

碩士學位論文

입원환자 특성분석을 위한
데이터마이닝

指導教授 李 尙 俊



濟州大學校 産業大學院

電子電氣工學科

李 아 나

2002年 6月

입원환자 특성분석을 위한 데이터마이닝

指導教授 李 尚 俊

이 論文을 工學 碩士學位 論文으로 提出함

2002年 6月 日

濟州大學校 産業大學院



電子電氣工學科
컴퓨터工學專攻

제주대학교 중앙도서관
JEJU NATIONAL UNIVERSITY LIBRARY

李 아 나

李아나의 工學 碩士學位 論文을 認准함

2002年 6月 日

審查委員長 金 壯 亨 印

審查委員 郭 鎬 榮 印

審查委員 李 尚 俊 印

목 차

Summary	1
I. 서 론	2
II. 데이터마이닝과 의료분야에서의 활용	4
1. 데이터마이닝 (Data Mining)	4
2. 의료분야에서의 데이터마이닝 활용	14
III. 환자특성분석을 위한 데이터마이닝 시스템	18
1. 연구모형	19
2. 데이터 전처리 과정	19
3. 목적설정, 항목설정 및 모형화를 위한 데이터마이닝 기법	24
IV. 결과 및 고찰	29
1. 시스템 환경과 틀	29
2. 데이터마이닝 수행 및 결과 해석	29
V. 결론 및 향후 연구	39
참고문헌	41
부 록	43

[표 차례]

Table. 1	환자특성분석에 이용될 데이터의 주요 레코드 속성	23
Table. 2	재입원패턴에 따른 군집분석 결과	33
Table. 3	재입원패턴에 따른 군집별 주요특성분석	35
Table. 4	병원이용빈도에 따른 군집분석 결과	35
Table. 5	병원이용빈도 군집별 평균연령의 비교	36
Table. 6	병원이용빈도 군집과 각 항목별 교차분석 결과	37

[그림 차례]

Fig. 1	데이터마이닝 프로세스 (KDD 프로세스)	5
Fig. 2	데이터마이닝 적용의 전체 구현 과정	18
Fig. 3	연구모형	20
Fig. 4	OCS 데이터베이스의 관련 테이블의 항목간 연관관계	21
Fig. 5	수집된 Excel 자료 파일	21
Fig. 6	분석에 사용될 정제된 Access DB 파일	22
Fig. 7	SAS 데이터분석기능을 이용한 데이터 변환	23
Fig. 8	환자특성분석에 사용될 주요변수	24
Fig. 9	재입원유무 모형평가 및 평가결과	25
Fig. 10	재입원패턴에 따른 군집분석 흐름도	26
Fig. 11	병원이용빈도에 따른 군집분석 흐름도	27
Fig. 12	병원이용빈도에 따른 군집결정에 이용된 변수 특성 및 중요도	28
Fig. 13	군집결정을 위한 특성변수의 중요도 평가 흐름도	28
Fig. 14	결정트리를 이용한 재입원 가능성 결정모형	32
Fig. 15	대화식 결정트리를 이용한 재입원 가능성 결정모형	33
Fig. 16	재입원패턴에 따른 군집별 제특성 비교 결과	34
Fig. 17	특정군집 결정을 위한 특성변수의 중요도 평가모형	38

A Study on Data Mining Application for Patient Analysis

Ah-Na Lee

*Department of Electrical and Electronic Engineering Graduate
School of Industry
Cheju National University*

Supervised by Professor Sang-Joon Lee



The advance of hardware and database leads the recent systems to save a large mass of data. Today, many methods appear to use them for the decision support. Data warehouse, OLAP and data mining are techniques to extract something important and find hidden patterns from the large data set. So, we can apply these techniques to hospital database, which saves several GBytes data every year, to get important information, to improve the quality of clinic and to study of medical science. In this study, we performed data mining on 8,300 patients' medical records who had been admitted at A hospital in cheju from Mar, 2001 to Jan, 2002. We could get characteristic information of admitted patients through the clustering for the admittance frequency and the analysis of factors letting a patient to be readmitted.

I. 서 론

다양한 형태로 분산되어 있는 방대한 양의 데이터를 수집하여 분석하기 쉬운 형태로 변환하고 데이터웨어하우스(DW: Data Warehouse)를 구축하며 이로부터 OLAP(On-Line Analytical Processing), 데이터마이닝(Data Mining) 등을 적용하여 고부가가치 데이터를 창출해내는 전반적인 기술을 지식공학(Knowledge Engineering or Business Intelligence) 이라고 하는데, 이러한 지식탐사 프로세스에 있어서 지식을 창출해내는 핵심작업이 데이터마이닝이다(Fayyad 등 1996).

데이터마이닝은 대량의 데이터로부터 패턴 인식, 통계적 기법, 인공지능 기법 등을 이용하여 숨겨져 있는 데이터간의 상호 관련성 및 유용한 정보를 추출하는 단계로 이 단계에서는 소비자 성향 파악을 통한 마케팅 및 판매 전략, 고객 지원 등과 같은 목적에서부터 의학치료방법 및 범죄 예방에 이르기까지 여러 가지 목적에 따라 알고리즘이 선택되고 원하는 결과를 얻기 위해 연속적으로 적용된다. 데이터마이닝 결과 생성되는 지식은 의사결정에 직접적인 영향을 미칠 수 있다(Michael 등 1997).

최근 우리나라 의료계는 의약분업, 의료기관 서비스 평가제 실시, 의약품 실거래가 상한제, 선택진료제도의 요건강화, 신용카드수납 의무화, 의료보험청구에 있어 전자자료교환(EDI: Electronic Data Interchange) 도입 등 의료정책 변화에 따라 병원경영의 여건이 어려워지고 있는 것이 현실이다. 또한 글로벌 경쟁사회의 도래 및 정보화·지식사회에서의 소비자(환자)의 요구가 증대되고 또 다양화되고 있어 이제 병원경영에 있어서도 새로운 방안을 모색하지 않을 수 없는 상황을 맞이하게 되었다.

또한 오늘날과 같이 변화된 비즈니스 환경에서는 기존 병원들에서 이루어지던 트랜잭션 중심의 시스템으로는 다양한 의사결정을 지원하고 정보 전략을 수립하는데 있어서 여러 가지 한계점들을 가지게 되었고, 따라서 이러한 문제점들을 해결하고자 최근 국내 병원들은 IT투자를 대폭 늘리면서 처방전달시스템(OCS: Order Communication System)의 상용화, 의료영상저장시스템(PACS: Picture Archiving and Communication System)의 구축, 전자의무기록(EMR: Electronic Medical

Record), ABC(Activity-Based Costing) 원가분석 등 통합의료정보시스템의 구축이 활발하게 이루어지고 있다.

이러한 정보시스템 구축의 일환으로 의료 데이터웨어하우스를 구축하고 다양한 보고서의 산출과 OLAP의 활용을 시도하고 있으며 보다 더 나아가 지식공학의 핵심인 데이터마이닝 기법을 적용하여 알려지지 않은 유용한 정보를 추출하고 이를 통한 다양한 의사결정을 지원할 수 있는 시스템구축이 요구되어지고 있다.

본 논문에서는 제주지역 A종합병원의 데이터베이스 시스템을 대상으로 환자의 의무기록정보를 기초로 하여 병원진료통계 및 의학연구를 목적으로 이용되어지는 기본적인 퇴원요약 관련 데이터베이스를 재구성하였다. 병원을 이용하는 입원환자를 대상으로 의료 이용횟수와 진료내역 등을 분석하고 환자의 의료기관 이용패턴을 분류하여, 고객중심의 통합되고 차별화된 진료서비스 체제를 갖출 수 있는 패턴추출 및 예측 등의 지식추출을 목적으로 분석할 수 있도록 데이터마이닝 기법을 적용하여 의사결정을 위한 기반을 마련하고자 하였다.

업무의 트랜잭션은 Unix 기반 운영체제 시스템으로 개발되었고, RDBMS를 사용하는 병원 데이터베이스 시스템으로 보다 다각적인 정보 분석을 할 수 있도록 구축되어 있으며 이는 데이터마이닝을 위한 기반으로 이용되었다. 이를 바탕으로 트랜잭션 중심의 운영 프로세스를 수행하는 시스템의 부하를 막고 개념적 분리를 하기 위하여 PC용 DB 프로그램인 Access를 이용하여 고객 분류 목적의 데이터를 선정(select)하였으며, 필요한 데이터를 추출, 정제, 전송, 로딩>Loading)하는 데이터 전처리 과정을 거쳐 데이터 집합(Data Set)을 생성한다. 자료의 분석은 통계패키지인 SAS와 데이터마이닝 툴인 SAS사의 Enterprise Miner 4.0을 이용하여 다양한 모델링 기법을 적용하고 모형을 구축하였다.

본 논문의 구성은 다음과 같다. II 장에서는 데이터마이닝에 대한 정의 및 기법과 의료분야의 데이터마이닝 활용형태에 대해서 기술하며, III 장에서는 환자 특성분석을 위한 시스템 연구모형 및 데이터 전처리 과정과 데이터마이닝 기법을 적용하는 시스템 설계를 하며, IV 장에서는 결과를 해석하며, 마지막으로 V 장에서는 결론 및 향후 연구방향을 모색하였다.

II. 데이터마이닝과 의료분야에서의 활용

1. 데이터마이닝 (Data Mining)

데이터마이닝 작업은 일반적인 규칙을 발견하고 의사결정처리를 향상시키기 위해서 과거의 데이터를 얼마나 효율적으로 사용하는가에 관한 문제이다(Tom M. Mitchell, 1999). 이러한 경향은 대규모의 데이터 저장 디바이스의 가격이 하락하고, 네트워크를 통해 데이터를 모으는 것이 용이하게 됨에 따라 데이터의 규모가 방대해지고 이를 분석하기 위한 방법들이 점차적으로 등장한 데서 기인한다.

1) 데이터마이닝의 정의

데이터마이닝에 대한 대표적인 정의들을 살펴보면 다음과 같다.

- ▶ 대량의 실제 데이터로부터 묵시적이고 전에는 알려지지 않았지만 잠재적으로 유용한 정보를 추출하는 것(Frawley et al, 1994).
- ▶ 대규모 데이터베이스 내에 존재하는, 그러나 대량의 데이터 사이에 숨겨져 있는 상호 관련성과 글로벌 패턴에 대한 탐색(Holshemier & Siebes, 1994).
- ▶ 대량의 데이터로부터 패턴인식 기술과 통계기법, 수학적 기법을 이용하여 의미있는 새로운 상관관계(correlations), 패턴 그리고 추세(trends)를 발견하는 과정(Gartner Group).

또한 데이터마이닝이란 자동화된 기능을 갖춘(automated and intelligent) 데이터베이스 분석기법으로 90년대 초반부터 지식발견(KDD: Knowledge Discovery in Database), 정보발견(information discovery), 정보수확(Information harvesting) 등의 이름으로도 소개되어 왔는데 일반적으로 ‘대량의 데이터로부터 새롭고 의미있는 정보를 추출하여 의사결정에 활용하는 작업’이라 정의된다. 용어에 ‘채굴하다’라는 의미의 ‘mining’을 포함시킨 이유는 데이터로부터 정보를 찾아내는 작업이 마치 금이나 다이아몬드를 발견하기 전에 수많은 양의 흙과 잡석들을 파헤치고 제거하는 것과 유사하다는 데에 기인한다(장남식 외, 1999).

2) 데이터마이닝 프로세스

데이터마이닝은 신경망모형이나 결정트리와 같은 특정 기법이 아니라 개념적인 정보추출의 방법론이며 일련의 과정(Process)으로, 실제 데이터마이닝의 수행은 Fig. 1과 같은 단계로 나누어 볼 수 있다.

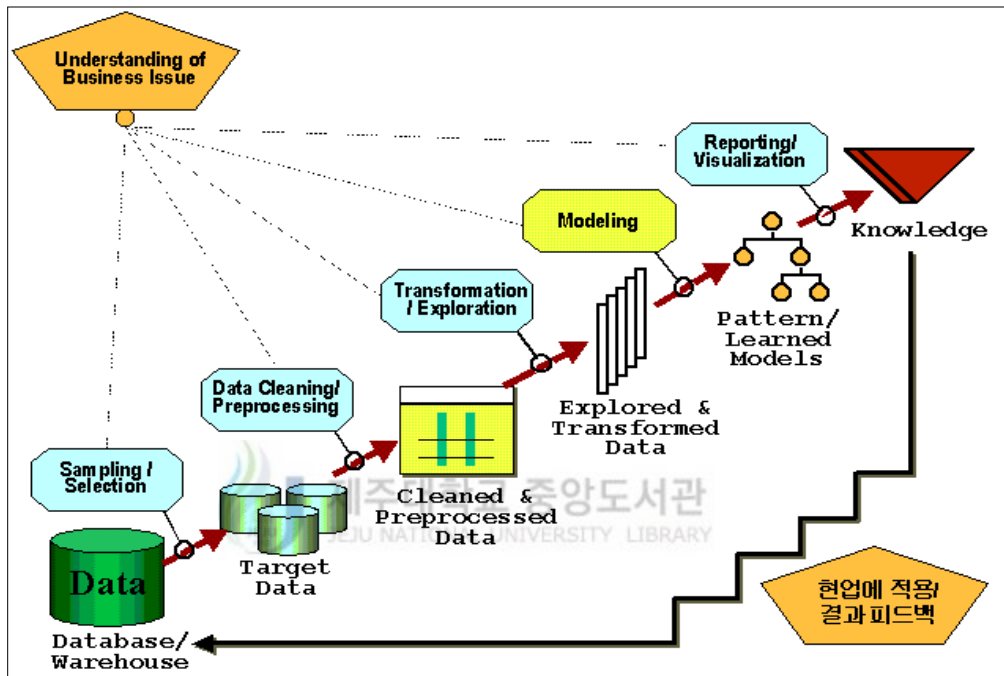


Fig. 1 데이터마이닝 프로세스 (KDD 프로세스)

(1) Sampling / Selection

데이터마이닝은 대체로 수십메가에서 수십기가에 이르는 대용량의 데이터를 기반으로 한다. 그러나 방대한 양의 데이터를 살펴보는 것은 시간의 측면에서만 보아도 많은 인내를 요하게 되는 작업이 될 수 있다. 이때 고려하여야 하는 과정이 바로 샘플링이다. 샘플링이란 방대한 양의 데이터(모집단)에서부터 모집단을 닮은 작은 양의 데이터(샘플:표본)를 추출하는 것이다.

(2) 데이터 정제 및 전처리 (Data Cleansing / Preprocessing)

데이터베이스에는 일관성이 없고 불완전하며 오류가 있는 데이터가 존재할 수

있다. 따라서 데이터 정제과정을 통해 데이터의 무결성과 질을 보장해주어야 한다.

(3) 탐색 및 변형(Exploration / Transformation)

데이터의 탐색과정에서는 이미 알고 있는 사실들을 확인하여 수치화하는 작업을 시작으로 하여 보유하고 있는 수많은 변수들의 관계를 살펴보는 단계이며, 탐색 단계에서 얻어진 정보를 기반으로 모형화 단계에서 모형의 성능을 향상시키기 위해, 그리고 데이터가 가지고 있는 정보를 효율적으로 사용할 수 있도록 변수변환, 수량화, 그룹화 같은 방법을 통해서 데이터를 변형하고 조정한다.

