의료 빅데이터를 활용한 질병 처방 예측 모델

고승완 ○, 강현태, 오영택, 박재호, 허의남
경희대학교 컴퓨터공학과
emakr518@khu.ac.kr, gusxo315@khu.ac.kr, vkdnej91@khu.ac.kr, qkrwoghwns@khu.ac.kr,
johnhuh@khu.ac.kr

Prediction Model of Disease Prescription Using Medical Bigdata

Seung-Wan Goh○, Hyun-Tae Kang, Young-Tak Oh, Jae-Ho Park, Eui-Nam Huh
Department of Computer Science and Engineering, Kyung Hee University

요 약

보건 의료 부문에서 빅데이터 활용의 기대 효과가 증가함에 따라 의료 데이터를 분석해 효과적인 치료 방법을 도출하는 연구에 대한 관심이 증대되고 있다. 의료 데이터 분석은 주로 어떤 질병에 대한 진단 및 처방 방법을 최적화하는 것에 목적을 두고 있으며, 이를 위해 코호트 연구 방법론을 기반으로 하는 다방면의 분석 및 예측 방법이 시도되고 있다. 따라서 본 논문에서는 의료 빅데이터를 분석해 질병 및 진단 관계를 분석하기 위해 코호트 기반의 의사 결정트리를 구성함으로써 최적화된 진료 추천 예측 모델을 도출한다. 본 논문에서 제시된 추천 모델은 모 병원에서 실측된 샘플 의료 데이터를 기반으로 구축되었으며, 사용되는 데이터의 수가 많을수록 정확도가 높아지는 것을 확인하였다.

1. 서 론

비즈니스 컨설팅 전문 업체인 McKinsey 에 따르면 보건 의료 부문이 빅데이터 활용의 용이성이 높고 경제 분야에서 큰 비중을 차지하고 있어 해당 분야에 대해 기대효과가 크다고 밝혔다[1]. 이에 따라 여러 분야에서 축적된 빅데이터에 관심이 증대되는 가운데, 국내외 의료 분야에서 빅데이터의 분석 기법 및 활용에 대한 연구가 활성화되고 있다. 의료 기관 및 관련 연구 기관에서는 환자 데이터를 지속적으로 축적하고 있으며, 모바일 기기와 IoT 의 발전과 함께 가정용 키트, 휴대폰과 자동차 등을 이용한 다양한 스마트 헬스케어 서비스에 대한 연구가 시도되고 있다. 다양한 헬스케어 모니터링 기기의 보급에 따라 수집할 수 있는 데이터의 양은 계속 증가하고 있으며, IDC 에서는 2020 년에는 약 25,000PB 까지 증가할 것으로 예상하고 있다[2]. 따라서 다양한 방면에서 수집되는 의료 빅데이터에 대한 분석 기술이 주목받고 있으며, 국외에서는 맞춤형 의료서비스를 위한 전자의료기록 분석, 질병 관리 및 예측을 위한 유전자 데이터 공유 분석과 그를 통한 질병치료체계 구축 등의 연구가 진행되고 있다[3]. 의료 데이터는 어떤 질병에 대한 진단과 처방이 가장 큰 비중을 차지함에 따라 코호트 연구 방법론에 의한 진료 기록 분석, 약물 분석, 처방, 역학조사, 예측 등의 분석이 시도되고 있다.

따라서 본 논문에서는 질병과 진단, 검사 결과와 처방의 관계를 분석해 의사 결정 트리를 구성하고, 환자의 효율적인 치료를 위하여 환자의 검진에 따른 최적화된 처방 추천을 하는 예측 모델에 대한 설계를 진행한다. 의사 결정 트리를 구성하기 위해 사용된 질병으로서 당뇨를 선택하였으며, 이로부터 추출된 의사 결정 트리를 검증하여 본 논문이 제안하는 질병에 따른 처방 예측 결과를 보일 것이다. 또한 검증에 사용할 데이터에 대한 무작위 추출을 여러 번 진행하고, 의사 결정 트리를 생성하는 표본 데이터의 크기를 변화시켜 데이터의 양 따른 예측 정확도를 도출할 것이다.

