대한의료정보학회지 제15권 제2호
J Kor Soc Med Informatics 2009;15(2):175-189

종설

임상에서의 데이터 마이닝 개념과 원칙

이선미[1], 박래웅[2,3]

가톨릭대학교 간호대학[1], 아주대학교 의과대학 의료정보학과[2], 아주대학교의료원 u-health 정보 연구소[3]

Basic Concepts and Principles of Data Mining in Clinical Practice

Sun-Mi Lee[1], Rae Woong Park[2,3]

The Catholic Univ. of Korea College of Nursing[1],
Dept. of Biomedical Informatics, School of Medicine, Ajou Univ.[2],
Institute for u-health Information Research, Ajou Univ. Medical Center[3]

Abstract

Recently, many hospitals have been adopting clinical data warehouses (CDW) as well as electronic medical records. These new hospital information systems are inevitably introducing very large amounts of clinical data that might be useful for further analysis. However, the electronic clinical data in the CDW are usually byproducts of clinical practice rather than the product of research. Therefore, they include inconsistent and sometimes erroneous information that might not have the specific context of the clinical situations. Data miners usually have various academic backgrounds such as electronics, informatics, statistics, biomedicine, and public health. If the complex situations surrounding the clinical data are not well understood, investigators performing data mining in clinical fields may have problems assessing the information they are confronted with. Here, we would like to introduce some basic concepts on the principles of data mining in clinical fields including legal and ethical considerations as well as technical concerns. ***(Journal of Korean Society of Medical Informatics 15-2, 175-189, 2009)***

Key words: Clinical Data Mining, Machine Learning

Received for review: June 15, 2009; **Accepted for publication:** June 22, 2009
Corresponding Author: Rae Woong Park, Department of Biomedical Informatics, School of Medicine, Ajou University, San 5, Wonchun-dong, Yeongtong-gu, Suwon 442-721, Korea
Tel: +82-31-219-5342, **Fax:** +82-31-219-4472, **E-mail:** veritas@ajou.ac.kr
DOI:10.4258/jksmi.2009.15.2.175

Ⅰ. 서론

데이터 마이닝은 “대량의 데이터에서 새롭고 유용한 지식을 창출하는 것”으로서, 데이터 더미에서 일반적인 사실을 의미하는 데이터가 아니라 의사 결정에 도움이 되는 유용한 정보를 포함한 지식을 추출하는 것이므로 ‘데이터 캐기(Data mining)’란 용어는 잘못된 용어라 볼 수 있다. 그런 관점에서는 ‘데이터에서 지식을 캐기(knowledge discovery from data: KDD)’라는 용어가 흔히 사용된다. 데이터 마이닝은 데이터 처리, 데이터 요약, 기계학습, 패턴인식, 시각화기술, 통계학, 지식추출기술 등 다양한 분야의 다학제적 기술을 필요로 한다[1)2)]. 기계학습은 데이터 마이닝의 주된 기술적 기반으로서 데이터베이스의 원천 데이터로부터 정보를 뽑아내는 방법론을 제공한다.

전통적으로 의학의 발전을 이끌어온 임상연구 방법론은 전향적 또는 후향적 자료를 관찰하여 분석하는 관찰연구와 실험을 통하여 가설을 증명하는 임상시험으로 분류할 수 있고, 연구결과 가설의 채택여부를 결정하기 위하여 다양한 모수적 또는 비모수적 통계방법을 이용한다. 이러한 연구들은 기본적으로 하나의 가설을 중심으로 연구를 설계하여 수행하고, 결과 데이터를 모으고 분석하여 통계적 검정을 거쳐서 가설을 채택하거나 기각하게 된다. 임상연구는 수행과정과 분석절차의 엄밀성과 확고한 방법론으로 인해서 의학적 연구방법론의 초석으로서 흔들리지 않는 지위를 가지고 있지만, 연구수행을 위하여 많은 노력과 시간, 그리고 비용을 필요로 한다[1)].

한편, 1990년대에 들면서 병원정보시스템이 전세계적으로 보급되고 있으며[3-7)], 최근 들어서는 데이터웨어하우스를 도입하는 병원이 늘어남에 따라서[8-11)], 병원 내에 임상데이터가 기하급수적으로 증가하고 있다. 이러한 데이터는 정교한 실험설계에 의하여 조직화된 데이터라기보다는 환자진료과정의 부산물로서, 전체적으로 보험청구를 중심으로 업무와 데이터 흐름이 설계되어 있어서 실제 환자의 임상적 상황을 잘 반영하지 못하는 경향이 크다. 또한 데이터의 생성과 측정 및 수집과정에 많은 실수와 일관성이 결여되어 신뢰성이 낮은 결과를 도출할 가능성이 많다. 그러나, 가설이나 목적을 전제하지 않아 데이터의 질이 상대적으로 낮고 비뚤림(bias)의 가능성이 큰 단점에도 불구하고, 지속적으로 생산되어 엄청난 양의 데이터가 쌓이고 있으며, 환자들에 대한 임상적 특성을 상당부분 내포하고 있다는 점에서 유용한 임상지식의 보고로서의 가치를 가지고 있다. 따라서, 그 속에 담긴 숨겨진 사실을 뽑아내어 구조화시킴으로써 기존에 알려지지 않았거나 발견하지 못했던 새롭고 유용한 지식을 창조할 필요성이 증가하고 있다[2)].

데이터 마이닝이 갖는 다학제적 성격으로 인해서 의료분야의 데이터 마이닝을 수행하는 연구자들도 전자, 정보통신, 통계학, 보건의학, 생물학 등 다양한 학문적 배경을 가지고 있다. 그러나, 의료분야의 데이터 마이닝은 데이터 자체가 갖는 기술적 문제로부터 윤리 및 사회학적인 문제까지 다양한 복잡한 문제를 안고 있어서[1)], 이러한 특징을 잘 이해하지 못할 경우, 비록 기술적으로 높은 성능을 보여준다 할지라도, 도출된 지식이 임상에서 사용되지 못할 가능성이 크다[2)12)].

유전체 등 omics 데이터를 주된 분석 대상으로 하는 생명정보학 분야에서도 데이터 마이닝은 필수적인 분석기술이다. 이러한 omic 데이터는 이미 임상정보와 결합되어 분석되고 있으며, 유방암이나, 혈액응고제 용량조절처럼 특정 유전자의 발현유무를 바탕으로 치료방향을 정하는 경우가 늘고 있다. 현실적으로는 임상자료와 분자생물학적 자료를 인위적으로 분리할 수 없지만, 기술(description)의 편의상 생물정보학 분야는 언급하지 않고, 임상분야에서의 데이터 마이닝을 중심으로 하여 데이터 마이닝 수행 시 고려하여야 할 일반적인 원칙과 절차를 소개하고, 수행과정에서 반드시 고려하여야 할 사항에 대하여 문헌고찰과 더불어 필자들의 소견을 제시하고자 한다.

Ⅱ. 의료 데이터 및 임상 데이터마이닝의 특징

임상에서 데이터 마이닝의 대상이 되는 의료 데이터는 타 분야의 데이터와 크게 구별되는 특징이 존재하는데, 데이터의 이질성과 복잡성, 부정확성과 오류가능성, 불완전성과 윤리 및 법적인 문제, 개인정보보호, 특징 선택의 제한, 모델의 투명성과 설명력에 대한 높은 요구도 등이 그것이다[1)12)]. 의료데이터의 이러

한 특징들은 마치 숨겨진 지뢰와 같아서, 이들을 잘 이해하고 적절한 절차를 밟아서 신중하게 진행하지 못할 경우 오류가 가미된 신뢰성이 떨어진 결과를 도출하거나 또는 법적, 윤리적 문제를 야기 할 수도 있다.

1. 자료의 이질성, 고차원성 및 복잡성

의료에서 다루는 자료는 저 차원의 데이터로부터 매우 고차원의 추상화 데이터까지 다양하고 이질적인 자료로 구성된다[1)12)]. 키나 체중과 같은 단순한 수치데이터나, CT, MRI와 같은 정지영상데이터, 초음파와 같은 동영상데이터, 환자감시장치나 생체모니터에서 연속적으로 생성되는 생체신호, 환자의 진술, 면담자료, 전문가의 임상소견 등 전혀 다른 속성을 가진 자료가 혼재되어 있기 때문에, 복잡할 뿐만 아니라 처리해야 할 데이터의 양도 많게 된다. 특히나 유전체나 단백체와 같은 'omics' 데이터의 경우 레코드 수는 수십-수 백 개 미만인데 비하여 변수의 수는 수 만 개를 넘는 경우가 흔하다. 환자감시장치와 같이 동일한 환자를 지속적으로 감시하는 장비의 경우 시간의 흐름에 따라 데이터 스트림이 누적되므로 분석해야 할 자료의 양이 매우 많아지게 된다[13)]. 특히 반복 측정되는 값들의 경우 어느 특정 시점의 값을 이용할 것인지, 평균이나 중위값을 사용할지 등 환자의 상태를 나타내는 대표값의 선정에 대하여 논란이 발생하게 된다. 임상문서의 경우 관찰되는 현상을 가능한 자세히 기술하는 진술문 형태를 띄고 있을 뿐만 아니라, 사용되는 용어들이 표준형을 가지고 있지 못해서 기관간에 차이가 있고, 심지어 같은 기관 내 임상의사들 간에도 동일한 개념에 대하여 서로 다른 용어를 사용하는 경우를 흔하게 볼 수 있다[1)].

