

신경망과 의사결정 나무를 이용한 충수돌기염 환자의 재원일수 예측모형 개발

정석훈¹, 한우석², 서용무³, 이현실^{4*}
¹(주) 위즈페니언, ²건양대학교 병원관리학과
³고려대학교 경영대학, ⁴고려대학교 보건과학대학

Length-of-Stay Prediction Model of Appendicitis using Artificial Neural Networks and Decision Tree

Suk-Hoon Chung¹, Woo-Sok Han², Yongmoo Suh³ and Hyun-Sill Rhee^{4*}

¹WIZPANION, ²Dept of Hospital Management, Konyang University

³Dept of Business School, Korea University

⁴Dept of Health Administration College of Health Sciences, Korea University

요 약 충수돌기염 환자의 LoS(Length of Stay)를 예측하는 것은 병상의 운영에 적지 않은 영향을 준다. 본 논문에서는 Neural Networks와 Decision Tree를 이용하여 LoS와 연관이 높은 입력변수들을 찾아 그 의미를 분석하며, 찾아낸 입력변수들을 이용하여 다양한 LoS 예측 모형을 개발하고 그 성능을 비교하였다. 모형의 예측 정확성을 높이기 위하여 Bagging과 Boosting 등의 Ensemble 기법도 적용하였다. 실험 결과, Decision Tree 모형이 Neural Networks 모형보다 좀 더 적은 수의 속성을 가지고도 거의 동일한 예측력을 보였으며, Ensemble 기법 중에서는 Bagging 기법이 Boosting 기법보다 좋은 결과를 보여주었다. 의사결정나무 기법은 Neural Networks 기법에 비해 설명력이 있으며, 충수돌기염의 LoS 예측에 매우 효과적이었고, 중요 입력 변수의 선정에도 좋은 결과를 보여줌에 따라 향후 적극적인 기법의 도입이 필요하다고 할 수 있다.

Abstract For the efficient management of hospital sickbeds, it is important to predict the length of stay (LoS) of appendicitis patients. This study analyzed the patient data to find factors that show high positive correlation with LoS, build LoS prediction models using neural network and decision tree models, and compare their performance. In order to increase the prediction accuracy, we applied the ensemble techniques such as bagging and boosting. Experimental results show that decision tree model which was built with less number of variables shows prediction accuracy almost equal to that of neural network model, and that bagging is better than boosting. In conclusion, since the decision tree model which provides better explanation than neural network model can well predict the LoS of appendicitis patients and can also be used to select the input variables, it is recommended that hospitals make use of the decision tree techniques more actively.

Key Words : Neural Network, Decision Tree, LOS, Appendicitis

1. 서론

병원 관리에 있어서 재원기간(LoS: Length of Stay)에 대한 연구는 한정된 병원자원의 효율적 활용을 위하여 의료 데이터 연구분야에서 매우 중요한 분야 중 하나로

자리매김 하고 있다[1]. 특히, 재원기간에 대한 정확한 예측은 효율적인 Medical Pathways를 개발하거나 Intensive Care Unit과 같은 병원의 자원을 효율적으로 운용하는데 많은 도움을 줄 수 있다[2]. 따라서 근래에는 환자의 재원기간이 병원 역량의 평가와 병원 진료 활동의 중요한 것

*교신저자 : 이현실(pridehyun@korea.ac.kr)

접수일 09년 05월 16일

수정일 09년 06월 15일

게재확정일 09년 06월 17일

대가 되고 있다[3,4]. 이와 더불어 한국에서의 포괄수가제(DRG: Diagnosis Related Group) 도입은 병원 자원의 효율적 배분 문제를 포함한 병원의 경영적 문제에 대해서도 많은 변화들을 가져 올 것으로 예측되고 있다. DRG 제도란 특정 질병군에 대하여 수행된 치료나 조치 및 LoS에 상관없이 일정하게 정해진 수가만을 청구 하도록 한 제도이다. 따라서, DRG의 도입으로 인해 과도하게 길어지는 LoS는 곧바로 병원의 비용으로 발생하게 되며 이러한 비용의 과도한 발생은 재정 상태를 악화시키는 결과를 초래할 수도 있게 된다.

지금까지 많은 연구자들이 LoS에 관련된 연구들을 수행하여 왔다. 특히, 각급 병원들에 정보시스템이 도입됨에 따라 의료 행위 등에 대한 대용량의 데이터가 확보되면서 LoS에 대한 좀 더 과학적인 연구들이 수행될 수 있게 되었다. 그러나 이러한 의료분야의 대용량 데이터 분석에서 기계학습과 같은 데이터마이닝 기술이 매우 효과적으로 이용될 수 있음에도 불구하고[5], 몇몇 연구를 제외하고는 LoS 예측에 대한 대부분의 연구는 간단한 회귀 모형이나 기술 통계 분석을 통해서 이루어진 것들이 주를 이루어 왔다[6-11].

신경망(NN: Neural Networks)를 LoS 예측에 사용한 연구들을 살펴보면, Dombi GW는 늑골 골절 환자들의 재원기간을 예측하기 위하여 feed-forward 방식을 이용한 NN 모형을 개발했으며[12], Pofahl WE는 췌장염 환자들의 LoS를 6일 이상과 그 이하로 나누어 예측함으로써 췌장염의 심각성 정도를 고려하는데 사용하는 연구를 수행하였다[13]. Lowell WE 등은 정신병 환자들의 LoS를 여러 개의 NN 모형을 이용하여 예측하는 모형을 개발하였다[14]. 그러나 상대적으로 Decision Tree를 적용한 연구는 아직 많이 이루어지지 않고 있다.