(4) 모형화(Modeling)

데이터마이닝 과정에서 가장 중요한 단계로서, 앞서 선행되었던 단계에서 선정된 주요한 변수를 사용하여 다양한 모형 - Neural Networks, CHAID, CART, 일반화선형모형 등의 전통적인 통계적 모형 등- 을 적합해 보는 단계이다. '신용카드의 도용방지모형은 Neural Network가 가장 적합하다' 등의 알려져 있는 모형이 있기도 하나, 그렇지 않은 경우에는 다양한 모형에 적합한 후 예측력이 가장 뛰어난 모형을 선택하는 것이 일반적이다.

(5) 보고 및 가시화(Reporting / Visualization)

데이터마이닝의 수행결과는 사용자들에게 보기 편하고 이해하기 쉬운 형태로 제공되어야 한다. 마이닝 결과를 그래프나 각종 차트 형태로 보여주는 것이 가시화이다. 가시화의 장점은 사전지식이 없이 동적인 관찰이 가능하고 인식의 한계에 대한 부담을 경감시킨다는 점이다.

3) 데이터마이닝 기법

데이터베이스로부터의 지식탐사과정에 적용되는 데이터마이닝 기법 및 알고리즘은 기계학습 알고리즘에 성능을 고려하여 동적인 실세계 데이터로부터 흥미있는 패턴을 발견하는 알고리즘으로 출발하여 지금은 다양한 기법과 혼합되어 사용되고 있다. 또한 순수 데이터마이닝을 목적으로 알고리즘들이 연구·개발되고 있다.

여러 가지 목적에 따라 알고리즘이 선택되고 원하는 결과를 얻기 위해 연속적으로 적용되는데 이러한 데이터마이닝 기법 중 군집분석, 결정트리, 신경망 등에 대해서 알아본다.

(1) 군집분석(Clustering)

군집분석(Clustering)이란 주어진 데이터 셋을 서로 유사성을 가지는 몇 개의 클러스터로 분할해 나가는 과정으로, 하나의 클러스터에 속하는 데이터 점들 간에는 서로 다른 클러스터 내의 점들과는 구분되는 유사성을 갖게 된다(Michael, 1997).

즉, p 개의 변수로 이루어진 n 개의 관측치를 그 변수들의 유사도(또는 근접도)에 따라서 q 개의 집단으로 나누는 방법으로, 군집의 최대목적은 대용량 데이터로부터 군집을 통해 잘 구분하는 것으로 동일한 군집의 개체들은 유사한 성격을 갖도록 서로 다른 군집에 속한 개체들 사이에는 그와 반대로 상대적으로 서로 다른 성격을 갖도록 군집이 형성되어야 한다.

▣ 유사성(Similarity)과 거리(Distance)의 척도

▶ 거리의 척도

서로 다른 개체 $X_i = (X_{i1}, X_{i2}, \dots, X_{ip})'$ 와 $X_j = (X_{j1}, X_{j2}, \dots, X_{jp})'$ 사이의 거리 $d_{ij} = d(X_i, X_j)$ 는 일반적으로 다음 조건을 만족시킨다

$$d_{ij} \geq 0, d_{ij} = 0$$

$$d_{ij} = d_{ji}$$

$$d_{ik} + d_{jk} \geq d_{ij}$$

위의 조건을 만족시키는 거리는 두 개체의 비유사성(Dissimilarity)의 척도를 기준하며 다음 방법들이 이용되어진다.

i) 유클리드 거리

$$d_{ij} = \sqrt{(X_i - X_j)'(X_i - X_j)}$$

ii) 마할라노비스(Mahalanobis) 거리

$$d_{ij} = (X_i - X_j)' S^{-1} (X_i - X_j)$$

iii) 민코우스키(Minkowski) 거리

$$d_{ij} = \left[\sum_{k=1}^p |X_{ik} - X_{jk}|^m \right]^{1/m}$$

▶ 유사성의 척도

두 개체의 유사성(similarity) : $S_{ij} = S(X_i, X_j)$ 는 일반적으로 상관계수를

많이 사용한다.

$$S_{ij} = \frac{\sum_{k=1}^p (X_{ik} - \overline{X_i})(X_{jk} - \overline{X_j})}{\sqrt{\sum_{k=1}^p (X_{ik} - \overline{X_i})^2} \sqrt{\sum_{k=1}^p (X_{jk} - \overline{X_j})^2}} \quad \text{단, } \overline{X_i} = \frac{1}{p} \sum_{k=1}^p X_{ik}$$

■ 군집방법과 종류

▶ 계층적 군집방법(Hierarchical Clustering Method)

i) 최단 연결법(Single Linkage Method)

두 군집 U 와 V 사이의 거리 d_{UV} 를 각 군집에 속하는 임의의 두 개체들 사이의 거리 중 최단 거리로 정의하여 가장 유사성이 큰 군집을 묶어 나가는 방법이다.

ii) 최장 연결법(Complete Linkage Method)

두 군집 U 와 V 사이의 거리 d_{UV} 를 각 군집에 속하는 임의의 두 개체들 사이의 거리 중 최장 거리로 정의하여 가장 유사성이 큰 군집을 묶어 나가는 방법이다.

iii) 평균 연결법(Average Linkage Method)

두 군집 U 와 V 사이의 거리 d_{UV} 를 각 군집에 속하는 모든 개체들의 평균거리로 정의하여 유사성이 큰 군집을 묶어 나가는 방법이다.

iv) 중심 연결법(Centroid Linkage Method)

군집 U 의 중심점과 V 의 중심점 사이의 거리 d_{UV} 를 두 군집사이의 거리로 정의하여 유사성이 큰 군집을 묶어 나가는 방법이다.

v) 중위수 연결법(Median Linkage Method)

두 군집 U 와 V 사이의 거리 d_{UV} 를 각 군집에 속하는 임의의 두 개체들 평균을 합하여 2로 나눈 값(군집의 크기를 고려하지 않은 단순평균)을 근간으로 정의하여 가장 유사성이 큰 군집을 묶어나가는 방법이다.

vi) Ward의 방법

두 군집 U 와 V 를 묶을 때 생기는 새로운 군집 Z 에 속해있는 개체들의 오차제곱합을 SSE_i 라고 하고 전체 군집의 오차제곱합을 $SSE = \sum_{k=1}^q SSE_i$

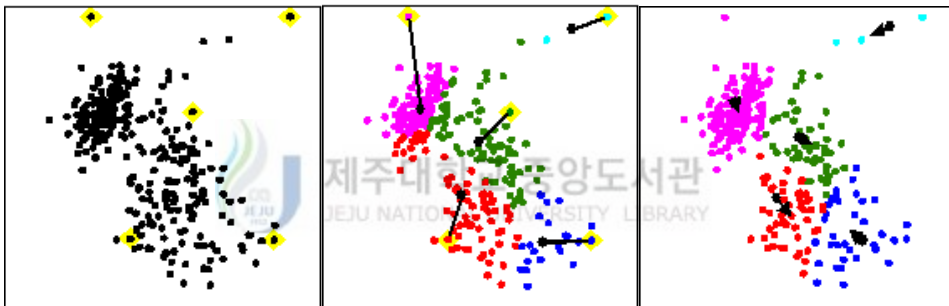
라고 할 때 새로운 군집으로 인하여 파생되는 SSE 의 증가량을 두 군집사이의 거리로 정의하여 가장 유사성이 큰 군집을 묶어나가는 방법이다.

▶ 비계층적 군집방법(NonHierarchical Clustering Method)

i) K-평균 군집법(K-Means Clustering)

각 개체를 가장 가까운 중심점에 할당하는 방법으로 일반적으로 다음 절차에 따라 군집을 형성

- K개의 초기 중심점을 선택한다.
- 각 개체를 가장 가까운 중심점을 갖는 군집으로 할당한 후 새로운 군집의 중심점을 계산한다.
- 각 개체의 할당에 변화가 없을 때까지 위의 단계를 반복하여 최종적으로 K개의 군집을 형성한다.



■ 군집분석의 장점

- ▶ 탐색적인 기법 : 대용량에 대한 탐색적인 기법이므로 사전적인 정보없이 의미있는 자료구조를 얻는다.
- ▶ 다양한 형태의 데이터에 적용가능 : 거의 모든 형태의 데이터에 적용가능
- ▶ 분석방법의 적용 용이성 : 변수들에 대한 역할정의를 필요가 없으므로 적용이 쉽다.

■ 군집분석의 단점

- ▶ 가중치와 거리정의 : 비유사성 거리정의와 가중치결정 어렵다.
 - ▶ 초기 군집수 설정 : K평균 군집분석의 경우 군집수 K가 적합치 않으면 결과가 나쁘다.
 - ▶ 결과해석의 어려움 : 사전에 주어진 목적이 없으므로 결과해석이 애매하다.
- 데이터마이닝에서 군집분석 방법은 기존의 통계학, 기계 학습, 패턴 인식에서 쓰

이던 방법에 부가적으로 데이터베이스 지향적인 제약 사항들(제한된 메모리 양, I/O 시간 최소화 등)을 첨가시킨 것으로써, 최근의 멀티미디어 데이터와 같이 혼합되고 다양한 다차원 데이터를 효율적으로 분류해 나가기 위한 방안으로 연구되고 있다.

(2) 결정트리(Decision Tree)

결정트리는 데이터마이닝의 분류작업에 주로 사용되는 기법으로, 과거에 수집된 데이터의 레코드들을 분석하여 이들 사이에 존재하는 패턴, 즉 부류별 특성을 속성의 조합으로 나타내고 분류모형을 트리의 형태로 만드는 것이다. 이렇게 만들어진 분류모형은 새로운 레코드를 분류하고 해당 부류의 값을 예측하는데 사용되며, 분석의 정확도보다는 분석과정의 설명이 필요한 경우에 더 유용하게 사용된다. 결정트리 분석을 수행하기 위한 다양한 방법들과 분리기준, 정지규칙, 가지치기 방법들이 제안되고, 이들을 어떻게 결합하느냐에 따라서 서로 다른 결정트리가 형성된다. 또한 정확하고 빠르게 결정트리를 형성하기 위해서 다양한 알고리즘이 연구되고 있고, 보다 개선된 알고리즘들이 계속 연구되어 발표되고 있다. 결정트리의 대표적인 알고리즘들은 다음과 같다.

▣ CHAID (Chi-squared Automatic interaction Detection) : 카이제곱-검정(이산형 목표변수) 또는 F-검정(연속형 목표변수)를 이용하여 다지분리(multiway split)를 수행하는 알고리즘으로 결과나무를 가지치기 할 필요가 없다. 여기서 다지분리란 부모마디에서 자식마디들이 생성될 때, 2개 이상의 분리가 일어나는 것을 허용함을 의미한다.

CHAID는 목표변수가 이산형일 때, Pearson의 카이제곱 통계량 또는 우도비 카이제곱 통계량(likelihood ratio Chi-square statistic)을 분리기준으로 사용한다. 여기서 목표변수가 순서형 또는 사전그룹화된 연속형인 경우에는 우도비 카이제곱 통계량이 사용된다.

카이제곱 통계량은 관측도수(frequency, f_{ij})로 이루어진 $r \times c$ 분할표(contingency table)로부터 계산된다. 분할표로부터, Pearson의 카이제곱 통계량은

$$\chi^2 = \sum_{i,j} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

과 같이 정의되고, 우도비 카이제곱 통계량은

$$\chi^2 = 2 \sum_{i,j} f_{ij} * \log e\left(\frac{f_{ij}}{e_{ij}}\right)$$

로 정의된다. 이 때 두 통계량의 자유도(degrees of freedom)는 $(r-1)(c-1)$ 로서 동일하다. 여기서 e_{ij} 는 분포의 동일성 또는 독립성의 가설 하에서 계산된 기대도수(expected frequency)를 말하며, 아래의 주어진 식

$$e_{ij} = \frac{f_{i.} * f_{.j}}{f_{..}}$$

와 같이 계산된다.

카이제곱 통계량이 자유도에 비해서 매우 작다는 것은, 예측변수의 각 범주에 따른 목표변수의 분포가 서로 동일하다는 것을 의미하며, 따라서 예측변수가 목표변수의 분류에 영향을 주지 않는다고 결론지을 수 있다. 자유도에 대한 카이제곱 통계량 값의 크고 작음은 p-값은 커지게 된다. 결국, 분리기준을 카이제곱 통계량 값으로 한다는 것은, p-값이 가장 작은 예측변수와 그 때의 최적분리에 의해서 자식마디를 형성시킨다는 것을 의미한다.

CHAID는 목표변수가 연속형인 경우에는 두 개 이상의 그룹에 대해서 평균치차를 검정하는 분산분석표(ANOVA Table: Analysis of Variance Table)의 F통계량을 분리기준으로 이용한다. F통계량의 자유도에 비해 매우 작다는 것은 예측변수의 각 범주에 따른 목표변수의 평균치 차가 존재하지 않다는 것을 의미하며, 따라서 예측변수가 목표변수의 예측에 영향을 주지 않는다고 결론지을 수 있다. 카이제곱 통계량과 마찬가지로 자유도에 대한 F통계량의 크고 작음은 p-값으로 표현될 수 있는데 F통계량이 자유도에 비해서 작으면 p-값은 커지게 된다.

CHAID에서는 계산된 F통계량의 p-값을 기준으로 명목형 목표변수인 경우와 유사하게 병합과 분리를 계속하여, p-값이 가장 작은 예측변수와 그 때의 최적분리에 의해서 자식마디가 형성된다.

▣ CART(Classification and Regression Trees) : 지니 지수(Gini Index) 또는 분산의 감소량을 이용하여 이지분리를 수행하는 알고리즘이다. 여기서 이지분리란 부모마디로부터 자식마디가 2개만 형성되게 한다는 것을 의미한다.

CART는 목표변수가 이산형일 경우 불순도를 측정하는 하나의 지수로서 지니 지수를 가장 감소시켜주는 예측변수와 그때의 최적 분리기준에 의해 자식마디가 형성된다.

$$\sum_{i=1}^r p(i)(1-p(i))$$

여기서 r 은 목표변수의 범주의 수를 나타내며 $P(i)$ 는 주어진 자료 중 i 범주에 분류될 확률을 나타낸다. 이의 최대값은 $P(i)$ 가 모두 같을 경우이다.

CART는 모든 가능한 분리자(splitter)를 조사하고, 그 중에서 가장 뛰어난 레코드를 찾아내는 것이고, 가지치기 할 때에는 최대한의 예측력을 가진 레코드를 찾기 위해서 가지치기를 한다.

■ C*.* : 범주화된 분리자를 다루고, 서브 트리를 평가할 때에 테스트 집합을 사용하지 않는다. CART의 가지치기와는 달리, 끝마디에서 에러율을 조사하여 최소한의 에러율을 얻기 위해서 가지치기를 한다.

이산형 목표변수의 분리기준으로 엔트로피지수(Entropy index)를 사용하는데 다항분포에서 우도비 검정통계량을 사용하는 것으로 이 지수가 가장 작은 예측변수와 그 때의 최적 분리기준에 의해 자식마디가 형성된다.

$$- \sum_{i=1}^r p(i) \log(p(i))$$

■ SPRINT(Scalable Parallelizable Induction of decision trees) : 메인 메모리 제한이 없어서 빠르고 방대한 양의 데이터 처리가 가능하며, 병렬화가 쉽다. 기존의 알고리즘들은 메인 메모리에 데이터를 저장하고 처리하였으나, SPRINT는 메인 메모리를 사용하고 부족한 공간은 디스크를 사용하여 유지한다(Shafer 등, 1996).