2. 자료의 부정확성과 오류, 불완전성과 중복

일상 의료행위 중에 수집되는 임상자료는 불완전하고 오류를 포함하는 경우가 많다[1)2)12)13)]. 환자의 동일한 특정 값도 상황에 따라 서로 다른 결과값을 보이는 경우가 흔하게 발생한다. 혈압측정을 예로 본다면, 환자의 체위, 이를테면 누워있거나 앉아있거나, 또는 안정상태에 따라 측정값이 나타날 수 있으며, 혈압을 측정하는 의료인의 성향이나, 또는 측정에 사용되는 혈압계에 따라서도 서로 다르게 측정될 수 있다. 정확히 측정하여도 입력과정에서 잘못 입력하는 경우도 있다: 예) 키를 입력하면서 단위를 착각하여 170(cm)를 1.7(cm)로 입력하는 경우. 결측값이 흔히 존재하는데 결측값의 의미가 우연에 의한 것인지, 또는 의도적으로 빠진 것인지 알 수 없는 경우가 대부분이다. 의료진들간의 직무교대나 진료과간 또는 과내 당직교대간에 업무 인계과정, 또는 직역간의 위계적 보고체계로 인해서 동일한 자료가 반복적으로 기록되는 경우가 많다. 방사선소견 또는 병리소견을 의무기록 내에 복제하여 사용하는 경우도 흔하게 볼 수 있다.

3. 윤리 및 법적인 문제

의무기록은 개인의 민감한 건강문제를 기록하고 있기 때문에 의료법상의 저촉을 받을 뿐만 아니라 고도의 윤리성을 전제로 하고 있다. 따라서 보안과 개인정보보호에 대한 각별한 주의가 필요하다. 데이터 마이닝 역시 관찰연구의 한 분야로서 기관윤리위원회(Institutional Review Board)의 심의를 필요로 한다. 다만, 'minimal risk data'인 경우 환자의 정보보호를 위한 충분한 절차와 신원정보를 제거하는 적절한 방법론이 제시될 경우에는 IRB의 심의를 면제 받을 수 있다. "minimal risk data란 해당 데이터가 환자의 일상적인 진단과 치료과정 중에 수집되어, 환자의 치료나 관리에 어떤 영향도 끼치지 않고, 특정 치료의 수용이나 거부에 대한 어떤 압력도 없으며, 추가적인 정보를 수집하기 위하여 환자나 가족들에게 정보를 요청하는 일이 없는 자료'로서 환자에 대한 기밀성(confidentiality)만이 문제가 되는 자료"를 말한다[14)]. 하지만 심의면제를 받더라도 의료정보를 다루는 모든 연구자들은 소정의 IRB 교육과 인증을 받을 필요가 있다.

4. 보안 및 개인정보보호

연구대상자의 개인정보보호를 위하여 연구대상 자료에 대한 무명화 또는 개인식별정보의 삭제가 필요하다. 무명화된 데이터(anonymized data)는 개인과 개

인의 기록을 연결하는 모든 연관을 비가역적으로 제거한 것으로서 개인과 기록간의 연결을 다시 수립하는 것이 거의 불가능 상태의 데이터를 말한다[1]. 하지만 이런 무명화 데이터는 자료의 오류를 검사하거나 추가적인 자료를 수집할 수가 없어서 연구 목적으로는 부적당한 경우가 많다. 개인식별정보가 제거된 자료는 초기 수집단계에서는 개인식별정보가 있지만, 적절한 단계에서 개인식별정보를 제거한 것이다[1]. 주의 할 것은, 각 필드 개별적으로는 신상정보가 아니지만 여러 필드의 정보를 서로 조합하면 대상자를 좁힐 수 있는 정보가 포함될 경우는 개인식별정보가 제거된 자료라고 할 수 없다. HIPAA(Health Insurance Portability and Accountability Act; http://www.hipaa.org/)의 개인정보보호 규정에 의하면 다음의 정보를 포함하지 말 것을 규정하고 있다: 주민번호, 병원등록번호, 이름, 주소, 우편번호 뒷자리, 날짜(생일, 입원일 등), 전화번호, 팩스번호, 이메일, 계좌번호, 면허번호, 얼굴사진 등. 그러나, 우편번호 앞자리나 날짜 데이터 중 년도에 해당하는 정보는 포함할 수 있다. 그러나, 미국의 HIPAA 규정을 국내 여건에서 완전히 반영하는 것이 타당한 일인지에 대해서는 논란의 여지가 있다. 기본적으로 정보보호에 대한 물리적 및 기술적 방법론을 정립하고, 데이터를 다루는 연구자들에 대하여 정보보호에 대한 문서화된 서약을 주기적으로 받는 것이 바람직하다.

5. 특징 선택(feature selection)의 제한

데이터에서 가용한 모든 특징을 사용하는 것이 분류모형의 성능을 최고로 향상시키는데 도움을 주는 경우가 흔히 있다. 하지만 의료분야에서는 가용한 모든 특징을 모델구축에 사용하기보다는 최소한의 변수만으로 최선의 성능을 추구하는 전략이 필요하다[12]. 실제 임상현장에서 필요한 데이터를 얻기 위해서는 많은 비용과 침습적 시술이 필요하며, 모델에 필요한 모든 변수가 동시에 가용한 경우도 매우 드물기 때문이다. 제안된 예측 또는 분류모델이 사용될 임상현장은 매우 분주하고 급박한 상황이 많이 벌어지기 때문에, 의료진의 업무흐름을 방해하지 않는 최소한의 변수로 구성할 필요가 있다.

6. 데이터 마이닝 모델의 투명성과 설명력

도출된 지식은 기존의 의학적 지식체계를 기반으로 하여야 하며, 해당 모델을 사용할 의료인이 합리적으로 이해하고 분석할 수 있는 형태라야 한다[12]. 예로서 유방암의 예후를 예측하는 모형 속에 임상적으로 관련성을 찾기 어려운 간효소치가 입력변수에 포함되어 있다거나 또는 적절한 입력변수가 선정되었더라도 이산화된 값이 이해하기 어려운 수준(예: Age>45.7)으로 되어 있다면 해당 모델이 임상의사들에게 받아들여질 가능성이 낮아지게 될 것이다. 데이터를 정제하고 변수를 추출하는 과정에서, 논란의 여지가 있겠으나, 임상적으로 설명하기 어렵거나 연관을 짓기 어려운 변수는 초기 특징 선택 단계에서 배제하고 분석을 진행할 필요도 있다. 도출된 지식을 이용한 의사결정 과정 역시 사용자에게 설명 가능하여야 한다. 아무리 우수한 성능을 보여주는 예측모형이라 할지라도, 이유를 모르는 결정을 따를 임상의사는 거의 없다고 보아야 한다[2]. 이런 면에서 신경망이나 서포트벡터머쉰(SVM, Support Vector Machine)과 같은 black-box 모형은 임상현장에서 받아들여질 가능성이 상대적으로 낮다고 할 수 있다.

III. 데이터 마이닝 수행절차

데이터마이닝 수행절차는 연구자에 따라서 다양하게 정의된다. Cios 등[1)15)16)]의 6단계 분류에 의하면 1) 대상 영역의 이해, 2) 데이터의 이해, 3) 데이터 준비, 4) 데이터 마이닝, 5) 평가, 6) 발굴한 지식의 사용으로 나눌 수 있으며, 한 번의 흐름으로 수행된다기 보다는 전체 과정을 지속적으로 반복하는 일련의 과정으로 볼 수 있다.

