계량(Chi-Square statistic)에 의한 분리기준을 사용하였다.

〈표 4〉에서는 진료과, 입원월, 나이, 종양여부 등이 의미 있는 변수로 선택된 것을 볼 수 있다.

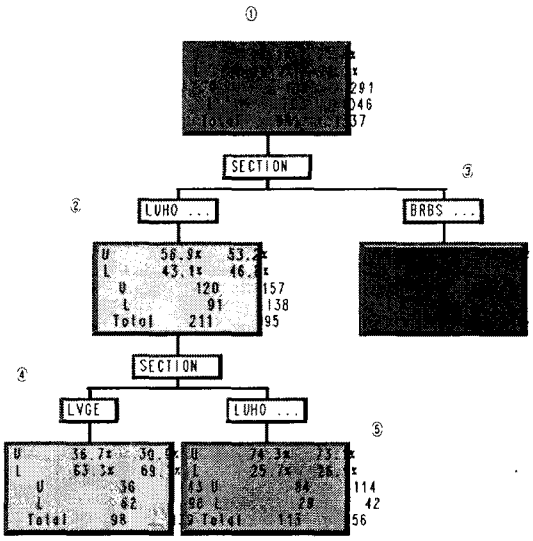
〈표 4〉 의사결정나무분석에 의한 변수선택 결과

| 변수 | 중요도 |
|---|---|
| 진료과 | 1 |
| 입원월 | 0.5821 |
| 나이 | 0.197 |
| 종양여부 | 0.1749 |
| 방사선치료여부 | 0.1538 |

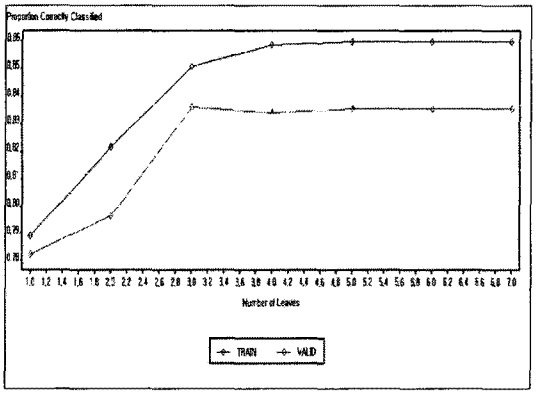
환자분류 시 진료과(Section)에 따라 LVHQ 이하와 BRBS 이상으로 나누어지며 LVHQ이하에서는 다시 LVGE와 LVHQ 이상으로 나누어져서 진료과(Section)가 중요한 변수임을 알 수 있다. 즉 전체가 U가 21%, L이 78%이지만 먼저 Section을 2군으로 나누면 한 쪽은 LVHQ이하로 여기서 U가 57%로 증가한다. BRBS이상인 군에서는 L만 87%로 거의 순수하게 L로 분류되는 것을 알 수 있다.

## 4. 신경망 분석

입력층의(input layer) 뉴런의 수는 Type노드에서의 지정에 따라 57개(input Layer : 57 neurons)로 구성하였으며, 첫번째 은닉층(hidden layer)에는 2개의 뉴런으로 구성하였고, 출력층(output layer)는 1개의 뉴런으로 구성 (목표필드: output, 등급)되었다. 여기서 등급은 I군에서 IV군까지의 환자 간호중증도 등급을 의미한다.




〈그림 2〉 의사결정나무의 구성요소




〈그림 3〉 의사결정나무 모델 결과

기본적으로 역전파 신경망을 사용하였으며, 신경망의 구조를 다음과 같이 선정하여 등급별 환자간호중증도 분류 예측문제를 해결하였다.