LoS를 직접 예측하는 연구와 더불어 LoS에 중요한 영향을 미치는 요인이 무엇인지 밝히고 분석하는 연구들도 수행되었다. Sheng 등은 ‘병원 내 감염’의 문제가 의료기관의 형태에 따라 치료비, LoS, 그리고 치료 결과에 대해 어떤 영향을 주는지 분석하였으며[15], Demir 등은 통계적 기법을 사용하여 군 훈련소의 병원 데이터에서 충수절제 환자의 재원기간에 영향을 미치는 요인들을 찾아내고자 하였다[16]. 이와 같이 LoS와 관련된 연구들이 대부분 통계적인 기법을 이용하고 있고, 몇몇 소수의 연구에서 비교적 간단한 형태의 NN을 이용하고 있다. 그러나 데이터의 형태가 숫자형이나 문자형이 혼재되어 있거나 강력한 설명력이 요구되는 의학 분야 데이터 연구에서는 숫자형 데이터만 처리할 수 있고, 또 데이터 특성에 가정이 필요한 전통적인 통계 기법이나 설명력이 떨어지는 NN 기법보다는 데이터의 형태에 제한이 없고 이해하기

쉬운 결과를 보여주는 Decision Tree 기법이 더 효과적일 수 있다. 또한 기존의 임상학적 해석을 사용하지 않고 데이터마이닝 기법을 사용하여 데이터에서 직접 LoS와 연관이 있는 요인들을 찾아내는 연구에서는, 좀 더 다양한 기법들을 적용하여 그 결과를 상호 비교하면서 신중하게 접근할 필요가 있다. 그러나 지금까지 LoS에 관련된 연구들은 보통 하나의 기법만을 이용하여 예측 모형을 개발하거나, 중요 요인을 분석하고 있다. 따라서 본 연구에서는 충수돌기염 수술 환자의 LoS에 영향을 미치는 요인들의 중요도를 다양한 기법을 적용하여 비교 분석하고, 최적의 중요 요인 집합을 찾아 LoS를 예측하는 모형을 개발하고자 한다. 재원일수 예측 모형의 개발에는 Boosting이나 Bagging과 같은 Ensemble 기법들을 사용하여 좀 더 예측력이 좋은 모형을 개발하고자 하였다.

충수돌기염은 충수의 폐쇄로 인해 염증반응이 일어나는데, 이것이 충수내의 압력을 더욱 증가시켜 정맥압(venous pressure)의 증가 및 모세혈관차단(capillary block)으로 인해 혈류(blood flow)의 정체현상이 초래되면서 이곳에 2차적인 세균감염이 발생하게 되는 것을 말한다. 충수돌기염의 병리학적 분류를 보면 초기의 것은 림프액의 순환장애로 인한 충수 부종 및 세균의 침입과 점막 궤양을 보이는 급성 병소성 충수염(acute focal appendicitis)과, 염증 반응이 더욱 진행되어 충수 내에서 점액분비(blenorrhagia)가 발생하여 압력이 증가되고 혈전(thrombus) 형성으로 충수의 부종과 허혈(ischemia)을 일으키는 급성 화농성 충수염(acute suppurative appendicitis)이 있다. 이 단계에서 염증이 더욱 심해지면 동맥혈 공급에 장애가 오게 되며, 조직의 괴사(necrosis)를 초래하여 괴저성 충수염(gangrenous appendicitis)으로 발전한다. 만약 남아 있는 정상의 점막에서 점액이 계속 분비되는 경우에는 압력이 더욱 증가하게 되어 괴사부위가 파열되며, 천공성 충수염(perforative appendicitis)을 일으키게 된다. 또한 충수염이 진행하여 국한성의 복막염(peritonitis)을 일으킨 경우에 우하복부에 근성 방어(muscular defense)를 볼 수가 있으며, 농양(abscess)의 형성으로 인한 종기(lump)를 촉지하는 경우도 있다. 혈액상(blood picture)은 백혈구(leukocyte)의 증가가 중요한 소견인데, 이것이 두드러지지 않은 경우라도 호중구 증가(neutrophilia)나 백혈구 백분율(leukocyte analysis percent)의 좌방이동이 도움이 된다. 이러한 특징을 갖는 충수돌기염의 재원일수 단축노력은 의료제공 체계의 효율적 운영을 위해서도 매우 중요하다.

이에 본 연구에서는 충수돌기염과 LoS의 관련성을 NN과 Decision Tree를 이용하여 분석하고자 하였다.