1. 대상영역 및 데이터의 이해

기계학습의 여러 알고리즘을 이용하여 준비된 데이터로부터 지식을 발굴하는 단계는 네 번째 단계인 데이터 마이닝 단계이겠으나, 가장 어렵고 노력이 많이 드는 단계는 첫 세 단계로서 전체과정의 60% 이상의 시간과 노력이 소모된다. 특히 연구주제가 바뀔 때마

다 해당 임상분야에 대하여 다시 새롭게 공부하여 이해하여야 하기 때문에, 심지어 의료전문가라 할지라도 매우 힘들고 어려운 단계라고 할 수 있다. 연구 대상영역을 이해하기 위하여 처음 문제를 제기한 임상전문가와 여러 차례의 예비모임과 본 모임을 가져야 한다. 관련분야의 교과서적인 내용을 공부하여 기초적인 용어와 병태생리학적 및 임상적 진단, 치료와 예후 등에 대하여 이해할 필요가 있다. 각 변수들의 단위와 참고치 및 이상치, 변수타입에 대하여 정리하여야 한다.

OCS나 EMR, 또는 임상데이터웨어하우스(Clinical Data Warehouse; CDW)에서 자료를 수집할 경우에는 Entity-Relation Diagram(ERD)과 테이블 정의서를 입수하여 관련된 자료가 어떤 테이블에 어떤 형태로 존재하는지 파악하여야 한다. 최근에 병원에 도입되고 있는 데이터웨어하우스는 전략적 의사결정을 지원하기 위한 전사적 관점에서, 동일한 데이터베이스 스키마 하에 여러 다양한 정보 출처로부터 수집된 정보의 종합 저장소라고 할 수 있다. 데이터웨어하우스 구축은 다양한 출처로부터 추출된 데이터를 정재, 통합, 변환, 적재하여 데이터웨어하우스를 구축한 후 주기적인 갱신과정을 거쳐서 이루어진다. 국내 병원에는 2004년부터 데이터웨어하우스가 도입되기 시작하였으나, 임상적인 연구를 위한 목적보다는 경영정보구축을 목적으로 하는 경우가 많았다. 최근에 들어서 EMR의 도입과 함께 임상연구를 지원하기 위한 CDW의 구축이 늘고 있는 실정이다.

2. 데이터 준비

대부분의 데이터 마이닝 전공서적은 기계학습알고리즘과 성능평가 부분에 초점을 맞추어 기술되어 있지만, 실제 문제가 발생하는 부분은 대부분 데이터 준비단계라고 할 수 있다. 이 과정은 크게 보아서 데이터 탐색과 데이터 정재로 나누어 볼 수 있지만 서로 분리된 과정이라기보다는 상호 반복되는 과정이다. 예를 들어, 분석의 원천이 되는 원자료는 통상 병원정보시스템에 시간의 흐름에 따라 데이터베이스관리시스템에 기록되는 트랜잭션 데이터인 경우가 많다. 이때 필드의 특성에 따라 수치형, 날짜형, 문자형이 구별되어서 기록되어야 하나 통상적인 병원정보시스템에서는 어떤 내용이라도 오류 없이 기록될 수 있도록 문자형으로 정의되어 있는 경우가 흔하다. 심지어 같은 필드 안에 숫자형과 문자형이 혼재되어 있는 경우도 있다. 분석을 위하여 SQL명령문 중 CAST나 CONVERT와 같은 형변환 문법으로 형을 변환시키게 되지만, 레코드 내에 문자가 존재할 경우 형변환이 안되어 오류가 발생하게 된다. 분석 대상 레코드 수가 적을 경우에는 수작업으로 변환작업을 수행하지만 레코드 수가 많을 경우 자동화된 방법 이외에는 대안이 없게 된다. 다음과 같은 SQL 쿼리문(MS-SQL기준)이 도움이 될 수 있다. 이후에 형변환이 되지 않는 레코드만 찾아서 수작업으로 변환하여야 한다: SELECT * FROM 테이블 WHERE ISNUMERIC(필드)=1 AND CAST(필드 AS NUMERIC(6,3))

또 다른 예로, 하나의 검사처방이 데이터베이스에 기록될 때에는 나뉘어져서 여러 개의 sub-order로 기록되는 경우(예: ‘CBC with differential count’)가 흔하므로, 다루고 있는 검사가 병원정보시스템 내에서 어떻게 기록되고 실행되는지 조사하여야 한다. 또한 약처방의 경우 투여된 약품의 용량을 계산할 때 반환된 약물을 고려하지 않을 경우 과처방으로 계산될 수 있으므로 주의하여야 한다. 반환처방의 경우 병원정보시스템마다 다르겠으나 투여된 약물용량이 음수이며, 원처방에 대한 고유번호를 가지고 있게 되므로, 원처방번호를 찾아서 서로 상계하여야 한다.

시간의 흐름에 따라 생성된 데이터를 각 환자별로 필드와 레코드를 치환하여 하나의 레코드에 한 명의 모든 필드가 존재하도록 변환하는 과정이 필요하게 된다. 단순한 형태의 경우 PIVOT을 이용할 수도 있으나 데이터베이스관리시스템에서 제공되는 간단한 명령문으로 해결되는 경우는 많지 않기 때문에, SQL의 cursor를 이용하거나 프로그래밍을 통하여 해결해야 하는 경우가 많다. 시간의 흐름에 따라 반복 측정하는 동일한 필드를 연속으로 표현해야 하는 경우에는 offset과 같은 기능을 이용하여 쉽게 해결할 수도 있다.

원자료에서 특정 값을 찾아서 채택할 때, 당일에 반복 측정한 자료가 여러 개 있을 경우, 예로써 환자의 신장기능을 모형에 넣기 위하여 creatinine값을 변

수로 채택하는 경우, 만일 당일에 여러 번의 측정치가 존재할 경우 시간적으로 마지막 값을 고를 것인지, 당일 측정치들의 평균값을 채택할 것인지 또는 중위수를 채택할 것인지에 대하여 사전에 결정하여야 한다. 임상상황에 따라서 최대값 또는 최소값을 채택하여야 하는 경우도 있다. INR(International Normalized Ratio)은 경구항응고제(warfarin) 복용환자의 혈액응고상태를 감시하기 위한 검사방법이다. INR의 목표치는 2-3 사이이나 당일에 반복 측정한 검사 값에서 8 이상의 높은 값이 기록되고 연이어 검사한 두 번째 측정치는 2-3 사이의 목표치를 보이는 경우가 있다. 이 경우는 과다한 경구항응고제 투여 상태를 되돌리기 위한 해독제인 vitamin K가 투여됐을 가능성을 고려하여야 한다. 해독제가 투여된 상황에서 검사값의 평균을 취할 경우 당시의 임상 상황을 전혀 다르게 해석하는 오류에 빠질 수도 있다.

3. 데이터요약

데이터 탐색은 기술적 데이터요약(descriptive data summarization)과정이라고 볼 수 있다. 기술적 데이터 요약은 대상 데이터의 일반적 특징을 살펴보는 것으로서 오류나 이상치를 발견하는데 도움이 된다. 데이터의 특징을 나타내는 대표적 방법으로 평균이나 중위수, 최빈값, 최소값, 최대값, 분산, 사분위범위 등이 유용하게 이용된다. 첨도나 왜도를 살펴봄으로써 분포의 대칭성이나 중심성향을 살펴볼 수 있지만, 히스토그램과 같은 그래프를 이용한 방식이 훨씬 직관적인 경우가 많다(Fig. 1).

Data Audit of [7 fields]

File Edit Generate

Field	Graph	Type	Min	Max	Range	Mean	Std. Dev	Skewness	Kurtosis	Valid
AGEYE...		Range	0	94	94	57.077	17.772	-0.760	0.251	376
CrCL		Range	9	249	240	101.585	39.877	0.361	0.483	349
Fibrinog...		Range	134	1484	1350	613.877	252.872	0.581	0.515	146
HEIGHT		Range	50.000	185.000	135.000	162.068	14.901	-3.829	22.935	225
SEX		Flag	--	--	--	--	--	--	--	376
WBC		Range	0	125	125	11.878	9.369	5.713	59.517	369
WEIGHT		Range	6.100	104.900	98.800	59.859	13.094	-0.218	2.714	225

Figure 1. An example of descriptive summary of some selected features

Data Audit of [7 fields] #1

File Edit Generate

Complete fields (%): 28.57% Complete records (%): 23.94%

Field	Type	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null...
AGEYEAR	Range	2	0	None	Never	Algorithm	100	376	0...
SEX	Flag	--	--	--	Never	Fixed	100	376	0...
Fibrinogen	Range	2	0	None	Never	Algorithm	38.83	146	230...
WBC	Range	3	1	None	Never	Fixed	98.138	369	7...
CrCL	<Default>	--	--	--	--	--	92.819	349	27...
HEIGHT	Range	2	2	None	Never	Fixed	59.84	225	151...
WEIGHT	Range	5	0	None	Never	Fixed	59.84	225	151...