〈표 5〉 신경망모델의 구조

| 역전파 3층 | | | | |
|---|---|---|---|---|
| 종류 | 입력층 | 은닉층 | 출력층 | 예측률 |
| 뉴우런수 | 57 | 2 | 1 | 84.06% |

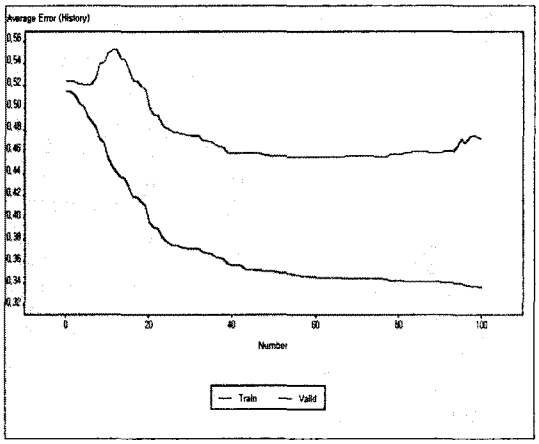
신경망 모델의 예측 정확도는 84.06%로 나타났다.

신경망의 성능평가와 예측 오차의 정확도는 평균제곱오차(MSE: Mean squared error)와 평균절대비율오차(Mean Absolute Percentage Error)를 사용하여 측정하였다.

MAPE는 오차의 절대치를 모두 더한 다음 이를 학습 자료의 수로 나눈 값이다.

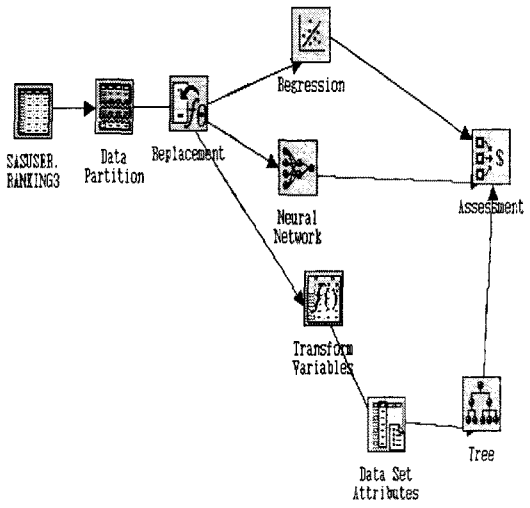
〈그림 4〉와 같이 환자간호중증도 분류 신경망의 오차곡선은 40-60번 내외에서 0.46 오차를 나타내고 있다.

입력변수의 기여도는 진료과, 진단명, 나이, 성별, 장소, 전이여부, 재원일수, 입원월은 문제해결에 기여하지만, 수술여부, 주진단여부, 분류항목, 방사선, 항암제치료여부 등은 문제해결에 대한 기여도가 낮음을 알 수 있다.




〈그림 4〉 반복에 따른 오차 함수

분석용 자료에 대한 오차함수 값은 반복회수가 증가함에 따라서 감소하지만, 평가용 자료에 대한 오차함수 값은 어느 정도 감소하다가 다시 증가하고 있다.



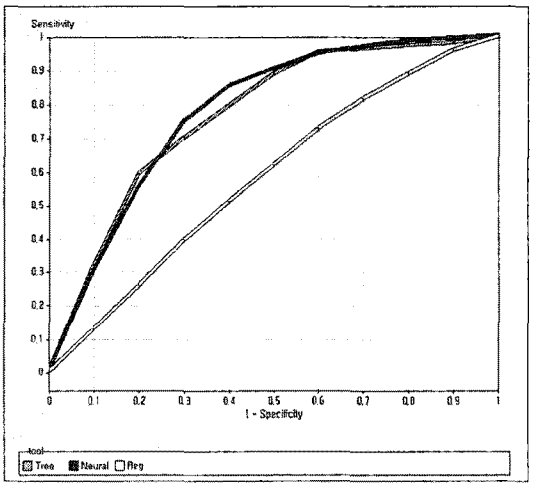

〈그림 5〉 모델 구축 과정

## 5. 모델의 평가

본 연구에서 구축된 모델에 평가용 데이터를 투입하여 나온 결과를 이용하여 세 모델의 정확도를 분석하였다(분류기준값 =0.5).