2. 연구대상 및 방법

2.1 연구대상

본 연구에 사용된 데이터는 일개 대학병원에서 수집한 충수돌기염 수술 환자를 대상으로 하였다. 변수 중에서 ‘수술 월’ 그리고 ‘수술 계절’과 같이 상호 높은 상관성이 예상되는 변수는 좀 더 상세한 정보(예, 계절 단위 보다는 월 단위)를 담고 있는 변수를 택하고 나머지 하나는 제거하였다. Null value를 갖고 있는 변수는 명목 변수의 경우에는 ‘NULL’이라는 새로운 변수 값을 부여하였으며, 연속형 변수인 경우에는 평균값으로 대체하였다. 우리가 예측하려고 하는 목표변수의 경우 총 LoS가 3일 이상인 경우에는 장기 입원자(long), 그 미만인 경우에는 단기 입원자(short)로 하여 두 가지 종류의 변수 값을 갖게 하였다. 충수돌기염의 LoS가 얼마 이상이면 장기인지에 대한 정의는 사실 매우 다양한 상황에서 각각 다르게 내려져 있다. Mohammed(2003)의 연구에 의하면 천공성 충수염이 있는 환자의 평균 LoS는 6.2일이며 천공이 없는 환자의 경우에는 3일로 보았는데, 이는 대한민국에서의 충수돌기염 평균 재원일수(3일)와 일치하는 수치임을 알 수 있다[17]. 그러나 Towfigh 등(2006)은 복강경 충수 절제술을 받은 환자의 LoS는 1.2일로 개복 충수 절제술을 받은 환자보다 짧아지는 것을 관찰하였으며, 특히 LoS는 파열된 충수염인지 아닌지의 여부에 따라 크게 영향을 받는다고 학계에 보고하기도 하였다[18] 이렇게 각 상황에 따라서 LoS의 장기 또는 단기에 대한 정의가 달라지지만 본 연구에서는 한국에서의 평균 재원일수인 3일을 사용하였다. 따라서 LoS가 3일 이상인 경우에는 장기 입원 환자로 보았으며 그 미만일 경우에는 단기 환자로 보았다. 위와 같은 전처리 과정을 거쳐 최종적으로 총 29개의 candidate input variable과 1개의 목표 변수를 구성하였다. Candidate input variable에는 ‘성’, ‘나이’ 등과 같은 환자 개인의 인구통계학적 데이터와 ‘재원 월’, ‘재원 과’, ‘수술 월’, ‘수술 요일’ 그리고 ‘재원 경로’ 등 환자의 행동과 관련된 데이터 그리고 ‘BUN’, ‘Creatinine’, ‘Protein’ 그리고 ‘AST_ALT’ 등 내원 직후 검사를 통하여 얻은 검사수치들을 포함하고 있다.

Raw data에서 사망환자 1인을 제외한 총 466명의 환자들에 대한 정보가 실험에 사용되었으며, 이 중 남자 환자는 239명 여성 환자는 227명이다. 환자들의 연령은 최소 3세에서 최고 85세로 평균 32세이고, 목표 변수의 경우 장기 환자는 215명이며 단기 환자는 251명으로 구성되어 있다.

2.2 연구방법

본 연구는 classification을 목적으로 가장 많이 사용되고 있는 데이터마이닝 기법은 NN와 Decision Tree으로써, 본 연구에서도 이 두 기법을 사용하였다. 아래에서 간략하게 이 두 기법을 소개하고자 하였다.

2.2.1 Neural Networks

NN 기법은 가장 많이 사용되고 있는 기계 학습방법 중 하나이다. 인간의 뇌세포들이 신경망을 통하여 주변의 다른 뇌세포들과 정보를 주고 받으면서 계산을 하거나 의사결정을 내리는 과정을 모방하고 있는 기법이다. 이 기법을 사용하기 위해서는 먼저 뇌의 구조를 모방한 형태로 input layer, hidden layer 그리고 output layer로 구성된 NN의 구조를 정의하고, 일정한 learning process를 거쳐 예측이나 분류를 할 수 있는 모형을 생성하였다.

우선, NN의 구조는 input layer의 node 수, hidden layer의 수와 각 hidden layer의 node 수, 그리고 output layer의 node 수에 의하여 정의된다. 그 다음, neural networks에서의 learning은 인접한 다른 층의 node와 node를 이어주는 arc들의 weight를 적절한 값으로 조정하는 과정을 통하여 이루어진다. 이러한 NN의 learning 과정에서 중요한 점은 weight들을 조절할 때 어느 정도씩 어느 방향으로 조절할 것인가를 결정하는 것과, 또 언제 이 learning 과정을 멈추도록 할 것인가를 지정하는 일이다.

Commercial data-mining tool에서는 여러 가지 옵션을 제공함으로써 사용자가 원하는 NN의 구조를 결정할 수 있도록 하고, learning과 관련된 내부 변수들의 값을 변경해 봄으로써 학습과정을 조절할 수 있도록 하며 학습을 멈추기 위한 기준도 사용자가 지정하도록 하고 있다. NN 기법은 noisy data에 크게 영향 받지 않는 장점이 있으나, Decision Tree에 비하여 설명력이 부족하다는 단점을 갖고 있다. 그러나 모형의 결과에 대한 설명력이 강력히 요구되는 의료 분야의 데이터 분석에도 많이 이용되고 있으며 LoS 예측에도 종종 이용되고 있다[12-14].

2.2.2 Decision Tree

Classification model을 생성하는데 많이 사용 되어온 Decision Tree 기법은 이질적인 사례들로 구성된 큰 집합을 점차 덜 유사한 사례들의 작은 여러 집합으로 분할하는 과정을 반복적으로 수행함으로써 decision making에 사용될 나무구조의 모형을 형성한다. 따라서 이 과정에서 고려해야 할 것은 나무구조를 형성하는 각 node에서의 분할 기준으로 사용되는 변수와 그 값의 선택은 어떻게 할 것인가와, 분할 과정은 언제 멈추게 할 것인가 그리고

작성된 모델의 분류 정확도가 떨어지면 어떻게 해결할 것인가 등이다.