Figure 2. Data quality and outlier summary

4. 이상치 검출 및 결측치 처리

이상치 검출을 위하여 box-plot(box-and-whisker diagram)과 같은 그래프를 이용한 방식과 통계적 방법이 같이 사용된다(Fig. 2). 이상치의 경우 임상적인 의미를 파악하여야 하므로 전 단계에서 임상적 지식을 필요로 하게 된다. 이상치를 파악할 때, 변수 하나만을 대상으로 수행하여서는 찾지 못하는 경우도 많다. 예로서 '키'를 조사할 경우 '65(cm)'라는 값이 있을 경우 이 수치만으로는 이상치 유무를 판단할 수 없으나, 키나 체중, 연령을 혼합하여 비율을 계산하여 이상치를 찾을 수도 있다.

자료에 포함된 결측치를 처리하는 데에도 주의가 필요하다. 관행적으로 결측치를 '999', '9999'로 코딩하는 경우가 있으므로 해당하는 숫자가 발견될 경우 정확한 의미를 파악하여야 한다. 결측치가 포함된 자료를 분석함에 있어서 여러 변수들 중에 결측값이 하나만 있어도 해당 증례를 제거하는 'complete case analysis(혹은 case deletion)'방법이 있다. 이 방법은 일변량 통계량 비교가 가능하며 간편하지만, 많은 표본수의 감소로 인해 대표성이 상실되어 검정력이 약화되는 단점이 있다. 각각의 변수에 사용 가능한 자료를 이용하는 'available case analysis'는 표본수 확보에 유리하지만, 표본의 기저가 분석마다 변하기 때문에 모수 추정시 수학적 문제 발생할 수 있다. 결측치를 대치하는 방법으로 평균이나 회귀분석의 예측값으로 대치하는 단순대치법과 복잡한 통계적 계산에 의한 다중대치법(multiple imputation)이 있다[17-19]. 전체적으로 결측치가 5%미만일 경우 complete case analysis를 사용할 수 있으며, 결측치가 15-20% 범위일 때에는 대치법을 이용하는 것이 좋다. 그러나, 결측치가 40%를 넘어서는 경우에는 해당 변수를 모형에 사용하기에는 무리가 있다고 보아야 한다.

Ⅳ. 데이터 마이닝 기법

데이터 마이닝에 주로 사용되는 기법으로 로지스틱 회귀분석, 신경망[20], 서포트벡터머쉰[21], 의사결정나무[22], 베이지안 네트워크[23-25], 연관성 규칙, K-평균, 코호넨 네트워크 등과 그 외 다양한 비교사학습방법(unsupervised learning)들이 있으며[26-28], 다음에서 그 중 일부를 간략히 살펴본다.

1. 로지스틱 회귀분석(Logistic Regression)

로지스틱 회귀분석은 전통적인 통계기법으로 이미 분류(Classification)와 예측(Prediction)에서 높은 정확도를 인정받은 기법이다[29]. 대용량의 입력변수를 다루기에는 여러 가지 제한이 따르지만[30][31], 이미 검증된 모델링 기법이므로 데이터 마이닝에서 수준점(Benchmark)으로 삼기 위해서 많이 사용된다.

2. 신경망(Neural Networks)

인간의 신경망 구조와 비슷한 원리에 의한 방법이며, 복잡한 구조를 가진 자료에서의 예측(prediction)과 분류(classification) 문제를 해결하기 위해서 사용되는 비선형모형(nonlinear models)의 하나이다. 인체의 신경망 구조에서 neuron에 해당하는 부분이 Figure 3에서의 Processing Element(PE)이며, PE는 많은 dendrites를 가지고 있고 입력변수(input variable)가 이에 해당된다. PE의 주요 기능은 입력변수를 통해 들어오는

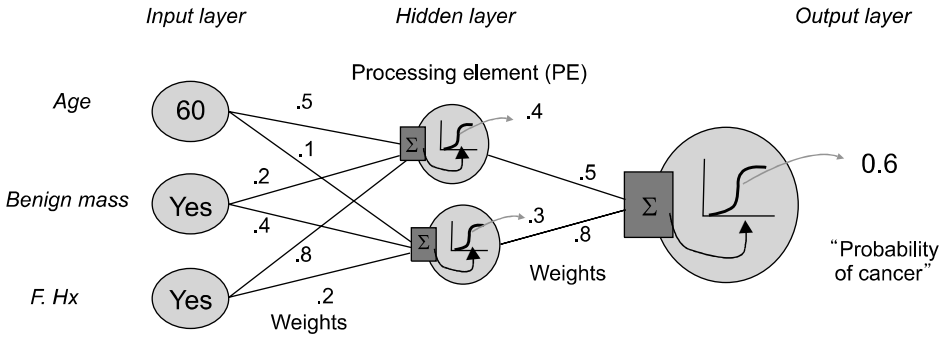

Figure 3. An example topology of artificial neural network model

정보에 대해 가중치가 적용된 summation function을 수행하고, 그 결과를 다시 activation function을 통해 수정하여 output layer로 보내는 것이다. Activation function의 예로는 logistic, linear threshold, hyperbolic tangent 등이 있다.

Figure 3의 예제는 암발생 예측을 위해 가상적으로 설정한 신경망 모델의 예이다. 이 모델에 의하면 연령(age)이 60세이고, 양성종양(benign mass) 과거력이 있고, 가족력(F. Hx)이 있는 경우 암 발생 위험률이 0.6으로 예측되었다. 이와 같은 위험률 예측은 기존 데이터를 통해 학습(machine learning)된 신경망 모델을 통해서 이루어진다. 인간의 학습능력과 기억력을 조절하는 신경전달 물질의 양이 어느 정도 필요한지에 대해 알려지지 않은 것처럼, 신경망 모델에서 각각의 input variable에서 제공되는 정보의 양과 예측과정은 쉽게 이해 할 수가 없다. 이러한 단점을 'Black Box' syndrome이라고 칭하기도 한다. 하지만 신경망 기법은 보건의료계에서 예측모델 개발에서 탁월한 예측력을 보여주어 왔다.

3. 서포트벡터머쉰(SVM, Support Vector Machine)

SVM은 classification이나 regression 기법으로 사용되며, 훈련데이터의 과적합(overfitting)없이 모델의 예측력과 정확도를 최대화 시켜주는 기술이다. 특히 의생명자료와 같이 레코드 숫자는 적으며 변수의 숫자가 많은 대용량 데이터(예: 수천 개 이상의 입력변수)를 분석하는데 적합하다[32)33)].

SVM은 Kernel 함수(Linear, polynomial, Radial basis function, sigmoid 등)를 통해 데이터의 선형변환뿐 아니라 비선형 변환과정을 거쳐 개체들을 최적으로 분류해주는 초평면(hyperplane)을 찾아내는 방법이다. 초평면을 중심으로 두 범주를 나누어주는 개체를 서포트벡터(support vectors)라고 하며 이 것들의 거리를 최적화 시켜 주는 것이 중요하다. Support vector들 간의 거리가 적을수록 모델은 과적합 될 수 있으며, 클수록 오분류되는 개체가 많아질 수 있기 때문이다. 그러므로 예측력이 가장 높은 최적의 모델은 관심 데이터를 대상으로 서로 다른 Kernel 함수를 적용하는 실험과정을 통해 얻을 수 있다.

4. 의사결정나무(Decision Trees)

의사결정나무는 Figure 4의 가상 시나리오 모델에서와 같이 의사결정규칙(decision rule)을 나무구조로 도표화하여 분류(classification)와 예측(prediction)을 수행하는 분석방법이다. Figure 4의 예는 첫번째 노드(node)에서는 분류기준 없이 모든 병원 입원 대상자 중 암환자가 차지하는 비율이 20%이었다. 그러나 두 번째 노드에서는 연령이 50세 이상인 경우 암환자는 35%로 증가되었다. 이와 같이 의미 있는 변수를 기준으로 계속적인 분류를 통해 암환자일 가능성 높은 군을 가려 내는 모델을 형성하게 된다. 즉 이 모델에 의하면 50세 이상이고 양성종양이 있고, 가족력이 있는 경우에는 암환자로 분류되는 경우가 20%에서 60%로 증가된 예를 보여준다. 이와 같이 의사결정나무 모델은 분류 또는 예측 과정이 나무구조에 의한 추론 규칙에 의해서 표현되기 때문에 의사결정과정을 쉽게 이해하고 설명할 수 있는 장점을 가지고 있다. 의사결정나무 분석을 위해 자주 사용하는 알고리즘은 C5.0, C&RT(Classification & Regression Tree), CHAID(Chi-squared Automatic Interaction Detection)이다. C5.0은 entropy나 정보이득(information gains)을 산정하여 가장 유의한 속성이 높은 변수 순서대로 가지를 쳐서 나무구조를 만들어 나간다.