ROC 곡선에서 알 수 있듯이 신경망 모델이 의사결정나무 모델과 로지스틱회귀분석모델에 비해 전체적으로 위쪽에 위치해 있으며, 이는 각각의 분류기준값에서의 민감도와 특이도가 높으며 모델평가 시 모델의 정확도가 높은 것으로 나타난다. 특히 분류기준값이 0.2-0.5에서 세 모델차이가 최대가 된다〈그림 6〉.

ROC 곡선은 사후확률과 각 분류값에 의해 오분류 행렬을 만든 다음, 특이도와 민감도를 통해 모델평가를 수행한 것이다.



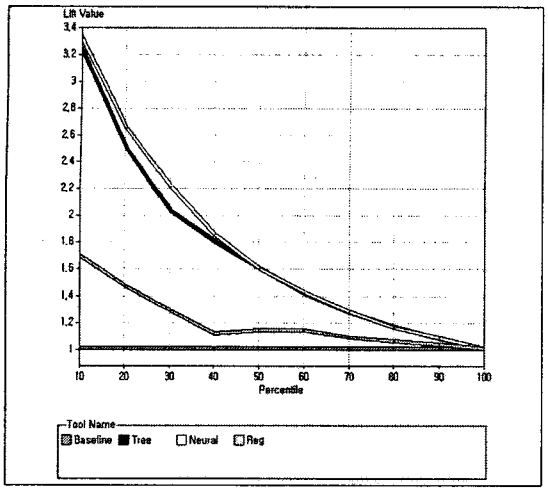

〈그림 6〉 예측모델의 ROC 곡선

이익도표는 예측모델의 성과를 비교하기 위해 LIFT라 불리는 측정치를 사용하는 것이다. LIFT가 실제 측정하는 것은 예측모델을 사용하여 모집단에서 목적에 의해 표본을 추출할 때 특정범주가 차지하는 비율의 변화이다.

〈표 6〉 예측모델의 이익도표

| 도구 | Root ASE | 오분류율 |
|---|---|---|
| 로지스틱회귀분석 | 0.408017587 | 0.210998878 |
| 의사결정나무 | 0.336640866 | 0.14365881 |
| 신경망모델 | 0.333303397 | 0.145903479 |

로지스틱회귀, 의사결정나무, 신경망분석의 LIFT Chart 분석결과 LIFT값은 상위 10%에 대해서 암 환자 중증도를 분류할 경우 기대 반응률(expected response rate)이 각각 1.7%, 3.2%, 3.3% 로 나타났다〈그림 7〉.

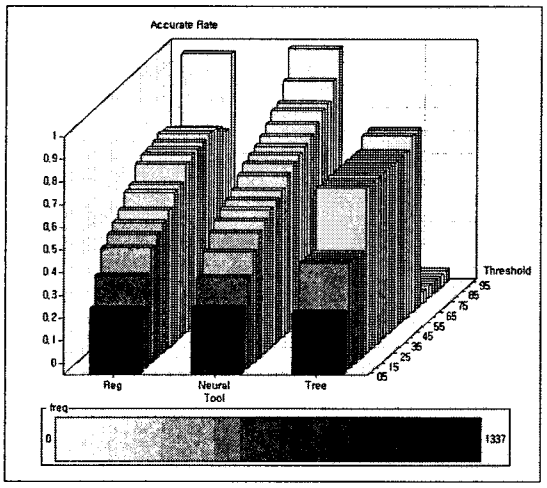



〈그림 7〉 예측모델의 LIFT 값

〈그림 8〉의 결과를 보면 다음과 같은 두 가지 결론을 내릴 수 있다.