위의 알고리즘들을 바탕으로 결정트리의 장점을 살펴보면 다음과 같다. 분류나 예측의 근거를 알려주기 때문에 이해하기가 쉽다. 데이터를 구성하는 속성의 수가 불필요하게 많을 경우에도 모형 구축시 분류에 영향을 미치지 않는 속성들을 자동으로 제외시키기 때문에 데이터 선정이 용이하다. 또한, 연속형이나 명목형 데이터 값들을 기록된 그대로 처리할 수 있기 때문에 지식발견 프로세스 중 데이터의 변환 단계에서 소요되는 기간과 노력을 단축시킨다. 마지막으로, 어떠한 속성들이 각각의 부류값에 결정적인 영향을 주는가를 쉽게 파악할 수 있으며, 모형구축에 소요되는 시간이 짧다. 이에 반하여 결정트리의 단점은 나이나 소득 등과 같은 연속형 데이터를 처리하는 능력이 신경망이나 통계기법에 비해 떨어지며 그 결과 예측력도 감소한다는 것이다. 또한, 부류가 연속형 변수의 형태를 취하며, 이것을 예측하는

모형을 구축하는 것이 목적일 경우에는 적합하지 않으며, 모형을 구축하는데 사용되는 표본의 크기에 지나치게 민감하다. 결정트리는 의학 진단 및 연구, 고객 관리를 위한 작업, 날씨 예측, 사기 행위 발견 등 분류를 필요로 하는 여러 영역에서 효율적으로 적용되고 있다.

(3) 신경망 (Neural Networks)

신경망은 예측 모델을 만들어내는 매우 강력한 도구로서 데이터마이닝에 적용되는 중요한 테크닉 중 하나이다(William, 1999). 신경망에 대한 보편적인 정의는 각각 작은 로컬 메모리를 가진 여러 개의 간단한 프로세서들, 즉 유닛들로 이루어진 하나의 네트워크라는 것이다. 각 유닛들은 보통 여러 방법으로 인코딩된 숫자 데이터를 운반하는 의사소통 채널들에 의해 연결되어져 있고 자신의 로컬 데이터와 연결을 통해 받은 입력들만 조작한다.

신경망에서 쓰이는 학습 방법은 크게 교사학습과 비교사학습 학습으로 볼 수 있다.

교사학습 학습의 경우, 학습 단계에서 뉴럴 네트워크가 얼마나 잘 수행했는지 또는 올바르게 행동한 것이 무엇인지를 판단하고 말해주는 교사를 두고 있다. 교사학습 네트워크들의 트레이닝 방법으로 주로 사용되는 것은 네트워크의 결과값과 목표값, 즉 원하는 결과값들 간의 차이를 측정하는 것이다. 이를 흔히 에러율 이라고 부르며, 다음 트레이닝에서 이 에러율이 얼마나 줄었는지를 또는 늘었는지를 측정하면서 목표에 점차 접근해 나가는 것이다. 이 알고리즘들은 분류 모델에 적합하며 피드백 프로세스를 가지고 있다.

비교사학습 학습의 경우는 네트워크에 있는 데이터들을 보고 그 데이터 집합의 특성들이 어떤 것인지를 찾고 이 특성들의 결과에서의 반영에 대해 학습하는 것이다. 이러한 특성들이 정확하게 무엇인지, 즉 네트워크가 인식하기 위해 학습할 수 있는 것이 무엇인지는 특정 네트워크 모델과 학습방법에 의존한다. 비교사학습을 따르는 대표적인 알고리즘들에는 Kohonen self-organizing map(Kohonen, 1997), adaptive resonance theory 등이 있다. Self-organizing map은 입력 공간의 클러스터로의 위상적인 매핑을 제공해주는 경쟁적인 네트워크로 되어 있다. SOM(Self-Organizing Map)은 다양한 인간의 감각적인 표현들이 뉴런들 사이의 공간적인 관계에 대응하는 자극체들 사이의 공간적인 또는 다른 관계들과 같은 인

간두뇌에 신경학적으로 매핑되는 식의 방법으로 만들어진다. SOM에서는 뉴런들로 grid의 조직을 이룬다. SOM 트레이닝에서는 neighborhood 크기의 선택이 가장 중요한 트레이닝 파라미터가 된다. 이 알고리즘들은 군집분석에 적합하다.

신경망의 경우 결과에 대해 어떻게 계산되었는지를 설명하기란 어렵다. 마치 블랙박스과 같아서 사용자들에게 결과를 납득시키기 힘들다는 문제점을 낳고 있다. 이를 해결하기 위해 최근에는 하이브리드 학습으로 퍼지 로직이나 유전자 알고리즘을 이용하여 사용되는 확장된 개념의 신경망 기법들이 사용되고 있다. 신경망의 적용분야는 모델의 특성에 따라 아주 다양하지만, 공통된 장점은 입력 선택이나 트레이닝 없이 다양한 입력 데이터를 다룰 수 있고 훌륭한 예측 모델이라는 것이다.

신경망의 유연한 비선형 회귀 모델 그리고 분류 모델은 통계학 또는 마이닝 틀 등에 유용하게 쓰이고 있으며, 분류, 군집분석, 모델링, 타임시리즈 예측의 영역에 적용된다. 신경망 알고리즘의 예측 모델로 데이터베이스 마케팅, 고객 관계 유지, 사기행위 적발, 파산 예측 등 그 응용 분야가 매우 다양하다[Bigus, 1996]



2. 의료분야에서의 데이터마이닝 활용

정보화 기술의 발전에 따라 병원에서는 처방전달시스템, 의무기록전산화로 방대한 양의 데이터가 축적되어 있는데 이들 데이터 속에서 의학연구나 병원경영에 유용한 정보를 찾아내는 것을 의료분야의 데이터마이닝이라고 할 수 있다.

1) 의료분야에서의 데이터마이닝 적용분야

데이터마이닝은 통계학, 컴퓨터과학, 인공지능, 공학에서 출발하였으나 다양한 경영환경에서 경쟁력확보를 위한 어떤 조직에서도 활용이 가능하며, 의료분야의 몇 가지 적용분야를 살펴보면 다음과 같다(최국렬 등, 2001).

(1) 의학연구

현재 가장 활발하게 활용되고 있는 분야이며, 특히 게놈 프로젝트와 같이 인간의 유전자 구조를 밝히는데 있어서는 데이터마이닝의 사용이 필수적인 요소로 되고

있고 그 사례도 다양한 형태로 나타나고 있다. 이 밖에도 알츠하이머 질환, 간염연구, 암 관리, 사고위험연구 등에서도 데이터마이닝이 효과적으로 활용되고 있다.

(2) 의료진단

현대 의료의 특성은 각종검사결과에 근거하여 의사들이 환자의 진단명을 정확하게 진단한다는 것이다. 즉, 데이터에 근거하여 환자의 진단을 내림에 따라 각종 검사결과를 기반으로 하여 데이터마이닝이 자동으로 진단명을 예측하게 함으로 의료진들이 보다 쉬우면서도, 체계적으로 환자에 대해 정확한 진단을 할 수 있게 되었다. 데이터마이닝을 이용하여 진단을 내리는 방법에 대한 연구로는 남성들의 불임 진단이나 EEG를 이용하여 진단명을 내리는 것에 대한 연구가 있다.

(3) 의료의 서비스 관리

현재 보건의료계가 당면하고 있는 가장 중요한 과제는 보다 적은 비용으로 양질의 의료서비스를 제공해야 한다는 것이다. 이를 위해서 국내에서도 병원의 감염관리 및 진료행위프로토콜 개발에 데이터마이닝이 활용되고 있다. 선진국에서도 합병증에 의해 발생하는 감염병관리, 사례관리자를 위한 훈련, 수술환자의 의료서비스 관리에도 활용되고 있다. 특히, 정보기술을 활용하는 의료의 서비스관리, 예컨대 증거지향의 의료(EBM)이나 진료행위 프로토콜(Critical Pathway) 분야에 데이터마이닝이 활용되기 시작하고 있다.

(4) 병원경영

정보화 사회에 병원도 지식경영을 할 필요성이 높다. 즉, 환자에 대한 마케팅, 재정투자, 시설배치 등 많은 분야에서 데이터마이닝 기법이 활용되고 있으며 병원경영전략수립에 있어서는 데이터마이닝의 활용이 매우 필요하다. 또한 현재 각광받고 있는 CRM(Customer Relationship Management)기법을 도입하기 위해서는 데이터마이닝이 가장 기본적인 요소기술이 되고 있다.

이에 따른 데이터마이닝의 이용형태를 보면 다음과 같다.

■ 환자 이탈방지 시스템 구축

OLAP와 간단한 통계분석을 이용하여 최근 몇 년간의 병원고객의 이탈 추이와 지역별, 시간별, 인구통계학적 등과 같은 변수별로의 이탈 현황을 살펴보는 탐색단계로 통해 향후 데이터마이닝을 위한 기본정보를 취득한다. 그리고 이미 이탈한 고객과 현재 이탈하지 않은 고객을 샘플링하여 각각의 특성을 파악하는 과정을 거쳐

나온 결과를 바탕으로 이탈하는 고객들의 패턴을 찾기 위한 데이터 마이닝 작업을 수행한다. 이탈분석을 위하여 필요로 하는 데이터는 고객에 대한 다양한 변수들을 갖고 있는 데이터이며 오류값이 정제되어 있어야 한다. 이와 같이 정제된 데이터로부터 데이터마이닝을 DB에 생성한다. 모델링을 위해 중요한 단계는 이탈 확률을 찾아내기 위하여 어떠한 변수들이 설명력을 갖는가에 대한 검증단계이다. 이를 위해서는 주로 Chi-square, 분산분석, 범주형 자료 등의 통계기법을 이용하여 유의성 있는 변수들을 찾아내어 이러한 변수들을 이용, 정밀도 높은 예측치를 도출해 내는 모델링 단계를 거친다. 이와 같이 데이터마이닝을 이용하여 도출된 모델을 전체 고객 데이터에 반영하여 이탈가능 확률이 점수화되며, 이러한 데이터베이스를 바탕으로 마케팅 계획이 수립되고 실행을 거쳐 다시 피드백 되는 과정을 거치게 된다.

■ 특정환자 타겟 마케팅(Target Marketing)

기존의 병원의 경영현실에서 보면 병원은 그저 방문하는 환자만을 고객으로 여기고 다른 병원의 방문환자나 잠재적 일반 고객에 대하여서는 관심을 보이지 않는 수동적인 자세를 보여 왔다. 그러나 앞으로 병원은 적극적인 자세로 보다 과학적인 접근방법을 통한 업무효율 및 고객 확보의 극대화에 대한 요구가 발생하였으며 이에 대한 조희시스템이 필요하게 될 것이다. 그 대표적인 예가 바로 특정 고객에 대한 타겟 마케팅 전략이다.

이는 우선 고객의 패턴을 모형화 하여 이를 실제 고객 마케팅업무에 반영하고 고객이 선호하는 서비스를 면밀히 조사하여 표적 고객집단을 선정하고 고객의 계층별 선호도를 이용하여 좀더 효과적이고 전략적인 시장세분화 전략을 수립하여야 한다. 또한 주제별 데이터 마트(Data Mart)를 구성하여 이를 이용한 OLAP시스템을 구축함으로써 지역별, 진료과목별 현황이나 심지어 담당의사나 환자의 정보 등 다양한 정보를 자유롭게 조회할 수 있는 시스템의 구축이 이루어져야 한다. 이와 같은 과정을 통해 얻어진 데이터를 이용하여 수량화, 최적척도변환 등을 이용한 대응분석 등을 통해 고객의 선호를 파악하고 우선순위를 설정하여 우선순위가 높은 고객에 대해서는 각 개인이 선호하는 서비스(특정 질환 예방검사, 예방 접종 등)를 적극 홍보함으로써 고객의 병원이용률을 유지시키고 우선순위가 낮은 고객의 유지 및 신수요의 창출을 위해서는 좀더 차별화된 서비스를 개발하기 위한 투자 등 보다 다양한 고객 서비스 전략을 구상해야 한다. 또한 자병원에 국한된 정보뿐만 아니라

다른 병원의 정보가 서로 연계되는 경우 이러한 전략은 보다 큰 효과를 기대할 수 있는 것이다. 즉, 이를 통하여 자병원의 특성에 맞는 환자를 유치할 경우 환자의 만족 증대는 물론 더 많은 수익을 가져다 줄 수 있기 때문에 고객과 잠재고객의 파악은 각 병원으로 하여금 보다 적극적인 마케팅 전략을 수립 및 실행시킬 수 있도록 할 것이다.

2) 병원 데이터베이스에서의 데이터마이닝을 활용한 관련연구

과거 환자만족도 분석이나 환자의 병원선택요인 규명에 초점을 두어 설문조사 수준의 병원마케팅의 연구를 벗어나 병원의 관계마케팅이나 데이터베이스 마케팅에 대한 실증적 연구로 부산지역의 I대학병원에서 의무기록정보를 기초로 하여 구축된 퇴원환자 데이터베이스 및 진료비 자료를 이용하여 병원이용빈도와 환자수익성을 유형화한 후, 그에 따라 시장세분화 및 목표환자군 선정을 통하여 관계마케팅 등 적절한 전략을 개발하는 과정을 분석하여 활용할 수 있도록 하였다(최길림, 2001).

또한 Y대학병원의 경우 병원의 의사결정지원을 위한 데이터웨어하우스 구축과 지식경영시스템 개발의 일환으로 병원 진료 및 경영지식을 추적, 조회 및 효율적 활용을 위한 데이터웨어하우스(Data Warehouse)를 구축하고, 데이터마이닝(Data Mining) 기법을 활용하여 고혈압 관리, 의료보험 사전심사, CQI(Continuous Quality Improvement)를 위한 의사결정지원시스템(Decision Support System)을 개발하였으며 이를 실제 임상에서 시범 운영하고 평가하기 위한 연구를 수행 중이다.

III. 환자특성분석을 위한 데이터마이닝 시스템

시스템 설계를 위하여 보건정보학 관련 논문에서 발표된 자료를 기반으로 병원 데이터베이스 시스템에서 환자패턴 분석을 위한 데이터마이닝 기법을 적용하는 구체적인 연구모형을 제시하며, 데이터 전처리 과정과 특성분석에 대한 목적설정과 항목설정 및 모형화 과정을 수행한다.

본 논문에서 데이터마이닝 적용의 전체 구현과정은 아래 Fig. 2와 같다.

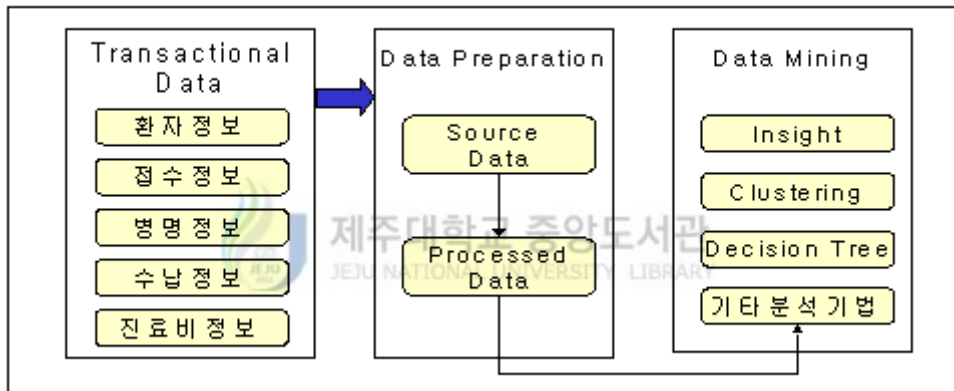


Fig. 2 데이터마이닝 적용의 전체 구현 과정

전체 시나리오 단계로는 각각의 트랜잭션 단위의 운영 데이터베이스 테이블에서 환자정보, 입원 및 외래 접수정보, 진료결과에 따른 진단병명 정보, 특정 환자들에 대한 진료비 감면 정보가 기록된 수납정보 및 처방내역 진료비 정보를 가지고 있는 분산된 테이블에서 필요한 항목을 추출하여 데이터를 수집하고 환자특성 분석을 위한 목적에 적절하게 필터 처리 등을 하여 분석대상의 데이터 집합(Data Set)을 만든다. 이후 SAS의 기본 통계정보 분석 및 E-Miner 데이터마이닝 툴이 지원하는 여러 기법과 알고리즘 및 통계분석 노드 등을 적용시켜 환자 분류 모형을 생성하고 규칙을 얻으며 다양한 분석 자료를 생성하도록 한다.