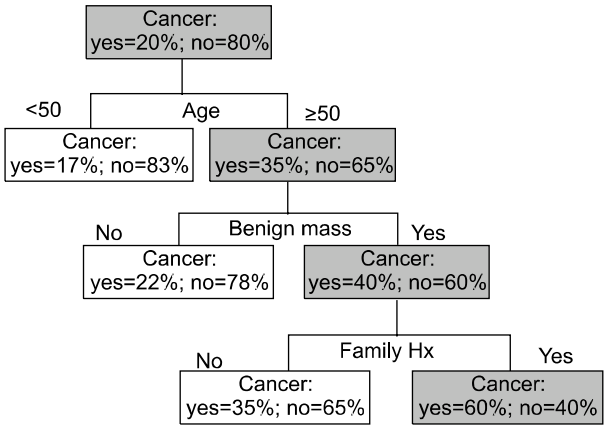

Figure 4. An examples of decision tree model

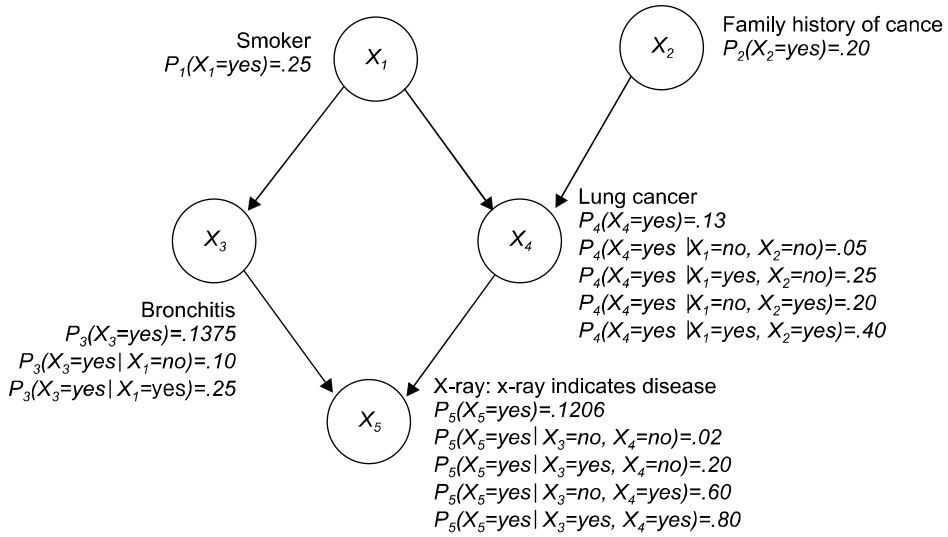

Figure 5. An example of bayesian network model

5. 베이지안 네트워크(Bayesian Network)

Figure 5는 베이지안 네크워크를 설명하는데 자주 인용되는 대표적인 예제이다[30)34)]. 이 모델은 다섯 개의 노드(node)로 표현되고 있으며, 각각의 근원 노드(root nodes/parents nodes; X_1 *smoker*와 X_2 *family history*)는 사전확률(prior probability) 분포만을 갖고, 비근원 노드(non-root nodes/child nodes; X_3 *bronchitis*, X_4 *lung cancer*, X_5 *X-ray*)는 사후확률(posterior 혹은 conditional probability) 분포도 가지고 있다. 예를 들면 사전 확률 $P(X_3)$은 환자가 흡연자(smoker)인지 아닌지 모를 경우 13.75%의 환자가 기관지염을 가질 확률을 뜻한다. 그러나 X_1에 대한 새로운 정보를 갖게 되었을 때 $P(X_3)$ 대신 X_3의 조건부 확률, $P(X_3|X_1)$로 표현되고 이는 흡연자인 경우 기관지염(bronchitis)일 가능성은 25%, 흡연자가 아니라면 10%라는 것을 의미한다. 폐암일 가능성을 예측하는 데 있어 베이지안 네트워크 모델의 임무는 관찰자료가 주어졌을 때 목표 노드(target node)인 폐암의 사후 확률을 추정하는 것이며, 사후 확률은 chain rule과 Bayes' rule에 의해 추정된다[26)].

베이지안 네트워크는 다음과 같은 장점이 있다.

1) 시각적 다이어그램으로 표현되므로 다른 기법보다 쉽게 해석할 수 있다. 이는 노드 사이의 상호작용을 쉽게 파악할 수 있기 때문이다; 2) 다른 기법은 일차적으로 데이터에만 의존해 예측모형을 개발할 수 있는 데 반해 베이지안 네트워크를 사용하면 연구자들의 전문지식을 모델 개발 과정에 접목시킬 수 있다. 즉 확률을 추정하는 데 있어 전문가의 지식(사전지식; prior knowledge)과 데이터에서 얻은 확률을 통합할 수 있다; 3) 2)의 이유로 표본이 상대적으로 적은 경우에도 모델링이 가능하다[35)].

6. 연관성 규칙(Association Rule)

If-then으로 표현되는 연관성 분석 결과는 이해하기 쉽고 적용하기도 용이하다. 주로 사용되는 것들로 Apriori와 GRI(Generalized Rule Induction)가 있다. 또한 시간에 의한 순서를 고려한 규칙을 찾아주는 알고리즘으로 순차규칙(Sequence)이 있으며, 시계열 자료에 유용하다. 연관성 규칙 알고리즘은 주로 지지도(support)와 신뢰도(confidence)를 산정하여 중요 규칙을 가려주게 된다.

7. 코호넨 네트워크(Kohonen Network)와 K-평균(K-Means)

코호넨 네트워크(Kohonen Network)와 K-평균은 비교사학습의 형태로 군집분석(clustering)을 수행하는 방법이다. 주어진 관찰치를 유사한 것들로 몇몇의 집단으로 묶은 후 각 집단의 성격을 파악함으로써 데이터 전체의 구조에 대한 이해가 가능한 분석방법이다. 대용량 데이터에서 개개의 관찰치를 요약하는 것보다

전체를 유사한 관찰치들의 군집(cluster)으로 구분하여 관찰함으로써 전체 데이터에 대한 의미 있는 정보를 요약적으로 얻을 수 있게 된다. 코호넨 네트워크는 신경망 알고리즘의 일종으로 비슷한 데이터끼리 모으는 군집모델에 사용된다. K-평균은 distance measures(예: Euclidean Distance)를 사용하여 각 군의 대표값과 거리가 가장 가까운 관찰치들을 같은 그룹으로 합치면서 군집들을 형성하는 방법이다.

Ⅴ. 모델 평가

모델의 성능 평가법은 discrimination과 calibration 능력을 측정하는 방법으로 분류할 수 있다. Discrimination 능력은 예측모델이 클래스를, 예를 들면 환자를 정상인으로부터, 얼마나 잘 분리해 내는가를 평가하는 것이고, calibration 능력은 예측치(predictive value)가 실제 결과(real outcome)에 얼마나 근접했는가를 평가하는 방법이다. 정확도는 대표적인 discrimination 방법으로서, 오차행렬(Confusion Matrix)을 이용하면 정확도와 민감도, 특이도, 양성예측도(Positive Predictive Value; PPV), 음성예측도(Negative Predictive Value; NPV)를 쉽게 계산해 낼 수 있다. 그런데, 정확도만으로 모델의 성능을 검정할 수 없다는 점에 주의하여야 한다. 의료 데이터는 'class imbalance', 즉 클래스간 데이터 개수의 불균형이 심한 경우가 많다. 예로써, 특정 암에서 예후가 좋지 않은 군과 예후가 좋은 군이 10 : 1의 비율이라고 가정하면, 분류모형이 모든 증례를 예후가 좋지 않다고 분류하여도 정확도는 90%가 될 수 있다. 또 이렇게 불균형이 심한 자료의 경우에는 모델 훈련시에도 비율이 낮은 클래스가 충분히 훈련되지 않을 수 있다. 따라서 모델 훈련시에 클래스간에 균형을 이루도록 넘치는 증례를 줄이거나 또는 부족한 증례를 늘려서 훈련을 함으로써 낮은 비율의 클래스가 충분히 학습되도록 하거나, 소수 클래스에 대해 잘못된 분류시 추가적인 벌점을 부과하는 오분류비용(miss-classification cost)을 이용해 볼 수도 있다.

Receiver Operating Characteristic(ROC) curve 분석은 정확도와 달리 유병율에 따라 계산 값이 달라지지 않는 장점이 있다[36)37)]. 각 예측모델의 ROC curve 분석에서 계산된 AUC(Area Under the ROC Curve)가 통계적으로 유의한 차이가 있는지 비교하기 위해 Hanley와 McNeil[36)37)]이 제시한 방법을 사용할 수 있다. 임계점(best cut-off point)을 (1－민감도)^{2}과 (1－특이도)^{2}의 합이 최소화되고 ROC curve에서 (0, 1)에 가장 가까운 지점으로 정하여 정확도, 민감도, 특이도, 양성예측치, 음성예측치를 구하여 분류기 간의 성능을 비교할 수 있다[38)].