먼저, 분할 기준은 큰 집합이 여러 작은 집합으로 분할된 후 이들의 확률적 순수도의 합이 분할 전 집합의 순수도 보다 가장 크게 하는 변수와 그 값을 찾게 되는데 이때 Gini index 또는 Entropy index 등을 활용한다. 이러한 분할 작업을 반복적으로 수행하다가 분할된 집합에 속하는 사례들의 수가 사용자가 지정한 수치보다 작거나, 그 집합의 순수도가 사용자가 지정한 수치보다 크게 되면 분할 작업을 멈추게 된다. 이렇게 형성된 의사결정나무 모델을 테스트 데이터 세트를 이용하여 분류 오류가 높은 node들은 pruning 과정을 통해 제거하는 작업을 수행하게 된다.

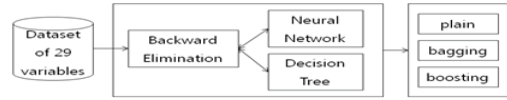
상용 툴마다 분할 기준을 정하는 데 사용하는 방법들을 다양하게 제공하고 있으며, 분할 과정을 멈추게 하기 위한 다양한 옵션을 제공하여 사용자가 지정하게 하고 있다. Decision Tree에서 생성된 규칙은 인간이 이해하기 매우 쉬운 형태로 만들어지기 때문에 그 설명력이 매우 뛰어나다는 장점이 있다[19] 그러나 아직 LoS 예측의 문제에 사용된 예가 극히 드물다.

2.2.3 모형 형성

Classification model을 생성할 때에는 candidate input variable들 중에서 target 변수에 영향을 미치는 주요 변수들을 선정하여 사용하게 된다. 특히 본 연구에서 사용한 데이터 세트는 환자들에 대한 기본 검사 결과자료가 입력변수들로 사용되었다. 따라서 본 실험에서의 변수 선정은 환자들에 대한 기본 검사의 필요성을 확인한다는 의미도 있을 수 있기 때문에 특별히 변수 선정에 주의를 기울이면서 변수 선정 기법에 따라 크게 두 가지로 실험을 진행하였다. 첫 번째 연구는 Backward Elimination 기법을 사용한 classification model 개발이며, 두 번째 연구는 Genetic Algorithm을 이용한 classification model 개발이다. 두 연구 모두 Classification을 위한 induction algorithm으로는 위에서 설명한 NN 기법과 decision Tree 기법을 적용하여 비교하였다. Kohavi의 분류방식에 의하면 첫 번째 실험은 wrapper 기법으로 볼 수 있으며, 두 번째 실험은 filtering 기법과 wrapping 기법을 함께 사용한 것으로 볼 수 있다[20].

Backward Elimination 기법은 최초로 모든 변수들을 사용하여 변수들의 중요도 순위를 산정한 후, 그 다음 단계에서는 가장 중요하지 않았던 변수를 하나씩 제거하면서 그 결과를 관찰하는 기법이다. 매 단계마다 가장 중요하지 않은 하나의 변수를 제거한 후 나머지 변수들끼리 새롭게 변수의 중요도를 다시 산정하기 때문에 보통

Feed-forward 기법보다 그 성능이 좋은 것으로 알려져 있다. 그림 1은 Backward Elimination 기법을 사용하여 입력 변수를 선정한 첫 번째 연구의 흐름도이다.



[그림 1] Experiment flow of research #1

첫 번째 연구에서는 최초 29개의 candidate input variable들에 대하여 NN 또는 Decision Tree를 이용하여 목표 변수의 예측에 영향을 미치는 정도의 순위를 매긴 후, 중요하지 않은 변수들은 하나씩 제거해 나가면서 가장 예측력이 높은 입력 변수의 집합을 최종 예측 모형의 입력 변수 세트로 사용하였다.

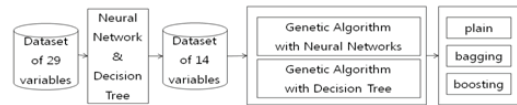
Genetic Algorithm을 사용한 두 번째 연구는 Genetic Algorithm을 이용하여 최적의 입력 속성의 집합을 찾은 다음 선정된 변수들을 induction algorithm에 적용시키는 방법이다. Genetic Algorithm을 사용할 경우 이론적으로 Sub-optimal set of variables를 비교적 빠른 시간에 찾을 수 있다는 장점이 있지만, 실제 적용함에 있어서는 maximum population, initial population 등 parameter 값을 어떻게 부여하는가에 따라서 그 결과에 많은 영향을 미칠 수 있으며 시간도 많이 소모될 수 있다. 따라서 두 번째 연구에서는 시간도 줄이고, 더 나은 결과를 유도하기 위하여 29개의 candidate input variable 중에서 덜 중요하다고 판단되는 변수들을 1차로 제거하였다. 이렇게 줄여진 변수들 중에서 다시 Genetic Algorithm을 이용하여 중요한 변수를 찾았다.