1. 연구모형

본 연구는 병원데이터베이스로부터 의무기록정보를 기초로 하여 환자의 의료기관 이용형태를 분석하고자 지식공학의 개념을 적용하고 있으며 연구의 흐름은 Fig. 3과 같다.

데이터마이닝을 위한 데이터 전처리 과정을 거친 후 구체적인 데이터마이닝 기법을 이용한 연구 형태는 다음과 같다.

첫째, 분석대상 환자들의 재입원패턴을 모형화하고 결정요인을 분석한다.

둘째, 재입원군을 세분화하고 유형별 관련 특성들을 분석한다.

셋째, 병원이용빈도를 기준으로 하여 유형화하고, 유형화된 환자군집들의 일반적인 인구사회학적 특성 및 진료관련 특성을 분석한다.

넷째, 병원이용빈도에 영향을 주는 특성변수들의 상대적 중요도를 평가한다.

2. 데이터 전처리 과정



본 연구에서 사용된 데이터는 제주지역 A종합병원의 입원환자에 대한 데이터로써, 2001년 3월 1일부터 2002년 2월 28일까지의 기간동안 입원진료 후 자료 추출시점인 2002년 3월을 기준으로 퇴원한 환자 8,300명의 자료를 특성 분석 대상으로 이용하였다.

일반적으로 최근에 병원정보시스템이 구축된 병원의 경우 환자 퇴원시 기록된 퇴원요약 DB를 이용하여 병원진료통계 및 의학연구의 목적에 이용되어지고 있지만, 분석대상병원의 경우는 1994년부터 운영되어온 RDBMS시스템으로 환자 특성 분석의 목적에 따른 자료 추출을 위해 다음과 같은 전처리 과정을 수행한다.

1) 데이터 준비

대상병원의 OCS(Order Communication System) DB에서 환자 분류에 필요한

데이터를 선정하여 각각의 업무별로 구성되어 있는 Fig. 4와 같은 환자 특성 분석과 관련된 테이블들의 주요 항목들의 연관관계를 살펴보고 관계를 결합하여 기본 데이터마트 구성 개념의 분석용 자료 추출을 위한 준비를 하였다. 자료 분석 결과 일관성이 없이 기록된 자료의 정제작업을 거쳤으며, 분석을 위한 missing value 처리와 데이터 변환작업 단계도 수행하였다.

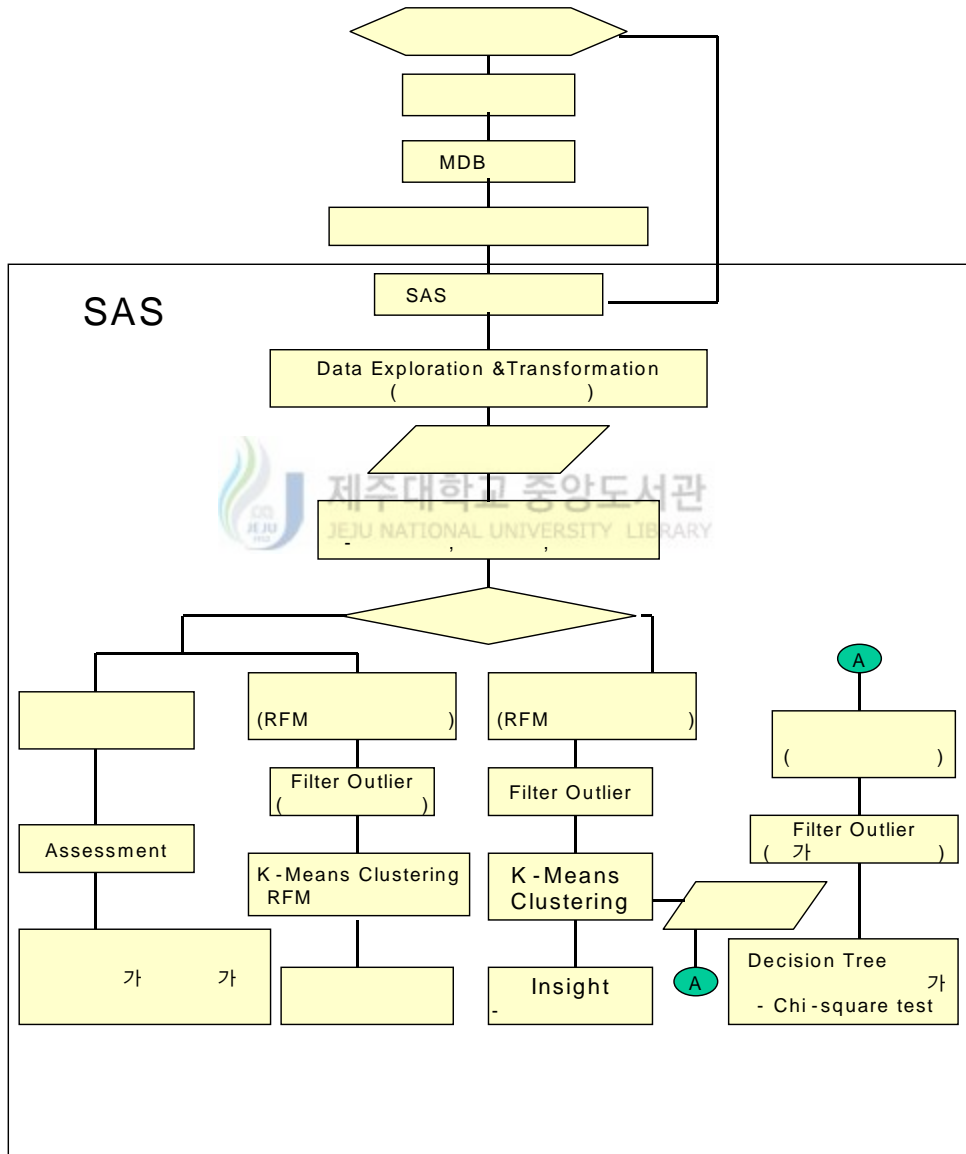


Fig. 3 연구모형

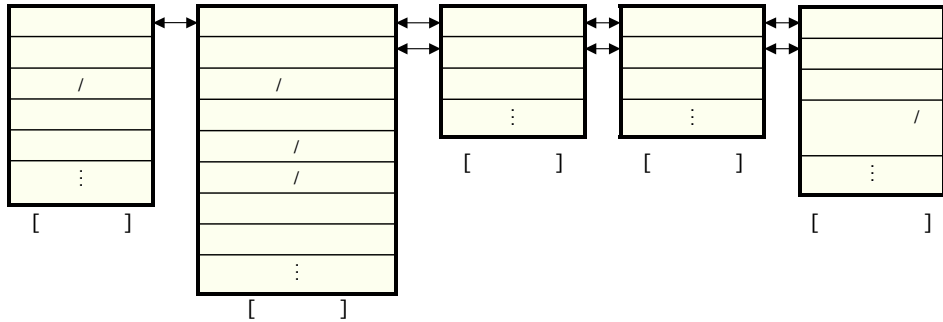


Fig. 4 OCS 데이터베이스의 관련 테이블의 항목간 연관관계

2) 데이터 수집

업무별로 분산되어 있는 운영시스템에서 환자 분류를 위한 소스데이터의 수집은 대상병원에서 사용 중인 업무프로그램 개발 툴인 Esprit를 이용하여 구현하였으며, 사용 중인 DB가 약 9년간 운영되어 자료의 양이 방대할 뿐만 아니라 업무적 특성 상 서버와 분리가 요구되어져 임시 텍스트 파일을 만들어 ftp를 이용해 PC의 Excel 파일로 변환하는 단계를 거쳤다. 이 가운데서 missing 값이 너무 많아 분석에 적절하지 않은 것 등은 정제작업을 거쳤으며, 분석기간 내에 퇴원환자의 정보들이 기록된 관련 테이블을 결합하여 입원자료 10,372건과 분석 기간동안 입원했던 환자들이 외래 진료를 받은 접수정보 및 진료비 관련 정보가 기록된 테이블에서 8,300명의 자료를 수집하여 Fig. 5와 같이 두 개의 Excel 파일을 생성하였다.

The screenshot shows two Excel spreadsheets. The first, '입원자료', has columns A through H with data including patient numbers, admission numbers, ward names, room numbers, medical history, age, sex, and hospital affiliation. The second, '외래자료', has columns A through E with data including patient numbers, admission numbers, ward names, room numbers, and medical history. Both tables show a list of patient records with various numerical and categorical values.

Fig. 5 수집된 Excel 자료 파일

노드를 이용하여 이상치를 제거하거나 범위를 재정의하는 과정 등을 수행하여 데이터 집합을 적절히 변환하였다.

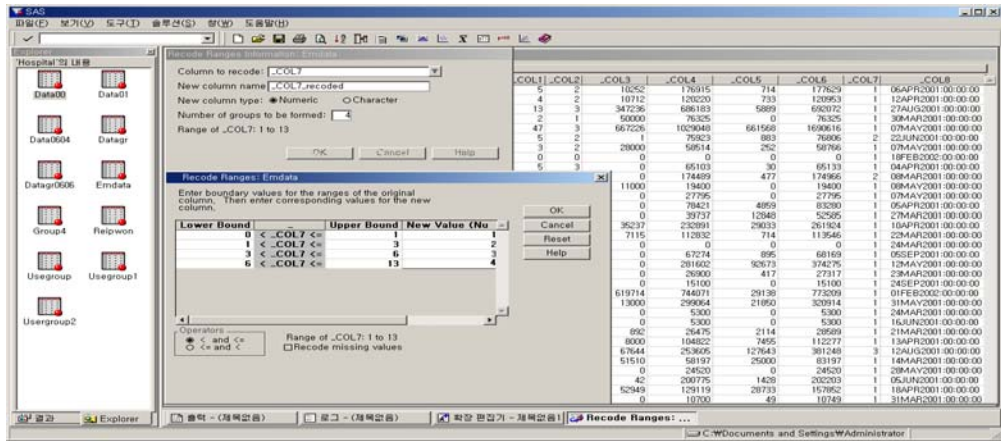


Fig. 7 SAS 데이터분석기능을 이용한 데이터 변환

분석에 이용되어질 데이터 집합의 주요 레코드 속성은 Table. 1과 같다.

Table. 1 환자특성분석에 이용될 데이터의 주요 레코드 속성

Variable Label	Type	Measurement	비고
병록번호	id	nominal	환자 ID
성별	char	binary	
나이	num	interval	1~100세까지 분포
우편번호	num	interval	우편번호코드의 앞3자리로 지역구분
환자유형	char	nominal	
과코드	char	nominal	
상병코드	char	nominal	환자의 주상병 기준
입원경로코드	char	nominal	
수술횟수	num	interval	
연령계층	num	nominal	9단계로 구분 (복지부 통계와 비교)
거주지역	num	nominal	제주시, 북군, 남군, 서귀포시, 기타지역
진료자격	num	nominal	일반, 보험, 산재, 자보, 보호, 기타
진료과	char	nominal	
질병군	num	nominal	KCD-3 대분류에 의해 21개로 분류
수술여부	char	binary	수술횟수를 기준한 무, 유
입원경로	num	nominal	1.응급실, 2.외래로 구분
총입원횟수	num	interval	
총외래방문횟수	num	interval	
총재원일수	num	interval	
퇴원후이용간격	num	interval	최종외래접수일 - 최종퇴원일 (월)
외래이용간격	num	interval	분석기준일(2002.3.28)-최종외래접수일(월)

3. 목적설정, 항목설정 및 모형화를 위한 데이터마이닝 기법

환자의 일반적 특성, 진료특성, 임상특성과 진료결과 변수를 분석에 사용하였으며 이러한 변수들을 체계화한 결과는 Fig. 8에 제시되어 있다. 진료결과 변수는 DB 마케팅에서 고객 구매 패턴의 세분화 기준으로 사용되고 있는 RFM(Recency, Frequency, Monetary)개념을 이용하였는데 병원이용간격은 구매의 최근성을, 총입원횟수와 총외래방문횟수는 구매의 빈도를, 총재원일수는 병원자원사용정도를 나타내는 지표로 삼았다.

SAS의 기술통계 중 빈도분석을 이용하여 환자의 특성들을 살펴보고, 연속형 반응변수의 경우 분산분석의 F -분포를, 범주형 자료의 경우 테이블 분석의 카이제곱 통계량(χ^2)을 이용하여 특성간 통계적 유의성 검정을 한다.

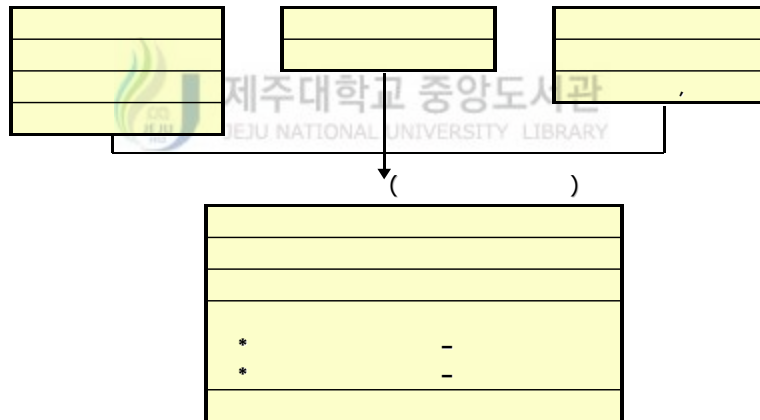


Fig. 8 환자특성분석에 사용될 주요변수

이를 토대로 재입원패턴과 병원이용빈도를 기준한 분류 목적을 설정하여 진행하였으며 다음은 각각의 목적과 이에 따른 변수의 설정 항목이다.

1) 재입원유무 결정요인분석

(1) 목적 설정

재입원패턴의 모형화와 결정요인 분석

(2) 항목 설정

목표변수
재입원유무
입력변수
성별, 나이, 진료자격, 거주지역, 입원경로, 수술여부, 진료과, 질병군

전체 재입원패턴을 모형화하고 그 결정요인을 분석하기 위해 Fig. 9와 같이 결정 트리, 회귀분석, 신경망기법, 대화식 결정트리 분석기법을 이용한 모형평가 (Assessment)를 하였으며, Root ASE(Asymptotic Standard Error) 값이 작을수록 오분류율이 작을 수록 좋은 검정 모형으로, 평가결과 결정트리와 대화식 결정트리 분석이 일반화가 용이하게 나타났다.