서로 다른 예측기(predictor)의 정확도를 측정하는 방법으로는 예측값이 결과값에 얼마나 근접했는지를 측정하기 보다는 예측값이 결과값에 얼마나 떨어져 있는지를 측정하는 방법을 쓰게 되며, 손실함수(loss function)는 예측값과 결과값 간의 차이를 측정하는데 사용된다. 대표적인 방법으로 'mean absolute error', 'mean squared error', 'relative absolute error', 'relative squared error'를 들 수 있다.

$$\text{Mean absolute error: } \frac{\sum_{i=1}^{d} |yi - y_i'|}{d}$$

$$\text{Mean squared error: } \frac{\sum_{i=1}^{d} (yi - y_i')^2}{d}$$

$$\text{Relative absolute error: } \frac{\sum_{i=1}^{d} |yi - y_i'|}{\sum_{i=1}^{d} |yi - \overline{y}|}$$

$$\text{Relative squared error: } \frac{\sum_{i=1}^{d} (yi - y_i')^2}{\sum_{i=1}^{d} (yi - \overline{y})^2}$$

yi: 실측값, *yi'*: 예측값, *y*: *yi*의 평균값

마지막으로 Calibration 능력은 Hosmer-Lemeshow goodness-of-fit Chi-square statistic으로 판정할 수 있다[39)40)]. 실무에서 모델의 효용성을 평가하기 위해 흔히 사용하는 방법으로는 Lift측정법이 있다. 이는 예를 들면 고위험군을 가려낼 수 있는 능력이 예측 모델이 없는 경우보다 어느 정도 효과적인지를 측정하는 방법이다.

분류기 또는 예측기 간의 성능을 평가하기 위해서 전체 데이터 중에서 일부를 이용하여 모델을 만들고 나머지 데이터로 모델을 평가하게 되며, 이러한 방법으로 'holdout', 'random subsampling', '*k*-fold cross-validation', 'bootstrap'법 등을 들 수 있다. 임상에서

구하는 자료는 흔히 데이터의 건수가 많지 않은 경우가 많아서, 주어진 자료를 훈련데이터와 검증데이터로 나눌 경우 충분한 검정력을 확보하기 어려운 경우가 많다. 이런 상황에서 자료를 최대한 이용하기 위해 주로 사용하는 방법이 *k*배 교차 검증법(*k*-fold cross-validation method)이다. *k*배 교차 검증법이란 데이터를 *k* 등분해 이 중 *k* 분의 *k*-1은 훈련데이터로 사용하고 나머지 *k* 분의 1을 검증데이터로 사용하여 모델의 예측력을 평가하는 방법이다. 총 *k* 차에 걸쳐 실시되어 모든 대상자가 한 번씩 검증 데이터셋에 속하고, 이를 통해 모든 대상자에 대한 예측치를 얻게 된다. 통상 10-fold cross-validation method가 모델 평가의 표준으로 자리 잡고 있다.

VI. 소프트웨어

예전에는 보건의료인들의 데이터마이닝 관련 연구가 알고리즘 개발에 중점을 두어왔다. 그러나 최근 10여 년 동안 연구자들이 쉽게 데이터마이닝 기법을 활용하여 데이터를 분석할 수 있도록 많은 소프트웨어들이 개발되었으며, 상업적으로 개발된 프로그램뿐만 아니라 freeware나 shareware도 인터넷을 통해 쉽게 접할 수 있게 되었다. 대표적인 상용 데이터 마이닝 툴은 SPSS사에서 나온 PASW modeler(예전의 Clementine)와 SAS사의 SAS Enterprise이다. 무료 소프트웨어로서 대표적으로 WEKA(http://www.cs.waikato.ac.nz/~ml/weka)를 들 수 있다. 이 소프트웨어는 상용 소프트웨어들이 일부 대표적인 기계학습 알고리즘만을 기능에 포함하고 있는데 비하여, 기존에 알려진 대부분의 알고리즘을 포함하고 있으면서 특징 선택에서 모델평가에 이르기까지 데이터 마이닝에 필요한 대부분의 기능을 가지고 있어서 학술적 목적으로 사용하기에 대단히 유용하다. 베이지안 네트워크 관련 소프트웨어로는 Hugin Expert(www.hugin.com), Netica(www.norsys.com) 등을 들 수 있다. 국내에서 개발된 프로그램으로는 ECMiner(www.ecminer.com)가 있다. R-Project나 Matlab 등은 기계학습을 위한 전용소프트웨어는 아니지만 대부분의 기계학습 알고리즘을 구현할 수 있는 확장 패키지를 가지고 있다. 흥미로운 것은 일부 데이터베이스관리시스템들도 몇 가지 기계학습 알고리즘을 구현하고 있다는 점이다. 데이터 마이닝을 수행하는 과정에서 한 가지 소프트웨어만을 이용하여 데이터 전처리에서 모델평가까지 모두 수행할 수도 있지만, 각 소프트웨어의 특징과 장점을 파악하여 단계별로 잘 조합하여 사용하는 것이 효율적인 경우가 많다.

VII. 의료에서의 사용례

의료분야에서 데이터 마이닝이 사용된 초기의 예로서 1980년대 초에 스탠포드 대학교에서 수행된 RX 프로젝트를 들 수 있다[41]. 연구자들은 ARAMIS기반의 데이터베이스를 이용하여 50명의 전신홍반루프스 환자들을 50회 이상 추적관찰한 데이터베이스를 이용하여 여러 임상변수와 결과들간의 시간적인 선후관계와 통계적 상관성을 연구한 바가 있다. 이후로 종양학[42], 간병리[43], 갑상샘질환[44-46], 류마티스[47], 심장[48], 신경정신, 부인과, 산과학, 전립샘암[49-58] 등 의학진단의 전 분야에 걸쳐서 기계학습이 적용된 바가 있다. 이처럼 의료에서 데이터 마이닝 기법을 이용하여 진단이나 예후를 예측한 연구는 매우 많이 있으나, 본 고에서는 그 중에서 낙상예측이나 자살예방을 중심으로 하여 극히 일부만을 언급하도록 한다.

Giles 등[59]은 로지스틱 회귀분석 방법을 이용하여 정보시스템에서 주로 사용되는 간호진단을 사용하여 낙상을 예측해주는 관련 간호진단을 찾아내어, 위험요인이 되는 각 간호진단별 낙상 위험률을 산정하여 낙상 예측 모델을 개발하였다. 총 10개의 간호진단이 선정되었으며, 가장 높은 오즈비는 'Urinary incontinence management'로 6.63이었고, 그 다음으로 'Risk management potential for falls', 'Care of the patient with impulsive behavior' 등이 선정되었다. Tiet 등[60]은 약물중독자들 중 자살시도를 생각해본 적이 있는 5,671명을 대상으로 실제적으로 자살을 시도할 위험이 높은 환자군을 찾아내기 위해서 의사결정나무 기법을 사용하였다. 의사결정나무 분석 결과 지난 30일간 자살을 시도를 생각한 적이 있는 환자들 중 자살위험이 가장 높은 환자군은 그 전에 자살을 실제로 시도한 적이 있었고, 지난 30일 내에 만취한 적이 있으며, 지난 3년 간 전일제 근무 경력이 있는 환자들이

었다. Modai 등[61]은 신경망 분석 기법을 이용하여 자살 위험이 높은 사람을 예측하기 위한 모델을 구축하였으며, 47개의 독립변수 중 6개의 주요 예측변수(독거, 치료에 불응, 약물중독이나 의존성향, 저하된 기능수준, 피해망상이 없는 경우, 직계가족 중 자살한 사람이 있는 경우)가 모델을 통해 선정되었다. 모델의 예측 정확도(accuracy)는 0.92로 높은 결과를 보였다. Anthony 등[62]은 신경망과 로지스틱 회귀분석을 이용하여 욕창 위험을 예측하는 모델을 구축하였고, 모델의 AUC(Area Under the Receiver Operating Characteristic Curve)는 각각 0.87과 0.85로 높은 결과를 보였다. Rapeli와 Botega[64]는 K-평균기법을 활용하여 자살을 시도하는 사람들의 특성을 파악하였다. 자살 시도자들의 특성은 세 그룹으로 나뉘었다: 1) 자살충동 그룹(주로 여자이고, 독극물로 자살 시도, 대부분 자살을 시도한 적은 있었으나, 자살 의도는 높지 않음), 2) 자살성향 그룹(남자, 모두 치명적인 독극물로 자살시도, 자살 경력이 적고, 자살 의도가 높은 편), 3) 자살 의지적 그룹(대부분 남자, 중장년층, 대부분 독신이며 자살 경력 없고, 치명적이고 폭력적인 방법에 의한 자살 시도, 매우 높은 자살 의지, 소생을 위해 수술이 필요한 경우가 대부분). Brossette 등[63]은 연관성 규칙을 이용하여 항생제 사용과 관련된 규칙을 병원처방 데이터에서 찾아내는데 활용하였다. 연구결과에서 보고된 주요 규칙 중 하나는 "R~Piperacilline → R~Imipenem"이다. 즉, Pseudomonas Aeruginosa 감염 환자에서 "Piperacilline에 resistance하면, Imipenem에 resistance하다"는 규칙이 생성되었다.