1차로 제거할 변수는 다음과 같이 결정하였다. 먼저 Tree 기법을 이용하여 29개의 후보 입력 변수들에 중요도 순위를 부여해 본 결과 상위 16개까지만 중요도가 산정되고 나머지는 모두 중요도가 0값을 보였다. 이 결과는 Tree기법으로 중요도를 산정했을 때 17위 밖에 위치한 변수들은 목표 변수의 예측에 아주 미미한 영향만을 준다는 것을 나타낸다. 따라서 이들 29개 입력 후보 변수들로부터 다시 NN를 이용하여 상대적 중요도 상위 16위 까지 만의 변수들을 선정한 후** SPSS 사의 Clementine 을 사용하였으며, 두 기법(Decision Tree and Neural Networks)에서 공통으로 선정한 변수들만을 Genetic Algorithm에 적용시켰다. 이러한 방법으로 찾아진 변수는 16위까지의 변수 중에서 age와 gender가 제거되고 총 14개가 선정됐으며 그 내용은 표 1과 같다. 표에서 변수들의 순서는 중요도 순위와는 상관이 없었다. 그림 2는 두 번째 연구의 흐름도이다. 확보된 환자들의 수가 다소 적기 때문에 두 실험에서 모든 모형들은 5-fold cross

validation 기법을 적용하여 그 결과에 일관성을 도모하였으며, 모든 모형은 예측력을 높이기 위하여 Ensemble 기법 중 하나인 Boosting과 Bagging 기법을 각각 적용하여 Ensemble 기법을 적용하지 않은 모형과 그 성능을 비교하였다. K-cross validation 기법은 데이터 세트를 k개로 나누는 후, 이 중 하나의 데이터 세트는 test set으로 사용하고 나머지 k-1개의 data set들은 training set으로 사용하는 방법을 말하는데, 이 때 모든 k 개의 data set은 돌아가면서 한 번씩 test set이 되어 k개의 결과들을 나중에 평균하는 방법이다. 이 방법을 사용하면 적은 데이터 세트에서 나온 결과에 대하여도 어느 정도 일관성을 부여할 수 있다. Ensemble 기법은 모형을 개발 할 때, 하나의 데이터 세트를 이용하여 한 개의 모형을 개발하여 사용하는 것이 아니라 데이터 세트에서 여러 개의 하위 데이터 세트를 만들어 내고 각 하위 데이터 세트마다 모형을 개발한 후, 그 여러 개의 결과들을 통합하여 최종 모형의 결과로 사용하는 방법이다. Bagging은 그 다음 하위 데이터 세트를 만들 때 무작위로 구성 하는 반면에, Boosting은 그 다음 단계의 하위 데이터 세트를 만들 때 전 단계에서 잘 예측하지 못한 instance들이 좀 더 많이 포함되게끔 조절해 주는 방법을 사용하고 있다. 본 연구에서는 여러 Boosting 방법 중 가장 널리 사용되고 있는 Ada-Boosting 기법을 사용하였다. Genetic Algorithm의 parameter는 Max Generation을 1,000, 그리고 초기 population의 수를 200으로 설정 해주었다** WEKA v3.4.12를 사용*.

【표 1】 Genetic Algorithm에 적용된 14개 변수

연번	변수명	의 미
1	진단코드	ICD-10를 토대로 환자에게 내려진 주진단명
2	입원과	환자가 입원한 당시 배정된 진료과
3	수술월	환자가 수술을 받은 월
4	입원월	환자가 입원한 월
5	Glucose	포도당 수치
6	입원경로	환자가 어떤 경로로 이송 및 입원 의뢰되었는지에 대한 정보
7	Blood	
8	입원일	
9	수술일	수술일자, 수술입실시간, 마취시간 및 수술이 끝난 시간 파악
10	HBS-Ag	B형간염표면항원
11	Protein	단백질
12	Wbc2	White blood cell(만성백혈병 혹은 재생불량성 악성빈혈 체크)
13	Ab_Screening Test	B형 간염보균자 확인 유무 상태
14	Anti HBs	B형 간염보균자 확인 유무 상태
15	Age	(탈락)
16	Gender	



【그림 2】 Experiment flow of research #2

3. 연구결과 및 토의

3.1 연구결과-1

Decision Tree를 사용하고 Backward Elimination을 적용한 모형의 경우 입력변수가 24개일 때와 9개일 때 예측률이 68.67%로 공동으로 가장 좋은 결과를 보여주었다. NN의 경우에는 입력변수의 수가 20개일 때 68.67%로 Decision Tree의 경우와 동일한 최고 예측력을 보여주고 있다. 따라서 최고 예측률로 보았을 때, 좀 더 적은 수의 입력 변수를 이용하여 동일한 예측력을 보여준 Decision Tree가 NN보다 좀 더 효과적이라고 할 수 있다. [그림3]은 9개의 최적 input variable로 만든 Decision tree의 모습이다. 그림에서 볼 수 있는 것과 같이 Decision Tree 기법은 결정 과정을 자세히 확인할 수 있는 tree를 만들어 주기 때문에 모형의 결과에 대한 이해도 및 설명력이 높았다. 따라서 Decision Tree가 NN보다 충수돌기염 환자의 LoS 예측에 적용하기에는 좀 더 적합할 수 있음을 보여준다고 할 수 있다. 표 2와 표 3은 Backward Elimination 적용 실험에서 변수를 하나씩 제거해 나가면서 얻어진 classification model의 예측률 변화를 보여주고 있다.