첫째. 모델의 안정성에 있어 신경망 모델의 결과가 보다 더 우수하다. 회귀모델이나 의사결정나무 모델은 막대의 변동이 몇 군데만 발생하고 있지만 신경망모델은 지속적으로 발생하고 있다.

둘째. 그래프의 막대높이는 해당 분류기준 값에서 분류기준 값 이상의 사후 확률을 갖는 범주 1의 빈도/전체 범주1의 빈도인 정확도를 나타내는데 바람직한 형태는 극단적으로 높거나 낮은 두 부류로 나누어지는 것이 좋다. 이러한 관점을 종합해보면 신경망모델이 더 우수하다는 것을 알 수 있다.




〈그림 8〉 예측모델의 정확도 비교

## V. 논 의

본 연구에서 로지스틱 회귀분석, 의사결정나무, 신경망 분석을 이용하여 구축한 예측모델에서 진료과, 나이, 입원월이 공통적으로 환자간호중증도에 영향을 미치는 변수로 나타났다.

박정호, 조현, 박현애, 한혜라(1995) 연구에서는 내외과별, 요일별로 간호업무량에 변화를 준다고 하였는데 본 연구에서 변수로 인식된 진료과와 입원월이 영향을 주는 것으로 나타나 요일과 해당월이라는 차이는 있으나 입원 시점이 환자 간호중증도에 영향을 주는 변수로 나타났다.

박정호, 박현애, 조현, 최용선(1996)에서는 환자간호중증도로 간호업무량을 예측함에 있어 각 병원의 간호활동량 조사 연구에서 언급된 동선, 병동의 간호사 수 및 구성, 간호사의 경력, 간호전달 방법의 차이로 인한 변이가 영향을 준다고 하였으나 이번 연구에서는 그에 대한 비교가 이루어지지 못하였다.

로지스틱 회귀분석과 의사결정나무는 목표변수가 2개 이상이면 예측이 불가능하므로 환자간호중증도를 I군, II군을 묶고 III군, IV군을 묶어 분석을 실시하였다. 신경망 분석 역시 타 모델과 형평성을 맞추기 위하여 I군, II군을 묶고 III군, IV군을 묶어 분석하여 비교 하였다.

세 모델의 평가에서는 환자간호중증도를 예측하는데 신경망 분석이 우수한 것으로 나타났다.

모델 평가 연구의 선행연구를 살펴보면, 김옥남 김윤희, 강성홍(2002)연구에서는 critical pathway 실행효과연구에서 로지스틱 회귀분석과 신경망분석, 의사결정기법의 3가지 기법을 비교하였고, 로지스틱 회귀분석이 대조군을 선정하는데 효과적이라고 평가 되었다. 이견직 정영철, 기미라(2003)연구에서는 의사결정나무 기법을 이용해 영향인자를 알아내어 신경망분석에 활용하였다. 정우진, 이선미, 김원훈 (2003)의 연구에서는 로지스틱 회귀분석의 구조와 최종모델에서 추정치를 이용하여 지역 보험에 속하는 각 세대의 지역 보험료 체납 확률 예측모델을 도출하였다.

Mueller 등(2004)연구에서는 조산아의 기관튜브발관결과 예측 연구에서 신경망 분석이 로지스틱 회귀분석에 비해서 예측 정확도가 높은 것으로 나타났다.

박현애(1994)는 인공지능분야에서 신경망모델이 최근 분류(classification)에 많이 사용되고 있다. 특히 간호분야에서는 분류과정에 적용된 규칙을 찾아내기 어렵고 사용된 데이터가 불안정한 경우 신경망을 이용한 분류가 규칙이나 직관에 의거한 다른 방법보다 더 효율적인 것으로 알려져 있다고 하였다.

본 연구에서 간호요구에 기초한 환자간호중증도 분류를 위한 신경망 모델을 구축하여 실험한 결과 모델이 내린 중증도 분류와 실제 적용된 경험에 의해 분류한 등급 분류간의 일치율이 84.06%로 나타났다.