VIII. 발전방향 및 결론

현재 한국의 병원은 전문요양기관을 중심으로 EMR과 CDW의 도입 등 의료정보화사업이 경쟁적으로 이루어지고 있다. 일찍 시작한 일부 대형병원에서는 이미 엄청난 임상데이터가 쌓이고 있고, 이를 활용하기 위해서 많은 임상전문가와 정보전문가들이 깊은 관심을 보이고 있다. 그러나 이러한 의료데이터가 가지는 복잡한 성격을 이해하지 못할 경우 비현실적인 기대감을 유발하거나, 또는 자칫 잘못된 결과를 도출할 수도 있다.

데이터 마이닝과 관련된 논란 중의 하나는 가설검정 없이 결과를 일반화하는 데이터 마이닝에 대한 전통적인 추론통계를 다루는 데이터 분석가들의 비판이다. 그러나 데이터 마이닝의 기본 전제는 표본분석이 아니라 모집단 데이터를 다루는 것이므로 이러한 이견을 반박하기도 한다. 통계학자들은 데이터 마이닝이 데이터에서 원하는 정보를 얻기 위해 데이터를 피싱(fishing)을 하는 방법이라고 비난하기도 한다. 그러나 데이터 마이닝이 기존 연구방법과 가장 크게 다른 점은 데이터가 존재하고 그 안에서 연구문제(research problem or question)를 생성한다는 것이며, 그 안에서 연구문제에 대한 답을 얻을 때까지 데이터를 반복해서 분석하는 것은 당연하다는 것이다.

병원 정보시스템을 설계하고 구현하는 과정에서 '좋은 자료가 수집되어 쌓일 수 있도록' 충분히 고려하여야 한다. 데이터 마이닝에서 올바른 데이터 항목을 규명하고 데이터를 전처리하는 과정에서 전체 시간과 노력의 60% 이상(심지어 70-80% 이상)이 소모되나, 그럼에도 불구하고 잘못된 수집된 데이터를 되돌릴 방법이 없다는 점을 깊이 생각하여야 한다. "Garbage in, Garbage out"은 데이터 분석의 원칙으로 아무리 많은 데이터가 쌓여있어도 쓸모 없거나 활용할 수 없는 형태로 수집되고 있다면 데이터 마이닝은 불가능하다는 점을 강조하는 말이다. 그러므로 병원정보시스템 구축 단계부터 수집되는 데이터를 향후 어떻게 이용할 것인가에 대한 계획을 포함하여야 한다. 만일 업무편의를 위하여 각종 측정장비의 결과치를 이미지로 스캔 하여 저장하기로 한다면, 아무리 많은 자료가 쌓여도 데이터 마이닝이 분석에 기여할 바는 거의 없을 것이다.

생성된 모델이나 규칙, 패턴 등의 '지식'을 임상의 사결정지원시스템(Clinical Decision Support System)의 형태로 활용하는 것에 대하여, 병원정보시스템 관리자들은 시스템의 과부하를 염려하고 있으며, 한편 임상전문가들은 번거로움과 불투명한 의사결정구조 등으로 인해 비관적인 또는 무관심한 입장을 취하고 있다. 아무리 좋은 지식이 데이터 마이닝에서 생성되었다 할지라도 임상에서 사용될 수 없다면 무용지물이 될 것이다. 정보시스템의 부하를 최소화 할 수 있는 방법, 사용자인 임상실무자들의 업무흐름 방해를 최

소화 할 방법, 의사결정과정을 투명하게 하여 설명력을 올리는 일 등이 데이터 마이닝을 의료계에서 활성화시키는 데 중요한 현안이 될 것이다. 또한 데이터 마이닝의 한계점을 해결할 많은 연구가 수행되어야 한다. 예를 들면, 임상 데이터의 특성상 앞에서 논의된 바와 같이 class imbalance가 있는 경우가 흔하며 이를 해결하기 위한 여러 시뮬레이션 연구가 수행되었지만 아직까지 신통한 해결책은 없는 실정이다.

EMR과 PACS의 보급으로 텍스트 및 이미지 데이터가 급격히 증가하고 있으나 자연어 처리를 이용한 의무기록분석이나 영상분석은 아직 초보적인 단계에서 벗어나지 못하고 있는 실정이다. 기존에 국내에서 주로 관심을 가진 데이터 마이닝 영역도 수치형 데이터를 이용한 분류, 예측, 규칙 발견 등이 주를 이루고 있다. 의무기록분석과 영상진단 분야에서, 혹은 욕창, 상처, 환자의 비정상적 행동이나 특정 문제를 발생 시킬 수 있는 외모의 변화에 대한 패턴 분석 등 text mining과 영상분석에서 활발한 연구가 진행되어야 할 것이다.

EMR과 CDW의 도입으로 급격히 쌓이는 임상자료를 어떻게 잘 이용할 것인지 많은 논의가 진행되고 있다. 무엇보다도 EMR 등 병원정보시스템 도입 초기 단계부터 질 좋은 정보가 진료업무의 '부산물'로 축적될 수 있도록 설계에 잘 반영하여야 할 것이다. 분석대상이 되는 의료자료를 잘 이해하면서 또한 동시에 고도의 정보학적 분석능력을 보유하기란 쉬운 일이 아니므로, 임상연구자들과 IT실무자, 그리고 데이터 마이닝 전문가들간의 상호존중과 긴밀한 협조가 성공적인 임상 데이터 마이닝의 필요조건이라고 하겠다.

Acknowledgement

The authors would like to acknowledge Professor Kyi Young Lee, Dept. of Biomedical Informatics, School of Medicine, Ajou University for his detailed review of the manuscript.

참고문헌

1. Cios KJ, William Moore G. Uniqueness of medical data mining. Artificial Intelligence in Medicine 2002; 26(1-2):1-24.
2. Lavrac N, Keravnou E, Zupan B. An overview. In: Lavrac N, Keravnou E, Zupan B, editors. Intelligent data analysis in medicine and pharmacology. Boston: Kluwer;1997. pp.1-13.
3. Simon SR, Kaushal R, Cleary PD, Jenter CA, Volk LA, Orav EJ, et al. Physicians and electronic health records: a statewide survey. Archives of Internal Medicine 2007;167(5):507-512.
4. Menachemi N, Perkins RM, van Durme DJ, Brooks RG. Examining the adoption of electronic health records and personal digital assistants by family physicians in Florida. Inform Prim Care 2006;14(1): 1-9.
5. Park RW, Shin SS, Choi YI, Ahn JO, Hwang SC. Computerized physician order entry and electronic medical record systems in Korean teaching and general hospitals: results of a 2004 survey. J Am Med Inform Assoc 2005;12(6):642-647.
6. Sittig F, Guappone K, Campbell E, Dykstra R, Ash J. A survey of USA acute care hospitals' computer-based provider order entry system infusion levels. Stud Health Technol Inform 2007;129(1):252.
7. DesRoches CM, Campbell EG, Rao SR, Donelan K, Ferris TG, Jha A, et al. Electronic health records in ambulatory care--a national survey of physicians. The New England Journal of Medicine 2008;359(1):50-60.
8. Dewitt JG, Hampton PM. Development of a data warehouse at an academic health system: knowing a place for the first time. Acad Med 2005;80(11):1019-1025.
9. Schubart JR, Einbinder JS. Evaluation of a data warehouse in an academic health sciences center. International Journal of Medical Informatics 2000; 60(3):319-333.
10. Silver M, Sakata T, Su HC, Herman C, Dolins SB, O'Shea MJ. Case study: how to apply data mining techniques in a healthcare data warehouse. J Healthc Inf Manag 2001;15(2):155-164.
11. Zhang Q, Matsumura Y, Teratani T, Yoshimoto S, Mineno T, Nakagawa K, et al. The application of an institutional clinical data warehouse to the assessment of adverse drug reactions (ADRs). Evaluation of amino-glycoside and cephalosporin associated nephrotoxicity. Methods Inf Med 2007;46(5):516-522.
12. Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. Artif Intell Med 2001;23(1):89-109.