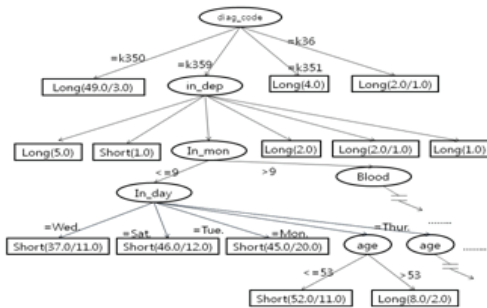
표에서 확인할 수 있는 것과 같이 ‘입원 월(in_mon)’이 재원일수 예측에 비교적 중요한 변수로 선정되었는데, Robin(2003)에 의하면 급성 충수돌기염의 연령과 성별-조정 발생률은 인구 10만 명 당 75명이었으며, 연령-정정을 성별 비는 1:1.4로 급성 충수돌기염은 타 계절에 비해 여름에 더 높게 나타나는 것을 관찰하였다[17].

3.2 연구결과-2

Genetic Algorithm에 의해 입력 변수를 먼저 선정한 후 Decision Tree와 NN를 적용시킨 경우에는 Backward Elimination을 적용시킨 모형보다 예측력이 다소 떨어지는 것을 관찰할 수 있었다. [표4]에서 그 결과를 확인할 수 있듯이, Decision Tree의 경우 총 10개의 입력변수가 최종 선정되었는데, 예측률은 plain model이 67.60%, Boosting 적용 모형은 63.09% 그리고 Bagging 적용 모형은 65.23%의 예측률을 보였다. Neural Networks 기법의 경우에는 이 보다 예측력이 더 떨어졌는데, 총 13개의 변

수가 선정되었으며 plain model은 60.51%, Boosting 적용 모형은 59.44% 그리고 Bagging 적용 모형은 63.30%의 예측률을 보였다. 이렇게 Genetic Algorithm을 사용한 모형의 예측률이 비교적 낮게 나온 이유는 충분한 량으로 parameter 값들을 설정해 주지 못해서 일 것으로 추정된다. 그러나 Max generation의 수나 초기 population의 수를 무작정 늘려 주는 것은 시간이나 비용상 현실적으로 불가능한 것이기 때문에 적절한 parameter 값들을 찾아내는 것에 대한 충분한 연구가 뒤따라야 할 것이다.

WBC의 경우에도 비교적 중요한 입력변수로 선택되었다. 따라서 전체적인 실험의 결과로 보았을 때, NN 보다는 Decision Tree의 성능이 다소 좋은 것을 관찰할 수 있었으며, Ensemble 기법 중에서는 Boosting 기법 보다는 Bagging 기법이 좀 더 좋은 성능을 보이는 것을 확인할 수 있었다.



[그림 3] 최적 의사결정나무 (일부)

[표 2] Results of Decision Tree Backward Elimination Model(%)

입력변수 개수	제거변수	plain	boosting	bagging
29	없음	64.16	64.16	65.24
28	AST_ALT	63.52	64.38	64.16
27	Protein-Total	64.60	61.16	63.73
26	Creatine	66.31	64.38	65.02
25	ALT-GPT	65.45	62.88	64.16
24	BUN	66.09	65.45	66.52
23	Wbc	65.67	65.02	68.67
22	Glucose1	65.45	65.45	66.31
21	in_num(입원횟수)	65.02	65.45	68.24
20	Hb	65.02	64.38	67.81
19	Oper_code(수술코드)	65.02	66.09	67.81
18	PLT	64.59	63.52	68.24
17	CPK	63.95	61.16	65.24
16	HCT	64.81	61.37	66.52
15	sex	65.67	59.66	67.60
14	Anti-Hbs	65.67	65.88	66.52
13	Ab_screening_Test	65.45	65.67	67.38

12	Wbc2	66.09	65.67	66.52
11	Protein	67.17	66.74	68.03
10	HBS_Ag	67.60	67.60	67.17
9	Oper_day(수술일)	66.52	66.52	68.67
8	in_day(입원일)	67.60	67.60	66.31
7	Blood	66.95	66.95	66.31
6	in_path(입원경로)	65.02	65.02	66.31
5	glucose	65.45	65.45	63.30
4	in_mon(입원일)	65.45	65.45	63.30
3	oper_mon(수술월)	65.24	65.24	63.52
2	Age(나이)	65.02	65.02	65.02
1	in_dep(입원과)	63.73	63.73	63.73
0	diag_code(진단코드)	X	X	X

[표 3] Results of Neural Network Backward Elimination Model(%)