결정트리의 경우, Pearson Chi-square값에 따라 부모마디에서 자식마디로 분리해나가는 방법으로, 유의수준을 0.05로 지정한 카이제곱검정으로 트리노드의 분리 기준을 결정하였으며, E-Miner의 경우 현장업무에 대한 경험을 바탕으로 한 좀 더 현실성 있는 모형 구축을 위한 대화식(Interactive) 모드를 지원하고 있어 질병군을 노드분리의 우선 규칙으로 지정한 대화식 결정트리를 이용하였다. 이는 보건학 관련 논문 및 실제 분석기간이 장기간일 경우 우선시 되어지는 규칙으로 판단되었기 때문이다.

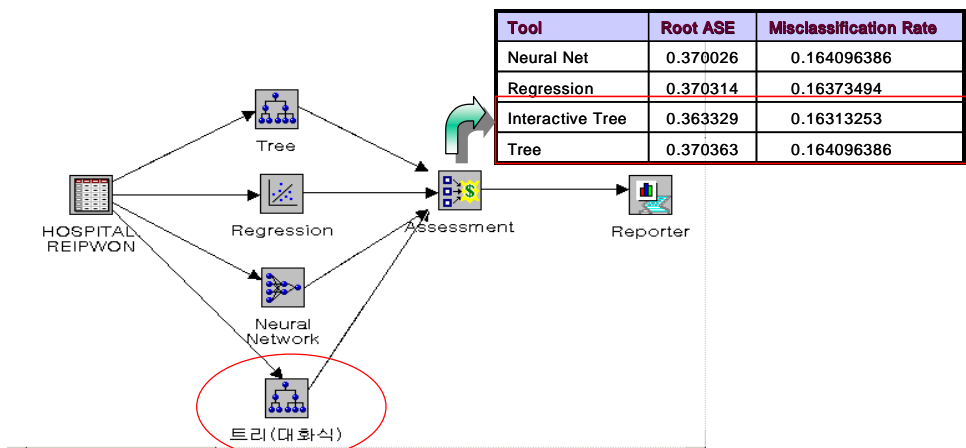


Fig. 9 재입원유무 모형평가 및 평가결과

2) 재입원패턴에 따른 군집별 특성분석

(1) 목적 설정

재입원패턴에 따른 군집별 특성분석

(2) 항목 설정

출력변수 (입원환자의 재입원패턴)	
군집1	단기입원환자군 (내과, 소아과 위주의 호흡기계 질환)
군집2	중기입원환자군 (내과, 외과 위주의 순환기 및 사고)
군집3	장기입원환자군
입력변수 (항목)	
총입원횟수	분석기간중 대상병원에 입원한 횟수
총재원일수	입원한 일수의 합계
평균재원일	1회입원 평균재원일
병원이용간격	2002년 3월 기준 환자의 최종외래내원일(외래진료예약일포함)과 최종퇴원일과의 간격

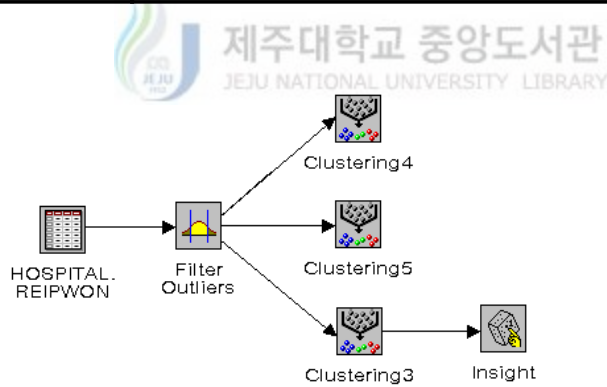


그림. 10 재입원패턴에 따른 군집분석 흐름도

결정트리의 모형화 결과만으로는 환자의 재입원에 대한 설명력이 미미하여 Fig. 10의 분석흐름도에 따라 Filter-Outlitter 노드를 이용하여 재입원한 환자만을 대상으로 군집화시켜 분류된 군집과 재입원유무를 결정짓는 특성변수들간의 관계를 살펴본다. K-Means Clustering 기법을 이용하여 여러 군집수 K를 이용한 테스트 결과, 군집수 3의 경우 가장 특성화된 군집들로 구성됨을 알 수 있었다.

3) 병원이용빈도에 따른 군집분석

(1) 목적 설정

병원이용빈도 특성에 따른 환자분류

(2) 항목 선정

출력변수 (입원환자의 병원이용빈도)	
군집1	입원다이용, 외래저이용 환자군 (입원진료 중심적)
군집2	입원/외래 저이용환자군. [일반적인 입원환자 이용형태]
군집3	입원/외래 다이용환자군. [장기입원환자군]
군집4	입원저이용, 외래다이용 환자군 (외래진료 중심적)
입력변수 (항목)	
총입원횟수	분석기간중 대상병원에 입원한 횟수
외래방문횟수	외래진료과를 방문한 횟수
총재원일수	입원한 일수의 합계
병원이용간격	분석기간중 최종내원일부터 2002년 3월까지의 기간(월) * 최종내원일은 예약진료 포함

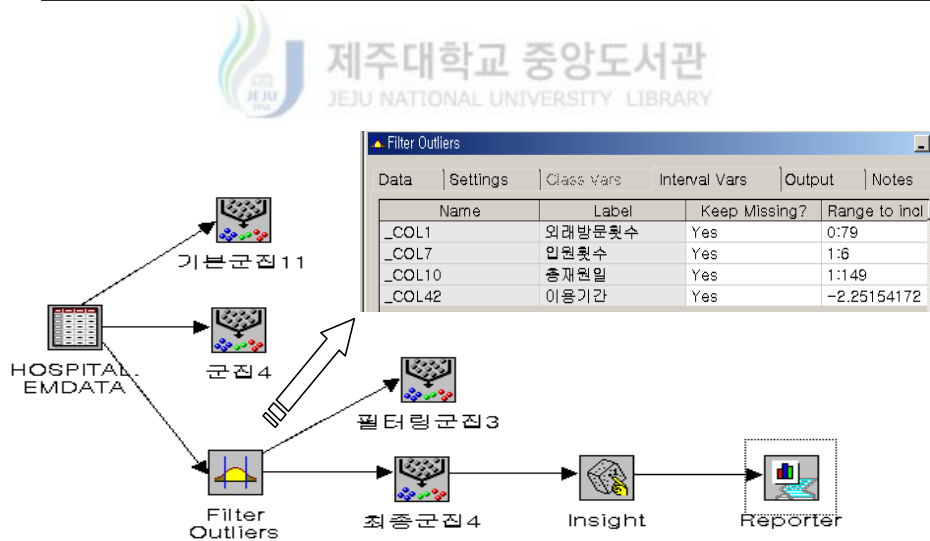


Fig. 11 병원이용빈도에 따른 군집분석 흐름도

Name	Importance	Measurement	Type	Label
_COL1	0.8095533704	interval	num	외래방문횟수
_COL10	1	interval	num	총재원일
_COL42	0	interval	num	이용기간
_COL7	0	interval	num	입원횟수

Fig. 12 병원이용빈도에 따른 군집결정에 이용된 변수 특성 및 중요도

진료결과 변수 중 병원이용빈도를 나타내는 변수인 총입원횟수, 외래방문횟수, 총재원일, 병원이용간격을 사용하여 Fig. 11의 과정을 거쳐 군집분석을 수행하여 병원이용형태를 살펴본다. 여러 군집수를 이용한 테스트 결과 군집 4개가 가장 특성화된 군집들로 구성됨을 알 수 있었으며, 군집결정에 이용된 변수의 특성 및 중요도는 Fig. 12에서 총재원일수와 총외래방문횟수임을 알 수 있다.

또한 Insight 노드를 이용하여 병원이용빈도를 기준으로 분류된 환자집단들의 일반적 특성 및 진료 특성 분석 결과를 볼 수 있다.

4) 병원이용빈도 군집관련 특성변수의 상대적 중요도 평가

앞에서 분류된 4개의 군집 중 Filter-Outlier 노드에 의해 군집2와 3을 제외한 군집1(입원다이용, 외래저이용 환자군)과 군집4(입원저이용, 외래다이용 환자군)만을 대상으로 군집결정에 영향을 주는 특성변수의 상대적 중요도를 Fig. 13의 분석흐름도를 따라 평가한다.

결정트리 알고리즘 중 분리기준을 지니지수(Gini Index)를 이용한 Gini reduction을 사용했으며 노드 깊이는 3으로 설정하였다.

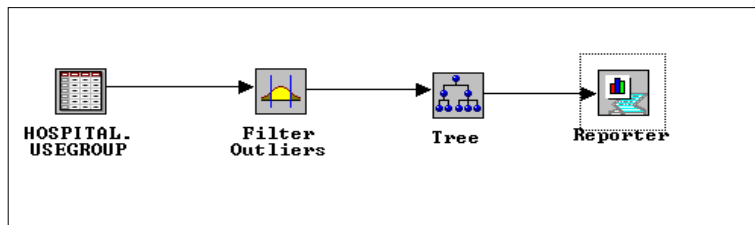


Fig. 13 군집결정을 위한 특성변수의 중요도 평가 흐름도

IV. 결과 및 고찰

1. 시스템 환경과 툴

연구에 사용된 시스템 환경과 툴은 다음과 같다.

분석자료 추출을 위한 병원 데이터베이스관리 서버시스템은 운영체제로는 SCO Unixware 7을 , RDBMS는 ORACLE 8.0.5로 운영되며, 데이터마이닝 적용을 위한 관련 데이터를 추출하기 위해 프로그램 개발 툴인 ESPrit(Embedded SQL PRogram Integrated Tool) 2.1을 사용하였다.

데이터마이닝 기법을 적용하여 패턴추출을 위한 시스템 환경은 Windows98 /Windows2000 시스템에서 사용하였으며, 분석테이블을 만들기 위해 PC용 DB프로그램인 Access와 통계패키지인 SAS, 데이터마이닝 툴인 SAS사의 Enterprise Miner 4.0을 이용하였다.



2. 데이터마이닝 수행 및 결과 해석

1) 입원환자의 특성분석 결과

(1) 분석대상자의 일반적 특성

우선 분석대상자의 연령계층, 성별, 진료자격 등의 분석은 전국의 의료기관을 전수조사를 했던 자료와 비교로 지역적 특성 및 분석대상병원의 일반적 특성을 알아볼 수 있었다.

퇴원환자의 남/여 성비는 1999년의 환자조사와 분석대상병원 모두 0.94로서 여자환자가 많은 경향을 보였다. 연령계층별 남/여 성비는 전반적으로 유사한 분포를 보이나, 35~44세 연령계층의 경우 전국 환자조사의 남/여 구성비 1.24에 비해 분석대상병원은 0.88로 여자가 많은 분포를 보였다. 연령계층별 퇴원환자수의 분포는

0~4세가 18.9%로 가장 많고, 25~34세 14.4%, 65~74세 11.9% 순으로 나타났다. 성별 분포로 보면, 남자의 경우 0~4세가 23%를 차지하고 있으며, 여자의 경우는 25~34세가 19.0%로 높게 나타났다. 4세 이하의 경우 전국 평균 7.9%의 약 2.5배인 18.9%를 차지하며, 65세 이상의 경우도 전국 조사의 14.6%보다 약 1.5배인 21.3%로 나타났다(부록 Table. A 참조).

퇴원환자의 진료자격별 분포는 의료보험 84.7%, 의료보호 9.63%, 자동차보험이 3.07% 순이었으며, 보건복지부에서 조사했던 1999년도의 자료와 부산지역 1개 대학병원의 특성분석 자료와 비교 결과, 의료보호 환자가 차지하는 비중이 높게 나타났다. 자동차보험의 경우 전국 비율보다 아주 낮게 나타났다(부록 Table. B 참조).

진료과별 이용 분포를 보면 소아과(24.1%), 내과(20.86%), 산부인과(12.57%), 일반외과(10.86%), 신경과(7.14%) 순이었으며, 소아과와 신경과의 과점유율이 전국 평균에 비해 상대적으로 높게 나타났다(부록 Table. C 참조).

주진단 질병군별 분포를 보면 호흡기계 질환(18.9%), 손상, 중독 및 외인에 의한 특정 기타 결과(10.96%), 임신, 분만 및 산욕의 합병증(8.64%), 특정 감염성 및 기생충성 질환(7.98%) 순이었다. 전국 10대 퇴원환자의 질병분포 및 부산지역 1개 병원의 질병분포와 비교했을 때 호흡기계 질환(18.9%)의 경우 약 3배 정도 많게 나타났다(부록 Table. D 참조).

(2) 제특성에 따른 재입원유무, 총입원횟수, 총재원일수

분석대상병원의 분석기간 1년 동안 입원환자들의 총입원횟수 평균은 1.25회, 총재원일수는 평균 10.5일, 표준편차가 18.49일 이었으며 그 변이가 큼을 알 수 있다. 재입원을 한 경우는 1,362명으로 전체의 16.4%였고 총재원일수가 181일 이상인 환자도 18명이나 되었다(부록 Table. E 참조).

입원횟수별로 1회입원 평균재원일수를 살펴보면 입원횟수가 6회까지는 평균재원일수도 증가함을 알 수 있으나 총입원횟수가 7회 이상일 때는 오히려 감소함을 볼 수 있다(부록 Table. F 참조).

▣ 일반적특성에 따른 재입원유무 및 총입원횟수, 총재원일수 평균비교

성별에 따른 재입원유무, 총입원횟수, 총재원일수는 남자가 여자에 비해 높게 나타났다. 연령에 따라서는 연령이 높을수록 총재원일수의 평균이 높아짐을 알 수 있

으며, 65~74세와 75세 이상에서 재입원유무(23.61%, 23.04%), 입원횟수평균(1.42회, 1.34회), 총재원일평균(16.22일, 16.24일)이 타 연령군에 비해 높게 나타났으며 통계적으로도 유의한 차이를 보였다. 진료자격별로 보면 산재(28.57%)와 의료보호(27.07%)인 경우 재입원율이 높았고, 입원횟수와 총재원일수도 차이가 났다. 거주지역별로는 남제주군(19.10%)과 북제주군(18.03%) 지역이 타지역에 비해 재입원율이 높게 나타났으며 재원일 평균도 높음을 볼 수 있으며 통계적으로도 유의한 차이를 보였다(부록 Table. G 참조).

▣ 진료특성에 따른 재입원유무 및 총입원횟수, 총재원일수 평균비교

입원경로별로는 외래를 통한 계획된 입원의 재입원율이 높음을 알 수 있다.

진료과에 따른 재입원율은 빈도수가 적은 가정의학과(22.22%)를 제외하면, 내과(21.72%), 안과(18.83%), 소아과(17.6%), 정형외과(16.9%), 일반외과(15.65%)순이며, 총입원횟수 역시 1.39회로 내과가 가장 높았으며, 총재원일평균은 정형외과(23.98일), 신경외과(21.61일)등 외과분야가 높게 나타났다(부록 Table. H 참조).

▣ 질병에 따른 재입원유무 및 총입원횟수, 총재원일수 평균비교

질병군에 따라서는 정신 및 행동장애(28.38%), 신경계(27.66%), 혈액 및 조혈기관 질환(26.25%) 순이었다. 총입원횟수의 경우 내분비, 영양 및 대사질환(1.74회), 혈액 및 조혈기관 질환(1.51회), 정신 및 행동장애(1.47회), 신경계 질환(1.4회)순으로 높게 나타났으며, 총재원일평균은 호흡기계 질환이 17.34일로 가장 길었고 질병이환 및 사망의 원인(16.6일), 정신 및 행동장애(16.55일) 순이었다(부록 Table. I 참조).