13. Lavrac N. Selected techniques for data mining in medicine. Artif Intell Med 1999;16(1):3-23.
14. Kopelman LM. Minimal risk as an international ethical standard in research. The Journal of Medicine and Philosophy 2004;29(3):351-378.
15. Cios KJ. Medical data mining and knowledge discovery. IEEE Eng Med Biol Mag 2000;19(4):15-16.
16. Cios KJ, Teresinska A, Konieczna S, Potocka J, Sharma S. A knowledge discovery approach to diagnosing myocardial perfusion. IEEE Eng Med Biol Mag 2000; 19(4):17-25.
17. Yuan YC. Multiple imputation for missing data: concepts and new development. Paper presented at: Twenty-Fifth Annual SAS Users Group International Conference 2000.
18. Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods 2002;7(2):147-177.
19. Harel O, Zhou XH. Multiple imputation: review of theory, implementation and software. Stat Med 2007; 26(16):3057-3077.
20. Haykin S. Neural networks and learning machines. 3rd ed. New York: Prentice Hall;2008.
21. Bishop CM. Pattern recognition and machine learning. 2nd ed. New York:Springer;2005. pp. 291- 358.
22. Rokach L, Maimon O. Data mining with decision trees: theroy and applications. Danvers, MA: World Scientific Publishing Company;2008.
23. Heckerman DE. Learning Bayesian networks: The combination of knowledge and statistical data. Redmond, WA: Microsoft Research;1995. MSR-TR-94-09.
24. Heckerman DE. Bayesian networks for data mining. Data Mining and Knowledge Discovery 1997;1:79-119.
25. Heckerman DE, Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R. Bayesian networks for knowledge discovery. Advances in knowledge discovery and data mining. Menlo Park, CA: The MIT Press; 1996. pp. 273-305.
26. Lee SM, Abbott P. Bayesian networks for knowledge discovery in large datasets: basics for nurse researchers Journal of Biomedical Informatics 2003;36(4/5):389-399.
27. SPSS. Clementine 12.0 modeling nodes. Chicago: SPSS;2007.
28. SPSS. Clementine manual-Basic. Seoul:SPSS;2007.
29. Menard SW. Applied logistic regression analysis. 2nd ed. London: Sage Publications;2001.
30. Lee SM, Abbott P, Johantgen M. Logistic regression and bayesian networks to study outcomes using large data sets. Nursing Research 2005;54(2):133-138.
31. Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. Journal of Clinical Epidemiology 1996;49:1225-1232.
32. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci U S A 2001;98(26):15149-15154.
33. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 2000;16:906-914.
34. Lauritzen SL, Spiegelhalter DJ. Local computations with probabilities on graphical structures and their application to expert systems. Journal of the Royal Statistical Society Series B 1988;50(2):157-194.
35. Eisenstein EL, Alemi F. A comparison of three techniques for rapid model development: an application in patient risk-stratification. Proceedings/AMIA Annual Fall Symposium 1996:443-447.
36. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143(1):29-36.
37. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148(3): 839-843.
38. Rowland T, Ohno-Machado L, Ohrn A. Comparison of multiple prediction models for ambulation following spinal cord injury. Proceedings/AMIA Annual Symposium 1998:528-532.
39. Hosmer DW, Lemeshow S. Goodness of fit tests for the multiple logistic regression model. Communications in Statistics 1980;A 9(10):1043-1069.
40. Lemeshow S, Hosmer DW. A review of goodness of fit statistics for use in the development of logistic regression models. American Journal of Epidemiology 1982;115(1):92-106.
41. Blum RL. Displaying clinical data from a time-oriented database. Computers in Biology and Medicine 1981; 11(4):197-210.
42. Elomaa T HN. An experimental comparison of inducing decision trees and decision lists in noisy domains. Paper presented at: 4th European Working Session on Learning; Dec 4-6, 1989; Montpeiller.
43. Lesmo L SL, Torasso P. Learning of fuzzy production rules for medical diagnoses. In: Gupta MM SE, editor.

Approximate reasoning in decision analysis. Amsterdam: North-Holland;1982. pp.249-260.
44. Hojker S KI, Jauk A, Fidler V, Porenta M. Expert system's development in the management of thyroid diseases. Paper presented at: European Congress for Nuclear Medicine; Sep, 1988; Milano.
45. Horn W. AI in medicine on its way from knowledge-intensive to data-intensive systems. Artificial Intelligence in Medicine 2001;23(1):5-12.
46. Quinlan R CP, Horn KA, Lazarus L. Inductive knowledge acquisition: a case study. In: JR Q, editor. Applications of expert systems. Boston:Addison-Wesley; 1987. pp. 137-156.
47. Zupan B, Dzeroski S. Acquiring background knowledge for machine learning using function decomposition: a case study in rheumatology. Artif Intell Med 1998; 14(1-2):101-117.
48. Cohen ME, Hudson DL. Neural network models for biosignal analysis. Conf Proc IEEE Eng Med Biol Soc 2006;1:3537-3540.
49. Chun FK, Karakiewicz PI, Briganti A, Walz J, Kattan MW, Huland H, et al. A critical appraisal of logistic regression-based nomograms, artificial neural networks, classification and regression-tree models, look-up tables and risk-group stratification models for prostate cancer. BJU Int 2007;99(4):794-800.
50. Rodriguez Alonso A, Pertega Diaz S, Gonzalez Blanco A, Pita Fernandez S, Suarez Pascual G, Cuerpo Perez MA. The utility of artificial neural networks in the prediction of prostate cancer on transrectal biopsy. Actas Urol Esp 2006;30(1):18-24.
51. Stephan C, Cammann H, Jung K. Artificial neural networks: has the time come for their use in prostate cancer patients? Nat Clin Pract Urol 2005;2(6): 262-263.
52. Gamito EJ, Crawford ED. Artificial neural networks for predictive modeling in prostate cancer. Curr Oncol Rep 2004;6(3):216-221.
53. Porter CR, Crawford ED. Combining artificial neural networks and transrectal ultrasound in the diagnosis of prostate cancer. Oncology (Williston Park) 2003; 17(10):1395-1399; discussion 1399, 1403-1396.
54. Schwarzer G, Schumacher M. Artificial neural networks for diagnosis and prognosis in prostate cancer. Semin Urol Oncol 2002;20(2):89-95.
55. Errejon A, Crawford ED, Dayhoff J, O'Donnell C, Tewari A, Finkelstein J, et al. Use of artificial neural networks in prostate cancer. Mol Urol 2001;5(4): 153-158.
56. Murphy GP, Snow P, Simmons SJ, Tjoa BA, Rogers MK, Brandt J, et al. Use of artificial neural networks in evaluating prognostic factors determining the response to dendritic cells pulsed with PSMA peptides in prostate cancer patients. Prostate 2000;42(1):67-72.
57. Gamito EJ, Stone NN, Batuello JT, Crawford ED. Use of artificial neural networks in the clinical staging of prostate cancer: implications for prostate brachytherapy. Tech Urol 2000;6(2):60-63.
58. Snow PB, Smith DS, Catalona WJ. Artificial neural networks in the diagnosis and prognosis of prostate cancer: a pilot study. J Urol 1994;152(5 Pt 2):1923-1926.
59. Giles LC, Whitehead CH, Jeffers L, McErlean B, Thompson D, Crotty M. Falls in hospitalized patients: can nursing information systems data predict falls? Computers, Informatics, Nursing 2006;24(3):167-172.
60. Tiet Q, Ilgen MA, Byrnes HF, Moos RH. Suicide attempts among substance use disorder patients: an initial step toward a decision tree for suicide management. Alcoholism: Clinical and Experimental Research 2006;30(6):998-1005.
61. Modai I, Valevski A, Solomish A, Kurs R, Hines IL, Ritsner M, et al. Neural network detection of files of suicidal patients and suicidal profiles. Medical Informatics and the Internet in Medicine 1999;24(4):249-256.
62. Anthony D, Clark M, Dallender J. An optimization of the Waterlow score using regression and artificial neural networks. Clinical Rehabilitation 2000;14(1): 102-109.
63. Brossette SE, Sprague AP, Hardin JM, Waites KB, Jones WT, Moser SA. Association rules and data mining in hospital infection control and public health surveillance. Journal of the American Medical Informatics Association 1998;5(4):373-381.
64. Rapeli CB, Botega NJ. Clinical profiles of serious suicide attempters consecutively admitted to a university-based hospital: a cluster analysis study. Revista Brasileira de Psiquiatria 2005;27(4):285-289.