입력변수 개수	제거변수	plain	boosting	bagging
29	없음	61.16	61.16	64.8
28	BUN	63.09	63.3	65.24
27	Creatine	66.31	66.31	65.02
26	in_num(입원횟수)	63.52	63.09	64.81
25	CPK	63.73	63.3	63.95
24	AST_ALT	63.73	63.73	65.24
23	sex	66.31	66.31	65.88
22	glucose1	65.88	65.88	66.52
21	Ab_Screening Test	67.6	67.6	66.74
20	AST_ALT	64.16	64.16	68.67
19	in_path(입원경로)	64.38	64.38	66.52
18	wbc	64.38	64.38	64.81
17	Protein_Total	61.16	61.16	65.88
16	HCT	64.16	64.16	65.02
15	oper_code(수술코드)	62.23	62.23	64.38
14	PLT	64.16	64.16	65.02
13	HBS_Ag	63.95	63.95	63.73
12	Protein	63.95	64.59	67.38
11	in_mom(입원월)	67.16	68.45	68.24
10	oper_mom(수술월)	63.95	64.16	66.31
9	Anti_Hbs	62.23	62.44	64.59
8	Age(나이)	63.95	61.37	65.24
7	Hb	65.45	65.67	66.09
6	wbc2	68.02	67.38	66.31
5	oper_day(수술일)	62.88	62.45	62.44
4	in_day(입원일)	65.67	65.23	66.09
3	Blood	65.02	65.24	65.45
2	glucose	65.02	65.02	65.02
1	in_dep(입원과)	63.73	63.73	63.73
0	diag_code(진단코드)	X	X	X

[표 4] Results of Research #2

입력변수	Decision Tree	Neural Networks
1	입원 월	입원 월
2	입원 일	입원 일
3	입원 과	입원 과
4	진단코드	진단코드
5	수술 일	수술 월
6	HBS_Ag	수술 일
7	Anti_HBs	입원 경로
8	Ab_Screening Tests	HBS_Ag
9	Blood	Anti_HBs
10	Wbc2	Ab_Screening Tests
11		Protein
12		Blood
13		Wbc2
Plain Model	67.60%	60.51%
Boosting Model	63.09%	59.44%
Bagging Model	65.23%	63.30%

4. 결론

정보화 사회를 앞두고 국민의 생활과 직결되는 보건복지 분야에도 정보화의 필요성이 그 어느 때보다 강하게 요구되고 있으며, 이러한 보건정보화를 통한 충수돌기염 환자의 신속한 치료방침의 결정과 해결 능력은 1차 진료를 담당하는 전문분야로서 해당 병원의 포괄적인 종합 관리능력의 지표로 활용될 수 있을 것이다. 따라서 실험의 결과에 의하면 의사결정나무 기법이 NN 보다 다소 효율이 좋거나 미미하지만 좋은 성능을 내 주는 것을 확인할 수 있었다. 또한 기대와는 달리 Boosting 기법은 본 데이터에서는 좋은 성능을 보여주진 못하였다. 오히려 plain model 이나 Bagging 모형이 더 좋은 성능을 보여주기도 하였다. 모형의 성능이 전체적으로 다소 떨어지는데 이것은 데이터의 양이 매우 적었기 때문인 것으로 추측해 볼 수 있다. 모형 성능을 발휘하는데 충분한 양의 학습 데이터를 사용할 수 없었기 때문일 것이다. 또한 데이터의 수집기간도 6개월 이내로 매우 짧은 기간 수집된 데이터이다. 좀 더 정확하고 의미 있는 모형을 개발하기 위해서는 좀 더 풍부한 데이터 세트를 수집해야 할 것이다.

본 연구는 충수돌기염의 진료비 예측 모형을 데이터마이닝 기법을 이용하여 개발하고 그 성능을 비교하여 보았다. 의료비 예측 분야에 데이터마이닝 기법을 적용시키고 의료비에 중요한 영향을 미치는 입력 변수들을 분석해 보았다는 것에서 본 연구의 의의를 찾을 수 있다. 이러한 재원일수 예측 등을 통한 재원일수의 단축노력은

의료제공 체계의 효율적 운영을 위해서도 매우 중요하다. 대부분의 종합전문요양기관들이 겪고 있는 입원환자 대기현상은 장기적으로 의료전달체계의 확립, 병상공급 증대 등의 정책을 통해 해결해야 하겠지만 단기적으로는 재원기간의 단축과 병상운영의 효율화를 통해 많은 부분을 해결할 수 있을 것이다. 또한 재원기간이 줄어드는 것은 환자의 입장에서는 의료비는 물론 불필요한 입원에 따른 제반 비용을 절감하는 방안이 될 수 있으며 국민의료비 차원에서는 낭비요인을 줄이는 방안이 될 수 있을 것이다.

참고문헌

- [1] Walczak S., Pofahl WE, Scorpio RJ, "A decision support tool for allocating hospital bed resources and determining required acuity of care", Decis Support Syst, 34. pp. 445-56, 2003.
- [2] Doering LV, Esmalian F., Imperial-Perez F, Monsein S, "Determinants of intensive care unit length of stay after coronary artery bypass graft surgery", Heart Lung, 30, pp. 9-17, 2001.
- [3] Lee AH, Ng SK, Yau KKW, "Determinants of maternity length of stay: a gamma mixture risk-adjusted model", Health Care Manage Sci. 4, pp. 249-55, 2001.
- [4] Wang K, Yau KKW, "Lee AH. Factors influencing hospitalization of infants for recurrent gastroenteritis in Western Australia", Meth Inform Med, 4, pp. 251-4, 2003.
- [5] Mitchell, Tom M, "Machine Learning and Data Mining", Communications of the ACM, 42, pp. 30-36, 1999.
- [6] Leung KM, Robert ME, Karen SR, Malik MH, Antonio PL, "Hospital- and Patient-Related Characteristics Determining Maternity Length of Stay: A Hierarchical Linear Model Approach", American Journal of Public Health, 88, pp. 377-381, 1999.
- [7] Zizza C, Amy HH, June S, Barry MP, "Length of Hospital Stays among Obese Individuals", American Journal of Public Health, 94, pp. 1587-1591, 2004.
- [8] Ottenbacher KJ, Pam MS, Sandra BI, Roger CF, Carl VG, "Length of Stay and Hospital Readmission for Persons with Disabilities", American Journal of Public Health, 90, pp. 1920-1923, 2000.
- [9] Epstein AM, Leighton R, Martha H. The Relation of Body Weight to Length of Stay and Charges for Hospital Services for Patients Undergoing Elective