2) 재입원유무 결정요인 분석 결과

(1) Decision Tree를 이용한 재입원 가능성 결정 모형

Fig. 14에서 보는 바와 같이 결정트리분석을 이용하여 재입원유무의 결정요인을 분류한 결과 나이, 진료자격, 수술여부, 진료과에 따라 분류군이 결정됨을 알 수 있다. 모형 결과를 보면 61세 이상이면서 외과·신경외과·성형외과·내과·가정의학과·치과·안과·비뇨기과 등에서 수술과 관련이 있는 진료를 받는 경우 대상병

원의 평균재입원율의 2배정도인 35.8%로 재입원율이 높음을 볼 수 있다.

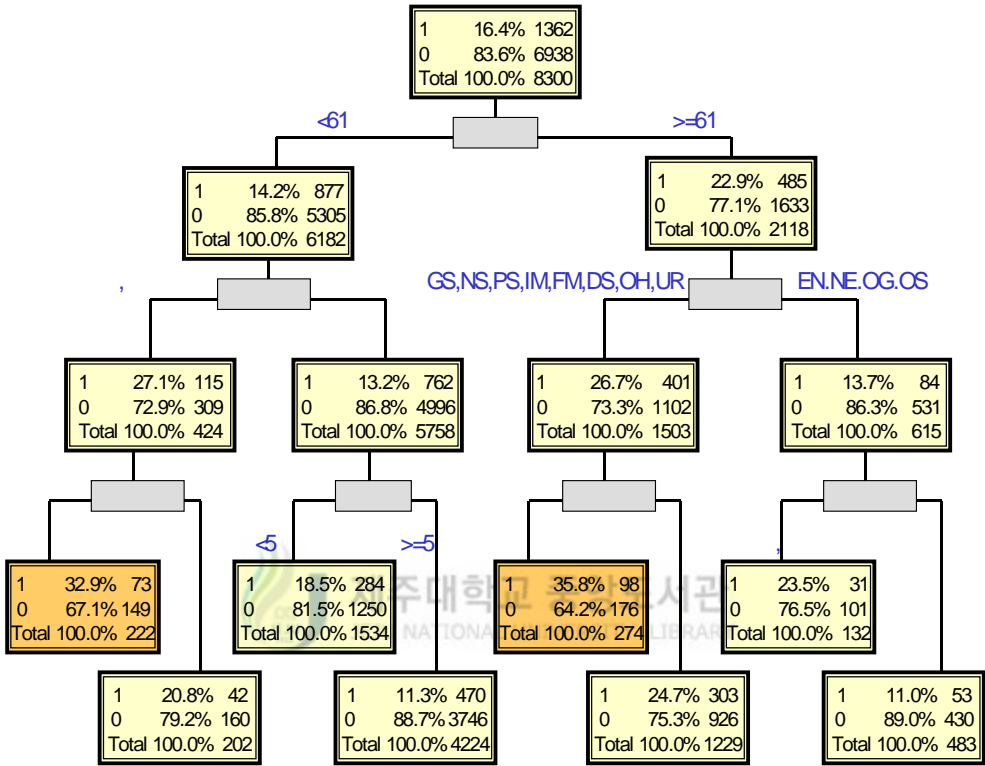


Fig. 14 결정트리를 이용한 재입원 가능성 결정모형

(2) 대화식 Decision Tree를 이용한 재입원 가능성 결정 모형

Fig. 15는 E-Miner가 제공하는 현장업무에 대한 경험을 반영할 수 있는 현실성 모형을 만들 수 있는 대화식(Interactive) 결정트리를 이용하여 재입원유무를 목적으로 하여 재입원 가능성을 분석한 결과이다. 질병군을 우선 규칙으로 지정하여 분석한 결과 다음 노드의 결정은 Fig. 14와 유사하게 나이와 진료자격 등에 따라 분류군이 결정되었다.

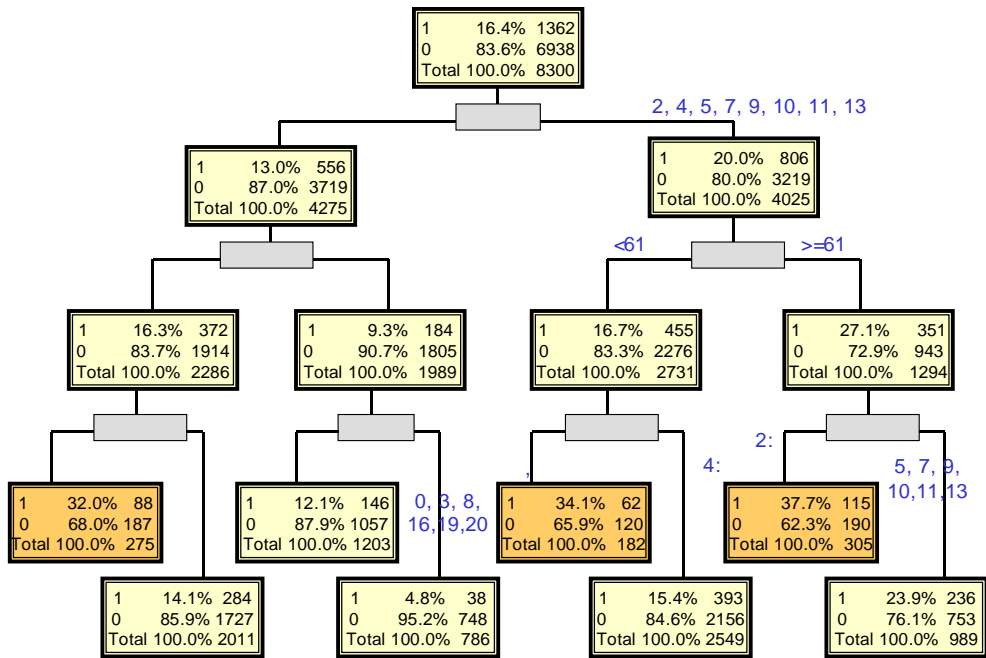


Fig. 15 대화식 결정트리를 이용한 재입원 가능성 결정모형

3) 재입원패턴에 따른 군집별 특성분석 결과

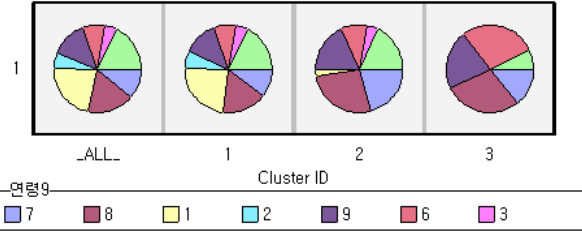
Table. 2에서 보면 재입원군이 3개 군집으로 분류되는데 먼저 빈도수가 가장 많은 군집1(1276명)은 총입원횟수 2.4회, 1회평균재원일수가 7.9일, 퇴원 후 최종외래 내원간격이 3.2개월로 재입원패턴의 일반적인 형태라 볼 수 있다. 분석결과 내과, 소아과 환자들이 많으며 호흡기, 소화기계 질환이 34%를 차지하였다.

군집2(72명)는 총입원횟수 3.4회였고 1회 평균재원일수가 38일로, 손상 및 중독, 외상(25%), 순환기계 질환(13.9%)순의 분포를 보였다.

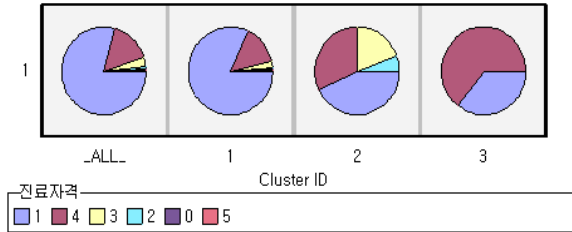
Table. 2 재입원패턴에 따른 군집분석 결과

구분 군집	빈도수	총입원횟수	총재원일수	퇴원후진료이용기간	1회입원평균재원일
군집1	1276	2.24	19.04	3.24	7.9
군집2	72	3.36	115.81	2.3	38.4
군집3	14	5.5	285.93	1.37	54.8

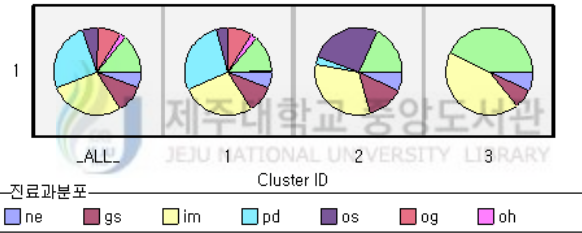
재입원유무



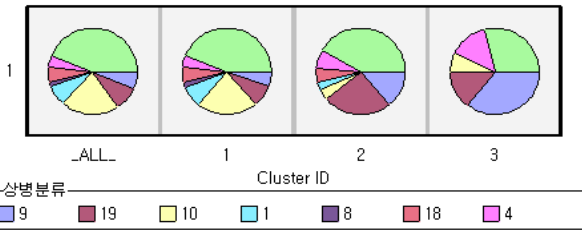
재입원유무



재입원유무



재입원유무



ne:	0:	7: 55 ~ 64	9:
im:	1:	8: 65 ~ 74	19: ,
gs:	2:	1: 0 ~ 4	10:
pd:	3:	2: 5 ~ 14	1:
os:	4:	3: 15 ~ 24	8:
oh:	5:	6: 45 ~ 54	18:
og:		9: 75	4:

Fig. 16 재입원패턴에 따른 군집별 제특성 비교 결과

각 군집별 연령계층, 진료자격, 진료과, 질병군 등을 기준한 분석결과를 Fig. 16에서 볼 수 있으며, Table. 3에 그룹별 대표적인 특성분석 결과를 볼 수 있다.

Table. 3 재입원패턴에 따른 군집별 주요특성분석

특 성	구 분	군집 1	군집 2	군집 3
질병군	2.신생물	8.9%	9.7%	
	4.내분비, 영양 및 대사질환	4.4%	6.9%	14.3%
	5.정신 및 행동장애	0.9%	1.4%	
	7.눈 및 눈부속기 질환	2.5%		
	9.순환기계의 질환	5.6%	13.9%	35.7%
	10.호흡기계의 질환	22.4%	4.2%	7.1%
	11.소화기계의 질환	11.6%	9.7%	7.1%
	13.근골격계 및 결합조직의 질환	2.9%	9.7%	
	14.비뇨생식기계의 질환	4.6%	2.8%	14.3%
	19.손상 및 중독 및 외인에 의한 특정 기타 결과	8.1%	25.0%	14.3%
진료과	내 과	27.2%	31.9%	42.9%
	소 아 과	27.4%	2.8%	
	정형외과	4.2%	26.4%	
	일반외과	10.2%	13.9%	7.1%
	신경외과	2.7%	12.5%	42.9%

4) 병원이용빈도에 따른 군집분석 결과

Table. 4 병원이용빈도에 따른 군집분석 결과

분류변수	군집 1	군집 2	군집 3	군집 4
외래방문횟수	10.5	6.4	13.5	36.8
총 재 원 일	50.5	7.3	120.2	10.7
이 용 기 간	2.6	4.1	1.3	-0.4
입 원 횟 수	2.1	1.2	3.2	1.7
빈 도 (%)	340(4.2)	7516(91.9)	37(0.5)	279(3.4)

분류된 군집 결과를 보면 군집2(7516명)가 가장 다수를 점하고 있는데 평균적으로 총입원횟수 1.2회, 총외래방문횟수가 6.4회이며, 분석시점을 기준하여 약 4개월 전까지 외래 진료를 받았던 환자로 분석대상병원을 이용하는 입원환자들의 일반적인 형태로 저이용 환자군으로 분류할 수 있다.

군집1, 군집2, 군집3의 경우는 군집2에 비해 이용빈도가 높은 환자군으로, 군집3(39명)의 경우는 빈도수는 적지만 장기입원환자로 볼 수 있으며, 군집1(340명)은 외래 진료보다는 입원진료 중심적 환자군, 군집4(279명)는 입원진료보다는 외래진료 중심의 환자군으로 분류 할 수 있다.

(1) 병원이용빈도 군집별 평균연령의 비교

Table. 5의 병원이용빈도 군집별 평균연령을 비교 결과 군집1, 군집3의 경우 평균연령이 58세 정도로서, 연령계층이 높은 환자들이 상대적으로 병원을 많이 이용하고 있으며, 군집4의 경우 평균연령이 30세 정도로 진료과별 교차분석의 결과와 비교해보면 산·소아과 분야가 많아 평균연령이 낮음을 볼 수 있다.

Table. 5 병원이용빈도 군집별 평균연령의 비교

구분	군집1	군집2	군집3	군집4	F
평균연령 (표준편차)	57.6 (21.8)	36.0 (26.3)	58.2 (18.5)	30.3 (31.6)	87.05 (p<.0001)

(2) 병원이용빈도에 따른 군집별 특성분석 결과

Table. 6은 병원이용빈도에 따른 군집과 각 특성항목을 교차분석한 결과이다. 성별과 교차분석 결과 입원 다이용환자군인 군집1, 군집3과 외래를 다이용하는 입원환자군집4의 경우도 남자의 빈도가 상대적으로 높음을 볼 수 있다.

진료자격에 따른 분석결과를 볼 때 군집2와 군집4의 경우 전체 환자의 진료자격 구성과 유사한 경향을 보이고 있는데 비하여 다이용환자군집인 군집1, 군집3의 경우 의료보호 환자가 각 22.94%, 32.43%로 입원 저이용환자군집에 비해 2배정도 많으며, 앞서의 군집분석 결과(Table. 6) 총재원일도 훨씬 긴 것으로 나타났다.

Table. 6 병원이용빈도 군집과 각 항목별 교차분석 결과

항 목	구 분	각 군집별 빈도수 및 칼럼백분율				명 (%)	χ^2
		군집 1	군집 2	군집 3	군집 4		
성 별	남 자	186 (54.71)	3581 (47.65)	21 (56.76)	163 (58.42)	19.4*** (p=0.0002)	
	여 자	154 (45.29)	3935 (52.35)	16 (43.24)	116 (41.58)		
거주지역	제 주 시	186 (54.71)	4457 (59.30)	16 (43.24)	184 (65.95)	30.1*** (p=0.0027)	
	북제주군	91 (26.73)	1580 (21.02)	12 (32.43)	56 (20.07)		
	서귀포시	9 (2.65)	315 (4.19)	1 (2.70)	0 (0.00)		
	남제주군	42 (12.35)	851 (11.32)	6 (16.22)	33 (11.83)		
	기타지역	12 (3.53)	313 (4.16)	2 (5.41)	6 (2.15)		
입원경로	응 급 실	168 (49.41)	3018 (40.15)	9 (24.32)	65 (26.30)	48.8*** (P<.0001)	
	외 래	172 (50.59)	4498 (59.85)	28 (75.68)	214 (76.70)		
진료자격	전 액 자 비	5 (1.47)	139 (1.85)	0 (0.00)	0 (0.00)	247.2*** (p<.0001)	
	의 료 보 험	203 (59.71)	6480 (86.22)	15 (40.54)	237 (84.95)		
	산 재 보 험	11 (3.24)	26 (0.35)	4 (10.81)	1 (0.36)		
	자동차 보험	43 (12.65)	201 (2.67)	6 (16.22)	2 (0.72)		
	의 료 보 호	78 (22.94)	642 (8.54)	12 (32.43)	37 (13.26)		
	기 타	0 (0.00)	28 (0.37)	0 (0.00)	2 (0.72)		

*** p < 0.01

5) 병원이용빈도 관련 특성변수의 상대적 중요도 평가모형

병원이용빈도관련 분류된 4개의 군집 중에서 군집1(입원 다이용, 외래 저이용 환자군)과 군집4(입원 저이용, 외래 다이용 환자군)만을 대상으로 데이터마이닝 기법인 결정트리를 이용하여 병원이용빈도의 영향 요인을 상대적 중요도로 분석한 결과 진료과, 나이, 거주지 등이 중요한 특성변수로 나타났다.(Fig. 17)

첫 번째 노드가 진료과로 분류되어 병원이용빈도를 설명하는 가장 중요한 변수로 제시되었다. 즉 이비인후과, 안과, 소아과, 비뇨기과의 경우 입원진료 이용빈도

에 비해 외래중심적 진료를 받는다고 볼 수 있으며, 외래진료보다는 입원중심적 진료를 받는 군집4의 경우는 내과, 산부인과 환자와 다른 외과분야의 경우 노인, 보호 2종 등 병원진료비 감면혜택을 받고 있는 환자들이 많은 비중을 차지하고 있다.