본 연구의 제한점으로 첫째, 자료자체가 연구목적으로 수집된 것이 아니라 환자사항과 환자분류등급간에 관계형 데이터 베

이스 형태로 되어있지 않아 많은 양의 자료를 정리하여야 했다. 중환자실과 병실에서의 환자간호중증도가 통일되지 않아 중환자실 입원환자를 정리하여야 했고, 실제 환자 사항과 환자분류등급을 연결하지 못하고 자료의 결측치가 많아 7,070명의 환자 중 단지 2,228명의 환자 데이터만을 사용하였다

Goodwin 등(2003)은 데이터 마이닝을 실행하는데에 자료의 질과 결측치, 공통되지 않은 언어가 문제라고 하였다. 본 모델을 사용하여 입원한 암 환자의 간호중증도 분류를 예측하기 위해서는 자료의 질을 높이기 위해 관계형 데이터베이스로 구성하는 것이 필수적이라고 할 수 있다.

둘째, 세 모델의 평가에서 목표변수 I군, II군과 III군, IV군으로 묶어 평가하였으므로 실제 자료와는 차이가 있을 수 있다.

셋째, 여기서 사용한 변수들은 병원자료의 제한점으로 폭넓게 선택할 수 없었으며, 실제 간호 요구도를 직접적으로 반영하기 보다는 간접적으로 반영하는 지표들 일 수 있다.

넷째, 데이터 마이닝 방법론에 따른 것으로 반복적인 시행으로 나름대로 최적의 모델을 구축하였으나, 모델설정 방법이나 각 방법론의 옵션 선택에 따라 결과가 조금씩 달라질 수 있다.

이러한 제한점에도 불구하고 본 연구는 기존의 환자중증도분류와 간호업무량의 측정이 후향적이였던 반면 본 연구에서는 예측모델을 이용해 환자중증도를 예측하여 간호업무량을 전향적으로 파악하여 간호인력배치에 탄력적으로 활용할 수 있도록 시도한 것에 의의가 있다고 할 수 있다.

그 동안 간호학 관련에서 데이터 마이닝을 이용한 국내연구로는 간호진단 신경망과 전문가 시스템간의 효과비교(김정애, 1998), 간호정보의 처리, 분석, 관리 기술개발(유지수, 유황빈, 박지원, 고일선, 1998), 역전파 신경망모델을 이용한 간호진단 자율 학습프로그램 개발(김정애, 1999) 등이 있다. 주로 간호진단 예측에 관한 연구가 대부분이다. 외국의 경우에는 Erisen, Turley, Denton (1997)의 간호행위 분류에 대한 연구, Goodwin 등(2001)의 데이터 마이닝을 이용한 지역사회에서 조산 위험군 예측인자 발견, Goodwin 등 (2003)의 조산 위험군 선별에 대한 연구, Muller, Martina, Wagner, Carol (2004)의 조산아의 기관튜브발관결과 예측 연구로 그 연구 영역을 확대 하고 있다.

데이터 마이닝을 이용한 대규모 간호자료에 대한 연구는 간호의 질을 향상 시키고 임상전문가들의 의사결정을 지원하며 건강 간호행정가들의 정책을 결정하는데 중요한 자료가 될 것이다 (Eriksen, Turley and Denton, 1997).

## 참 고 문 헌


김옥남, 김윤희, 강성홍 (2002). BSC와 데이터 마이닝 기법을 이용한 Critical Pathway 실행효과 연구. *대한의료정보학회지, 8*(2), 51-67.

김정애 (1998). 간호진단 신경망과 전문가시스템간의 효과비교 *대한의료정보학회지, 4*(1), 75-81.

김정애 (1999). 역전파 신경망모델을 이용한 간호진단 자율학습프로그램 개발. *대한의료정보학회지, 5*(1), 67-76.

박정호, 조현, 박현애, 한혜라 (1995). 환자분류에 의한 일개 2차 의료기관의 간호업무량 조사. *간호행정학회지, 1*(1), 132-146.