Surgery: A Study of Two Procedures. American Journal of Public Health, 77, pp. 993-997, 1987

[10] Hyman I, Tonia F, Janice DM, Sarah O, Marsha MC, "The Association Between Length of Stay in Canada and Intimate Partner Violence Among Immigrant Women", American Journal of Public Health, 96, pp. 654-659, 2006.

[11] Daniel WW, Edward NB, J. Paul C, "The use of Demographic Characteristics in Predicting Length of Stay in a State Mental Hospital", American Journal of Public Health, 58, pp. 938-948, 1968.

[12] Dombi GW, Nandi P, Saxe JM, Ledgerwood AM, Lucas CE, "Prediction of rib fracture injury outcome by an artificial neural-network", J Trauma, 39, pp. 915-21, 1995

[13] Pofahl WE, Walczak SM, Rhone E, Izenberg SD, "Use of an artificial neural network to predict length of stay in acute pancreatitis", Am Surgeon, 64, pp. 868-72, 1998.

[14] Lowell WE, Davis GE, "Predicting length of stay for psychiatric diagnosis-related groups using neural networks", J Am Med Inform Assoc, 1, pp. 459-66, 1994.

[15] Sheng WH, Wang JT, Lu DCT, Chie WC, Chen YC, Chang SC, "Comparative impact of hospital-acquired infections on medical costs, length of hospital stay and outcome between community hospitals and medical centres", Journal of Hospital Infection, 59, pp. 205-214, 2005.

[16] Demir CC, Celik Y, Gider O, Yagci G, Sahin B, Turgut T, Col AA, Gen DS, "The Factors Length of Stay of the Patients Undergoing Appendectomy Surgery in a Military Teaching Hospital", Military Medicine, 172, pp. 634-639, 2007.

[17] Mohammed AO, Muhammad MM, Robin SM, "Epidemiologic features of acute appendicitis in Ontario, Canada", Canadian Journal of Surgery, 46, pp 263, 2003.

[18] Towfigh S, Chen C, Mason R, Katkhouda N, Chan L, Berne T, Endosc S, "Laparoscopic appendectomy significantly reduces length of stay for perforated appendicitis", 20, pp 495-499, 2006.

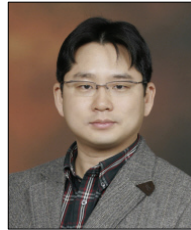
[19] Berry MJA, Linoff G, "Data Mining Techniques, John Wiley & Sons", Inc, 1997.

[20] Kohavi, R, "A study of cross-validation and bootstrap for accuracy estimation and model selection, In S. Wermter, E. Riloff, & G. Scheler(Eds.)", The fourteenth international joint conference on artificial

intelligence (IJCAI), Montreal, Quebec, Canada, San Francisco, CA: Morgan Kaufman Publishing, pp. 1137-1145, 1995.

정 석 훈(Suk-Hoon, Chung)

[정회원]



- 2001년 2월 : 아주대학교 경영정보학 석사
- 2001년 2월 ~ 2001년 9월 : (주) 대우증권, IT센터 근무
- 2003년 9월 ~ 2009년 2월: 고려대학교 경영학 박사
- 2009년 2월 ~ 현재 : (주)위즈페니언 근무

<관심분야>

의료 데이터마이닝, Fraud Detection Model

한 우 석(Woo-Sok Han)

[정회원]



- 1993년 8월 : 미네소타주립대 이학석사
- 2008년 3월 : 건양대학교 경영대학원 (경영학박사)
- 2000년 2월 ~ 2008년 2월 : 건양대학교 병원 행정부장
- 2008년 3월 ~ 현재 : 건양대학교 병원관리학과 교수

<관심분야>

병원 경영

서 용 무(Yong-Moo Suh)

[정회원]



- 1974년 3월 ~ 1978년 2월 : 학사 서울대학교 수학교육
- 1978년 3월 ~ 1980년 2월 : 석사 한국과학기술원 전산학
- 1983년 9월 ~ 1989년 5월 : 석사 Univ. of Texas at Austin 전산학
- 1989년 9월 ~ 1992년 12월 : 박사 Univ. of Texas at Austin 경영정보

- 1996년 9월 ~ 현재 : 고려대학교 경영대학 교수

<관심분야>

데이터마이닝, Semantic Web, Ontology

이 현 실(Hyun-Sill Rhee)

[정회원]



- 1990년 2월 : 서울대학교 보건대학원 (보건학석사)
- 2005년 2월 : 경산대학교 보건대학원 (보건학박사)
- 2004년 3월 ~ 현재 : 고려대학교 보건과학대학 보건행정학과 교수

<관심분야>
보건정보관리