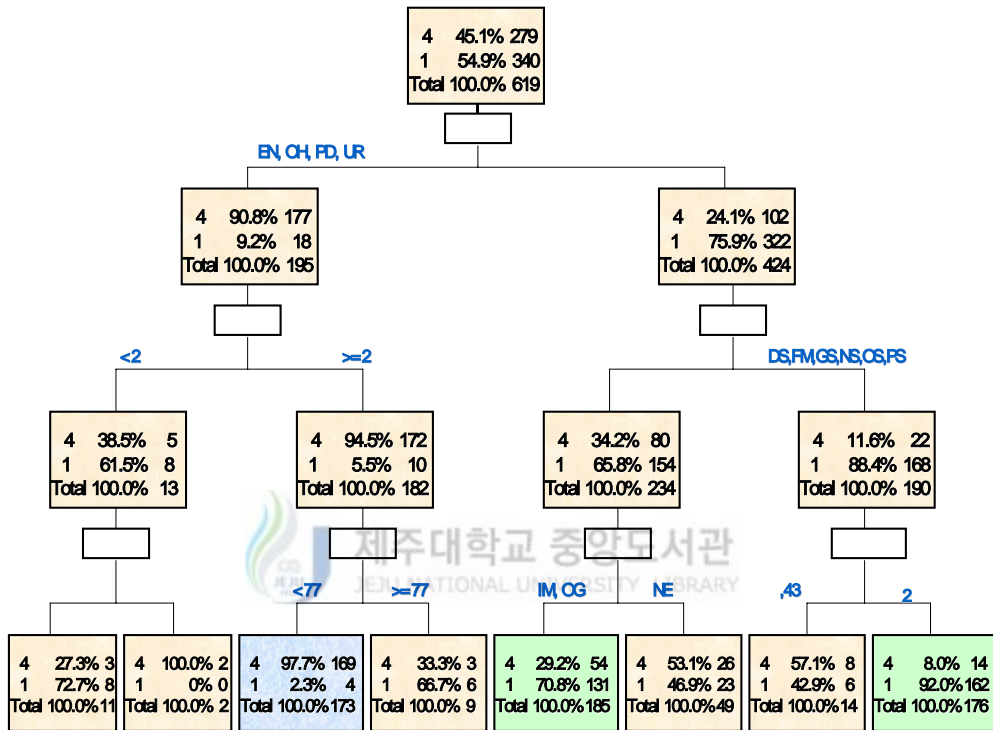


Fig. 17 특정군집 결정을 위한 특성변수의 중요도 평가모형

V. 결론 및 향후 연구

본 논문은 병원데이터베이스로부터 입원환자의 특성분석을 위해 데이터마이닝 프로세스를 응용한 모형화 작업으로 새롭고 유용한 패턴과 지식을 얻기 위한 목적으로 수행하였으며, 어떠한 속성들이 각각의 부류값에 결정적인 영향을 주는가를 쉽게 파악하는 데는 결정트리 분석기법이, 또한 유사성을 갖는 몇 개의 군집으로 나누는 과정에는 대용량의 데이터에 대한 탐색적 기법인 사전적인 정보없이 의미있는 자료구조를 얻을 수 있는 군집분석 기법이 적용되어지는 형태를 알 수 있다.

제주지역의 A 종합병원을 대상으로 2001년 3월부터 2002년 2월까지 입원진료를 받고 퇴원한 환자와 관련된 DB를 연구목적에 맞게 다각적으로 재가공하여 분석하였으며, 다음과 같은 입원환자의 특성분석 결과를 얻을 수 있었다.

첫째, 분석대상자의 일반적 특성은 연령계층별로는 소아와 노인 연령계층이 많은 비율을 차지하고 있었으며, 진료자격별로는 의료보호 환자가 전국평균에 비해 많이 차지하고 있다. 진료과별 특성을 보면 소아과의 경우 전국평균과점유율(10.2%)의 약 2.4배인 24.1%를, 신경과의 경우 전국평균 2.3%의 약 3.5배인 7.14%로 이들 진료과가 전문화되어 있고 인지도가 높음을 알 수 있다. 질병별로는 호흡기계 질환이 18.9%로 가장 높게 나타났다.

둘째, 재입원패턴을 분석해본 결과 1년 동안의 전체 재입원율은 16.4%이며, 1회 입원시 평균재원일수는 7.7일이었다. 결정트리와 대화식 결정트리 분석기법을 적용하여 재입원 가능성 결정모형을 분석 결과 나이, 진료자격, 수술여부, 진료과 등이 특성변수로 나타났으며, 그 중 나이가 가장 중요한 변수임을 알 수 있었다. 재입원 패턴에 따라 3개의 군집으로 분류할 수 있었으며, 각 군집과 관련 특성들을 교차분석한 결과 장기입원환자군인 군집3으로 갈수록 연령계층이 높았으며, 의료보호 자격권자와 순환기계 질환의 내과 진료의 분포가 많았다. 소아가 많은 군집1의 경우 호흡기계 질환이 대표질환임을 알 수 있었다.

셋째, 병원이용빈도에 따른 군집분석 결과 4개의 군집으로 분류되었으며, 군집2의 경우 입원/외래 저이용환자군으로 일반적인 입원환자의 병원이용형태로 볼 수 있으며, 군집3은 장기입원환자군이다. 분류된 군집 중 군집1(외래저이용, 입원다이

용 환자군)과 군집4(입원저이용, 외래다이용 환자군)만을 대상으로 결정트리 분석 기법을 이용하여 군집결정 영향 요인을 상대적 중요도로 분석한 결과 진료과, 나이, 거주지역 등이 중요한 특성변수로 나타났다.

본 연구에서의 제한점으로는 1개 종합병원의 자료를 기준으로 하여 최근 1년 동안의 입원환자에 한해서 분석이 이루어졌으며, 향후 연구에서는 외래환자 및 장기간의 입원환자를 대상으로 분석이 요구되어진다. 입원환자란 어느 정도 중증도가 있는 환자군으로, 이를 기준한 단기간의 환자의 병원이용패턴분석결과를 의사결정에의 반영은 실제 적용함에 있어 신중해야 할 것이다.

또한 제주지역 병원들의 통합 데이터베이스를 통한 제주지역 환자들의 특성분석을 위한 연구가 진행되면 의료의 질적 서비스 향상에 많은 도움이 될 것이다.

본 연구는 지금까지의 병원정보시스템 흐름상 단순 통계에 의한 비교분석에 의한 병원의 의사결정을 행하고 있는 업무처리 시스템에 좀 더 체계적이고 타당성 있는 의사결정시스템을 위한 모형을 제안하였다는 점과 최근의 제주지역 병원자료를 체계적으로 분석, 구체화하였다는 데에 의의를 두고자 한다. 또한 제주지역에서는 처음으로 병원데이터베이스를 활용한 데이터마이닝 기법을 적용하여 병원이용환자의 특성을 분석하였다는 점에도 의의가 있다.

참고문헌

- 강성홍, 강수원, 김병수, 이상열, 2001, "메뉴방식의 SAS를 이용한 병원통계분석학", 청구문화사.
- 강현철, 한상태, 최종후, 김은석, 김미경, 2001, "SAS Enterprise Miner 4.0을 이용한 데이터마이닝 방법론 및 활용", 자유아카데미.
- 김유미, 2000, 퇴원요약 DB를 이용한 데이터베이스 마케팅, 인제대학교 보건대학원 석사학위논문.
- 김종호, "의료데이터웨어하우스에 대한 개념적 이해", (주)비트컴퓨터.
- 김현희, 이정원, 김호숙, 최지영, 이진숙, 최윤정, 김민정, 이혜수, 김백선, 윤혜성, 박승수, 이상호, 용환승, 2000, 최근 데이터마이닝 알고리즘의 경향 및 추세, 이화여자대학교 과학기술대학원 EIST 연구보고서. EIST-TR-CSE-00003.
- 김혜숙, 채영문, 탁관철, 박현주, 호승희. 2001, 데이터마이닝을 이용한 의료의 질 측정지표 분석 및 의사결정지원시스템 개발, 한국의료QA학회지. 8권 2 호
- 한국보건사회연구원, 2000, "1999년도 환자조사보고서", 보건복지부 용역보고서 2000-64
- 서두성, 1998, 의사결정나무의 평가와 응용, 고려대학교 대학원 석사학위논문.
- 이진숙, 2001, 금융 환경에서 데이터마이닝을 활용하는 방안에 관한 연구, 이화여자대학교 과학기술대학원 석사학위논문.
- 장남식, 홍성완, 장재호, 1999, "데이터마이닝", 대청미디어, P.26.
- 장동인, 1998, 실무자를 위한 데이터웨어하우스, ORACLE.
- 조우현, 이선희, 이해중, 전기홍, 1999, "의료서비스마케팅", 퇴설당.
- 최국렬, 조대현, 이상열, 석경하, 박일수, 김유미, 김옥남, 김병수, 강성홍, 2001, "데이터마이닝 이론과 실습 -보건의료데이터 중심-", 청구문화사.
- 최길림, 2001, 병원이용빈도와 진료수익성에 따른 환자군집별 특성과 데이터베이스 마케팅의 활용성, 인제대학교 대학원 박사학위논문.
- 최종후, 한상태, 강현철, 김은석, 김미경, 2000, "SAS Enterprise Miner를 이용한 데이터마이닝 기능과 사용법", 자유아카데미.

- Fayyad,U.M., Piatetsky-Shapiro,G., Smyth,P.,and R.Uthtusamy, 1996, ADVANCES IN KNOWLEDGE DISCOVERY AND DATA MINING, AAAI/MIT Press.
- Han,J., and Micheline Kamber, August 2000, DATA MINING : CONCEPTS AND TECHNIQUES, Kaufmann Publishers, ISBN 1-55860-489-8.
- J.P. Bigus, 1996, Data Mining with Neural Network, McGraw-Hill.
- J. Shafer, R. Agrawal, and M. Mehta, 1996, SPRINT: A scalable parallel classifier for data mining.
- Jhannes Gehrke, Venkatesh Ganti, and Raghu Ramakrishnan. 1999, "BOAT: Optimistic Decision Tree Construction", In proc. of the ACM SIGMOD Conference on Management of Data, Philadelphia.
- Kohonen, T., 1997, "Self-Organizing Maps", Berlin: Springer-Verlag. Second edition.
- Michael J. A. Berry, and Gordon Linoff, 1997, Data Mining Techniques : For Marketing, Sales, and Customer Support, John Wiley & Sons, Inc.
- Piatetsky-Shapiro,G., Fayyad,U.M., and Smyth,P., 1996, From Data Mining to Knowledge Discovery, in ADVANCES IN KNOWLEDGE DISCOVERY AND DATA MINING, AAAI/MIT Press, pp 1-34, CA.
- Tom M. Mitchell, 1999, "Machine Learning and Data Mining", Communication of the ACM, Vol. 42, No. 11.
- William J. E. Potts, 1999, "Generalized Additive Neural Networks", SAS Institute Inc., ACM KDD-99.

부 록

▣ 분석대상병원 입원환자의 특성분석 결과

Table. A 연령계층별 퇴원환자수 및 남/여 구성비

구분 연령	1999년도 전국대상 환자조사보고서							분석대상병원						
	1999년 9월 1일 ~ 9월 30일							2001년 3월 1일 ~ 2002년 2월 28일						
	환자수(명)			구성비(%)			성비 남/여	환자수(명)			구성비(%)			성비 남/여
	계	남	여	계	남	여		계	남	여	계	남	여	
0~4	28,915 (100.0)	17,251 (59.7)	11,664 (40.3)	7.9	9.7	6.1	1.48	1,566 (100.0)	924 (59.0)	642 (41.0)	18.9	23.0	15.0	1.44
5~14	16,478 (100.0)	10,354 (62.8)	6,124 (37.2)	4.5	5.8	3.2	1.69	776 (100.0)	440 (56.7)	336 (43.3)	9.3	11.0	7.8	1.31
15~24	37,506 (100.0)	19,077 (50.9)	18,429 (49.1)	10.2	10.7	9.7	1.04	603 (100.0)	311 (51.6)	292 (48.4)	7.3	7.7	6.8	1.07
25~34	85,911 (100.0)	27,586 (32.1)	58,325 (67.9)	23.4	15.5	30.8	0.47	1,195 (100.0)	379 (31.7)	816 (68.3)	14.4	9.4	19.0	0.46
35~44	57,676 (100.0)	31,904 (55.3)	25,772 (44.7)	15.7	17.9	13.6	1.24	810 (100.0)	379 (46.8)	431 (53.2)	9.8	9.4	10.1	0.88
45~54	42,149 (100.0)	24,080 (57.1)	18,069 (42.9)	11.5	13.5	9.5	1.33	774 (100.0)	424 (54.8)	350 (45.2)	9.3	10.6	8.2	1.21
55~64	45,033 (100.0)	24,683 (54.8)	20,350 (45.2)	12.2	13.8	10.8	1.21	808 (100.0)	453 (56.1)	355 (43.9)	9.7	11.3	8.3	1.28
65~74	34,711 (100.0)	15,969 (46.0)	18,742 (54.0)	9.4	9.0	9.9	0.85	991 (100.0)	461 (46.5)	530 (53.5)	11.9	11.5	12.4	0.87
75세 이상	19,302 (100.0)	7,510 (38.9)	11,792 (61.1)	5.2	4.2	6.2	0.64	777 (100.0)	243 (31.3)	534 (68.7)	9.4	6.1	12.5	0.46
전체 (%)	367,681 (100.0)	178,414 (48.5)	189,267 (51.5)	100	100	100	0.94	8,300 (100.0)	3,771 (48.4)	3,752 (51.6)	100	93.9	100	0.94

Table. B 진료자격별 환자분포

진료자격	구분	1999년도 환자조사보고서 퇴원환자 비율(%)			부산지역 A병원*		분석대상병원	
		종합병원	병원	전체	빈도수	비율(%)	빈도수	비율(%)
전액자비		2.5	3.7	2.7	39	1.0	144	1.73
의료보험		82.6	68.3	76.7	3,768	92.6	7,030	84.70
산재보험		1.6	1.5	1.5	66	1.6	42	0.51
자동차보험		5.3	14.8	11.7	118	1.8	255	3.07
의료보호		7.2	10.3	6.6	741	2.9	799	9.63
기타		0.8	1.4	0.9	4	0.1	30	0.36
전체		100.0	100.0	100.0	4,069	100.0	8300	100.0

* 부산지역 820병상의 종합전문요양기관으로 2000년 9월~2000년 12월까지 재입원환자 4,069명 대상

Table. C 분석대상병원의 진료과 분포

진료과	전국평균* 과점유율(%)	부산지역병원** 과점유율(%)	빈도	비율(%)
내과	28.7	19.1	1,731	20.86
신경과	2.9	2.3	593	7.14
가정의학과	0.3	0.0	36	0.43
일반외과	10.2	10.4	901	10.86
신경외과	3.9	8.5	339	4.08
정형외과	6.5	6.9	432	5.20
성형외과	2.3	4.0	187	2.25
산부인과	12.5	13.7	1,043	12.57
소아과	10.2	10.4	2,000	24.10
안과	4.0	8.2	154	1.86
이비인후과	5.8	4.2	527	6.35
비뇨기과	3.0	2.9	272	3.28
치과	0.8	1.3	85	1.02
계	91.1	91.9	8,300	100

* 전국평균 - 17개 3차의료기관 1998년 연보자료.