박정호, 박현애, 조현, 최용선 (1996). 환자분류에 의한 간호인력 산정 및 배치과정 전산화. *간호학회지, 26*(2), 399-412.

박정호, 송미숙 (1990). 종합병원에 입원한 환자의 간호원가 산정에 관한 연구. *대한간호학회지, 20*(1), 16-37.

박찬형 (2001). 한국의 암정책과 종양전문간호사의 역할. *종양간호학회지, 1*(2), 231-245.

박현애 (1994). 신경망모형을 이용한 간호진단시스템. *한국보건통계학회, 11*(1), 106-107.

안지현 (2002). 데이터 마이닝을 이용한 병원이용 고객 세분화 분석. *서울대학교 보건대학교 석사학위논문, 서울.*

이견직, 정영철, 김미라 (2003). 신경망모형을 이용한 외래환자 만족도예측 및 민감도분석. *병원경영학회지, 8*(1), 81-93.

정우진, 이선미, 김원훈 (2003). 국민건강보험 지역 보험료 체납 결정요인 및 체납 확률 예측모형. *보건행정학회지, 13*(2), 85-100.

Eriksen, L. R., Turley, J. P., Denton, D. (1997). Data mining : a strategy for knowledge development and structure in nursing practice. *Stud Health Technol Inform, 46,* 383-388.

Mass, Meridean, L., Delaney, Connie (2004). Nursing Process Outcome Linkage Research. *Lippincott Williams & Wikins, 42*(2), 40-48.

Mueller, Martina, Wagner, Carol (2004). Predicting extubation Outcome in Preterm Newborns. *International Pediatric Research, 56*(1), 11-18.

Goodwin, L., Iaunnacchione, Mary Ann, Hammond, W. E. (2001). Data Mining Methods Find Demographic Predictors of Preterm Birth. *Nursing Research, 50*(6), 340-345.

Goodwin, L., Van Dyne, M., Lin, S., Talbert, S. (2003). Data mining issues and opportunities for building nursing knowledge. *Journal of biomedical informatics, 36*(4-5), 229-231.

국립암센터 (2004. 6). http://www.ncc.re.kr

- Abstract -

# An Analysis of Nursing Needs for Hospitalized Cancer Patients : Using Data Mining Techniques

*Park, Sun-A**

**Back ground**: Nurses now occupy one third of all hospital human resources. Therefore, efficient management of nursing manpower is getting more important. While it is very clear that nursing workload requirement analysis and patient severity classification should be done first for the efficient allocation of nursing workforce, these processes have been conducted manually with ad hoc rule.

**Purposes**: This study was tried to make a predict model for patient classification according to nursing need. We tried to find the easier and faster method to classify nursing patients that can help efficient management of nursing manpower.

**Methods**: The nursing patient classifications data of the hospitalized cancer patients in one of the biggest cancer center in Korea during 2003.1.1-2003.12.31 were assessed by trained nurses.

This study developed a prediction model and analyzing nursing needs by data mining techniques. Patients were classified by three different data mining techniques, (Logistic regression, Decision tree and Neural network) and the results were assessed.

**Results**: The data set was created using 165,073 records of 2,228 patients classification database.

Main explaining variables were as follows in 3 different data mining techniques.

1) Logistic regression : age, month and section.

2) Decision tree : section, month, age and tumor.

3) Neural network : section, diagnosis, age, sex, metastasis, hospital days and month.

Among these three techniques, neural network showed the best prediction power in ROC curve verification. As the result of the patient classification prediction model developed by neural network based on nurse needs, the prediction accuracy was 84.06%.

**Conclusion**: The patient classification prediction model was developed and tested in this study using real patients data. The result can be employed for more accurate calculation of required nursing staff and effective use of labor force.

**Key words : Nursing need, data mining**

* National Cancer Center Quality Improvement