** 부산지역 일개 3차의료기관의 진료과 분포 (인제대학교 석사학위논문, 김유미, 2000)

Table. D 질병별 성별 평균입원일 및 주진단 질병군 분포(KCD-3 대분류)

상 병 분 류	평균입원일				부산 병원 비율(%)	분석대상		전국*** 10대 상병(%)
	전국평균*		분석대상병원**			빈도	비율	
	남	여	남	여				
1.특정감염성 및 기생충성 질환	8.9	6.5	5.6	5.6	4.4	662	7.98	4.8
2.신생물	14.0	12.1	9.7	9.0	15.0	461	5.55	7.4
3.혈액, 조혈기관 질환 및 면역기전을 침범하는 특정장애	10.8	9.4	4.0	6.8	0.9	35	0.42	
4.내분비, 영양 및 대사질환	13.5	11.6	9.7	9.2	2.5	222	2.67	
5.정신 및 행동장애	72.6	55.6	13.5	8.7	2.1	47	0.57	3.2
6.신경계 질환	21.5	13.7	6.4	6.4	2.3	237	2.86	
7.눈 및 눈부속기 질환	3.7	2.7	3.8	3.8	6.8	168	2.02	
8.귀 및 유양돌기의 질환	6.8	6.2	5.4	5.4	0.8	262	3.16	
9.순환기계 질환	13.2	13.2	10.9	10.7	8.8	481	5.80	5.6
10.호흡기계의 질환	8.4	6.8	6.0	5.9	6.6	1,569	18.90	6.5
11.소화기계의 질환	8.4	7.4	8.1	7.6	8.8	824	9.93	10.3
12.피부염 및 피하조직의 질환	13.0	9.8	9.4	10.3	1.0	89	1.07	
13.근골격계 및 결합조직의 질환	19.8	15.7	12.6	14.9	4.0	253	3.05	4.7
14.비뇨생식기계의 질환	8.3	7.0	6.3	6.6	5.0	500	6.02	3.8
15.임신, 분만 및 산욕의 합병증	-	4.0	-	5.2	9.8	717	8.64	17.7
16.주산기에 관련된 일정한 병태	8.8	8.9	7.7	7.7	2.1	182	2.19	
17.선천기형, 변형 및 염색체이상	8.8	8.5	7.7	5.3	4.2	34	0.41	
18.달리 분류되지 않는 증상, 징후와 임상감사의 이상	5.6	5.4	5.9	6.1	0.8	537	6.47	
19.손상 및 중독 및 외인에 의한 특정 기타 결과	16.4	16.1	12.2	14.5	12.0	910	10.96	21.9
20.질병이환 및 사망의 외인	6.3	4.6	5.3	4.6		59	0.71	
21.건강상태, 보건서비스 접촉에 영향을 주는 요인	7.6	6.9	6.0	5.4	2.0	9	0.11	

* 퇴원환자 평균입원일수 (자료: 보건복지부, 「1999환자조사」, 2001.)

** 분석대상병원의 질병별 남/여 평균입원일수

*** 퇴원환자 10대 상병순위 (자료: 보건복지부 환자조사, 1996)

Table. E 분석대상환자의 입원 특성

특 성	구 분	빈도수(명)	비율(%)	평균 (표준편차)
재입원유무	무	6,938	83.59	
	유	1,362	16.41	
총입원횟수	1회	6,938	83.59	1.25 (0.71)
	2회	957	11.53	
	3~6회	386	4.65	
	7회 이상	19	0.23	
총재원일수	7일 이내	5,082	61.23	10.5 (18.49)
	8~14일	1,900	22.89	
	15~30일	852	10.27	
	31~60일	318	3.83	
	61~90일	79	0.95	
	91~120일	25	0.30	
	121~180일	26	0.31	
181일 이상	18	0.22		
계		8,300	100.00	

Table. F 입원횟수별 1회 평균 재원일수

입원횟수	빈도수	1회입원 평균재원일수				F (P)
		부산지역 1개병원*		분석대상병원		
		평 균	표준편차	평 균	표준편차	
1회	6,938	13.93	21.02	7.28	7.89	61.8 (<.0001)
2회	957	16.26	20.06	8.81	8.98	
3~6회	386	18.31	16.76	12.87	13.32	
7회 이상	19	14.22	12.71	11.08	11.16	
계	8,300	14.62	20.55	7.72	8.45	

Table. G 일반적특성에 따른 재입원유무 및 총입원횟수 · 총재원일수 평균비교

특 성	구분	재입원유무		χ^2 (P)	총입원횟수		총재원일수	
		무	유		평균	F (p)	평균	F (p)
		빈도	빈도		표준편차		표준편차	
		비율 %	비율 %					
성 별	남	3293	721	13.659 0.0002	1.28	17.41 <.0001	11.21	11.77 0.0006
		82.04	17.96		0.76		20.09	
	여	3645	641		1.22		9.82	
		85.04	14.96		0.66		16.83	
연 령	0~4	1268	298	141.149 <.0004	1.27	17.91 <.0001	6.90	48.44 <.0001
		80.97	19.03		0.67		8.00	
	5~14	694	82		1.12		5.64	
		89.43	10.57		0.41		5.38	
	15~24	544	59		1.15		7.88	
		90.22	9.78		0.72		10.75	
	25~34	1055	140		1.16		7.59	
		88.28	11.72		0.52		10.73	
	35~44	708	102		1.19		9.90	
		87.41	12.59		0.62		17.76	
	45~54	658	116		1.25		12.81	
		85.01	14.99		0.73		23.66	
	55~64	656	152		1.30		14.23	
		81.19	18.81		0.83		25.25	
	65~74	757	234		1.42		16.22	
		76.39	23.61		0.97		26.87	
75세 이상	598	179	1.34	16.24				
	76.96	23.04	0.76	24.42				
진 차 료 격	0.전액자비	135	9	90.963 <.0001	1.06	22.34 <.0001	7.90	97.00 <.0001
		93.75	6.25		0.24		9.00	
	1.의료보험	5951	1080		1.23		8.88	
		84.64	15.36		0.67		13.46	
	2.산재보험	30	12		1.40		32.24	
		71.43	28.57		0.73		31.68	
	3.자동차보험	211	44		1.21		23.60	
		82.75	17.25		0.50		28.47	
	4.의료보호	582	216		1.49		20.06	
		72.93	27.07		1.04		37.94	
5.기 타	29	1	1.03	6.10				
	96.67	3.33	0.18	4.81				
거주지	1.제 주 시	4126	787	28.068 <.0001	1.24	5.20 0.0004	9.85	5.89 <.0001
		83.98	16.02		0.71		18.20	
	2.북제주군	1450	319		1.28		11.92	
		81.97	18.03		0.73		18.84	
	3.서귀포시	282	48		1.20		10.53	
		85.45	14.55		0.57		22.43	
	4.남제주군	771	182		1.30		11.75	
		80.90	19.10		0.77		18.77	
	5.기타지역	309	26		1.12		8.90	
		92.24	7.76		0.55		14.76	

Table. H 진료특성에 따른 재입원유무 및 총입원횟수·총재원일수 평균비교

특 성	구분	재입원유무		χ^2 (<i>P</i>)	총입원횟수		총재원일수				
		무	유		평균	<i>F</i> (<i>p</i>)	평균	<i>F</i> (<i>p</i>)			
		빈도	빈도		표준편차		표준편차				
		비율 %	비율 %								
입 원 경 로	1. 응급실	2844	450	30.076	1.21	1.22	10.22	49.23			
		86.34	13.66		0.66		13.19				
	2. 외 래	4094	912		<.0001		1.27		0.2696	10.68	<.0001
		81.78	18.22		0.74		21.26				
퇴 원 결 과	계 속	6860	1318	34.876	1.24	94.42	10.26	9.62			
		83.88	16.12		0.70		18.08				
	사 망	78	44		<.0001		1.70		<.0001	26.55	<.0001
		63.93	36.07		1.25		32.78				
진료과	내과 (IM)	1355	376	77.644	1.39	9.62	13.36	55.25			
		78.28	21.72		0.93		23.37				
	신경과 (NE)	512	81		1.20		11.24				
		86.34	13.66		0.66		19.21				
	가정의학과 (FM)	28	8		1.25		11.03				
		77.78	22.22		0.50		10.62				
	일반외과 (GS)	760	141		1.26		12.45				
		84.35	15.65		0.79		17.94				
	신경외과 (NS)	289	50		1.26		21.61				
		85.25	14.75		0.77		43.84				
	정형외과 (OS)	359	73		1.22		23.98				
		83.10	16.90		0.54		26.27				
	성형외과 (PS)	164	23		1.14		10.82				
		87.70	12.30		0.39		15.23				
	산부인과 (OG)	925	118		1.14		6.61				
		88.69	11.31		0.45		5.83				
	소아과 (PD)	1648	352		1.25		6.41				
		82.40	17.60		0.71		7.26				
	안과 (OH)	125	29		1.21		3.81				
		81.17	18.83		0.47		3.82				
	이비인후과 (EN)	468	59		1.14		5.53				
		88.80	11.20		0.44		4.00				
	비뇨기과 (UR)	230	42		1.22		7.72				
		84.56	15.44		0.67		8.61				
치과 (DS)	75	10	1.12	8.11							
	88.24	11.76	0.32	9.31							
수 술 유 무	무	5280	989	7.498	1.24	3.09	9.37	95.73			
		84.22	15.78		0.70		15.93				
	유	1658	373		0.0062		1.27		0.0786	13.96	<.0001
		81.63	18.37		0.74		24.44				

Table. I 질병에 따른 재입원유무 및 총입원횟수 · 총재원일수 평균비교

상 병 분 류	재입원유무		χ^2 (P)	총입원횟수		총재원일수	
	무	유		평균	F	평균	F
	빈도	빈도		표준편차	(p)	표준편차	(p)
	비율%	비율%					
1.특정감염성 및 기생충성 질환	39	3		1.14		10.05	
	92.86	7.14		0.65		22.71	
2.신생물	560	102		1.20		7.19	
	84.59	15.41		0.53		10.31	
3.혈액, 조혈기관 질환 및 면역기전을 침범하는 특정장애	340	121		1.51		14.31	
	73.75	26.25		1.14		17.09	
4.내분비, 영양 및 대사질환	30	5		1.74		7.14	
	85.71	14.29		2.52		6.44	
5.정신 및 행동장애	159	63		1.47		16.55	
	71.62	28.38		0.94		30.87	
6.신경계 질환	34	13		1.40		16.28	
	72.34	27.66		0.74		22.79	
7.눈 및 눈부속기 질환	199	38		1.23		8.48	
	83.97	16.03		0.70		12.48	
8.귀 및 유양돌기의 질환	136	32		1.21		4.68	
	80.95	19.05		0.47		5.76	
9.순환기계의 질환	237	25		1.14		6.16	
	90.46	9.54		0.52		4.36	
10.호흡기계의 질환	394	87		1.32		17.34	
	81.91	18.09		0.89		35.98	
11.소화기계의 질환	1279	290	123.30	1.26	8.41	7.82	20.57
	81.52	18.48		0.62		11.47	
12.피부염 및 피하조직의 질환	668	156	<.0001	1.33	<.0001	11.16	<.0001
	81.07	18.93		0.89		16.55	
13.근골격계 및 결합조직의 질환	76	13		1.22		14.18	
	85.39	14.61		0.62		27.63	
14.비뇨생식기계의 질환	209	44		1.27		18.35	
	82.61	17.39		0.74		21.33	
15.임신, 분만 및 산욕의 합병증	437	63		1.18		8.88	
	87.40	12.60		0.62		22.23	
16.주산기에 관련된 일정한 병태	630	87		1.16		6.24	
	87.87	12.13		0.48		6.43	
17.선천기형,변형 및 염색체이상	169	13		1.09		8.88	
	92.86	7.14		0.35		14.97	
18.달리 분류되지 않는 증상, 징후와 임상검사의 이상	30	4		1.12		7.65	
	88.24	11.76		0.33		10.31	
19.손상 및 중독 및 외인에 의한 특정 기타 결과	461	76		1.22		7.98	
	85.85	14.15		0.67		11.94	
20.질병이환 및 사망의 외인	787	123		1.17		16.60	
	86.48	13.52		0.48		24.68	
21.건강상태, 보건서비스 접 촉에 영향을 주는 요인	56	3		1.05		5.20	
	94.92	5.08		0.22		4.65	
0.기 타	8	1		1.11		6.67	
	88.89	11.11		0.33		5.60	

감사의 글

2000년의 시작을 변화에 대한 호기심과 두려움 속에 대학원 생활을 동반한지 2년 6개월이라는 기간이 흘렀습니다. 마무리 단계인 지금까지도 많은 아쉬움이 남아 있지만 이 아쉬움은 보다 나은 도전에 대한 희망이 됩니다.

본 논문이 완성되기까지 따뜻한 관심과 배려로 지도해주신 이상준 지도교수님께 깊은 존경과 감사의 마음을 드립니다. 아울러 논문심사를 맡아 세심한 배려와 학문연구의 진정한 의미를 일깨워주신 김장형 교수님, 곽호영 교수님께 감사들이며, 결실을 맺기까지 항상 관심과 지도편달을 아끼지 않으신 안기중 교수님, 변상용 교수님, 송왕철 교수님께도 깊은 감사를 드립니다.

항상 시간의 부족함을 느끼는 저에게 모자란 부분을 채워주기 위해 충고와 조언을 아끼지 않고 많은 힘이 되어준 우리 연구실의 김영민 선생님, 박충희 선생님, 이종헌씨, 정은경씨, 문남월씨와 광석, 세진, 은주에게도 고마움을 전하며, 같이 시작한 동기 변태보 선생님, 허동진 선생님과 행정적 절차를 도맡아 해주신 양영수 선생님에게도 고마움을 전합니다. 그리고 대학원 생활을 하며 마주친 선배님, 동기들과 후배들에게도 깊은 감사를 드립니다.

학업을 할 수 있도록 배려해주신 제주대학교병원의 홍강의 원장님을 비롯하여 직장 상사 및 모든 동료들에게 머리 숙여 마음속 감사를 전해드리며 보다 더 나은 모습으로 노력하겠다는 마음을 전해드립니다.

직장생활과 대학원 공부를 병행하면서 힘들었던 시간들이었지만, 나의 부족함까지 채워가며 늘 함께한 영원한 동반자 형준씨와 아직 어려 엄마품이 그리다는 안스러운 말과 열심히 하라는 응원까지 해준 아들 태완, 대훈에게 고마움을, 그리고 묵묵히 지켜봐주신 부모님들께도 깊은 감사를 드립니다. 더불어 많은 희망을 주신 친지분들께도 감사를 드립니다.

이 글을 통해 미처 감사드리지 못한, 저와 인연이 닿았던 많은 분들께 진심으로 감사를 드립니다.