2. 관련 연구 – 결정 트리 (Decision tree)

결정 트리는 데이터의 패턴을 분석하여 예측 가능한 규칙을 도출하고 그 결과를 트리 구조로 도식화한 데이터 예측 모델이다. 트리의 각 내부 노드들은 하나의 입력 변수에 해당하며, 이를 기반으로 적절한 기준에 따라 데이터 집합을 부분 집합으로 분할함으로써 최종적인 목표 변수의 값을 예측할 수 있다.

2.1 정보 이론 – 엔트로피 및 지니 불순도

정보 이론은 데이터를 정량화하기 위한 응용 수학의 한 분야로서, 의사 결정 규칙을 명확하게 만드는 과정에서 불순도를 측정하는 데 흔히 사용되는 엔트로피와 지니 지수에서 차용되는 이론이다.

엔트로피(Entropy)는 분류된 부분집합 속 목푯값의 집합에 불확실성이 얼마나 포함되어 있는 지를 나타내는 데 사용되는 측정 방식이며, 데이터의 정보량을 의미한다. 한 분류만 포함한 부분집합은 목푯값이 100% 확실한 결과이므로 엔트로피가 0 이다. 반면 가능한 종류를 모두 포함하는 부분집합은 해당하는 모든 종류의 목푯값이 존재하므로 불확실성이 상승하여 엔트로피는 높은 값을 갖게 된다. 부분집합이 N 개의 분류로 이뤄져 있고, p_i 가 분류 i 에 속하는 표본의 비율이라면 엔트로피는 다음 [그림 1]의 공식으로 정의된다.

$$I_E(p) = \sum_{i=1}^{N} p_i \log\left(\frac{1}{p}\right) = -\sum_{i=1}^{N} p_i \log(p_i)$$

[그림 1] 엔트로피(Entropy) 공식

지니 지수(Gini Index)는 부분집합 내에서 무작위로 선택한 표본을 무작위로 선택해 분류했을 때 그 분류가 틀릴 확률이다. 엔트로피와 같이 단일 분류만 있는 부분집합의 경우 지니 불순도의 값은 0 이 되며, 분류의 수가 많아지고 데이터의 수가 대등해질수록 값이 증가한다. 지니 지수는 다음 [그림 2]의 공식으로 도출할 수 있다.

$$I_G(p) = 1 - \sum_{i=1}^{N} P_i^2$$

[그림 2] 지니 불순도 (Gini Index)

2.2 결정 트리에 사용되는 알고리즘

2.2.1 CART (Classification and Regression Tree)

CART 알고리즘은 의사 결정 트리 생성 방법론 중 가장 잘

알려진 방법론 중 하나로, 전체 데이터셋에서 두 개의 자식 노드를 생성하기 위해 모든 예측 변수를 사용해서 부분집합을 쪼갬으로써 의사 결정 트리를 생성한다. 불순도를 측정할 때, 이산적이지 않은 범주형일 경우 지니 지수를, 수치형은 분산(Variance)을 사용한다. CART 알고리즘은 후보 트리를 여러 개 생성한 후에 그 중에서 최적의 트리를 찾아내는 기법을 가장 큰 강점으로 갖는다.

2.2.2 C4.5 & C5.0

C4.5/C5.0 알고리즘은 범주형 뿐만 아니라 수치로 나타나는 목표 변수를 기준으로 분류할 수 있는 알고리즘이다. 수치형 속성을 사용할 때 트리의 모양이 복잡해지는 문제에 대해 보완책으로 자식 노드가 여러 개로 분기되는 다지 분리(multiple split)가 가능하다, 또한 속성 별 가중치 설정과 결측값 처리가 가능하며, 불순도 측도는 엔트로피를 사용한다.

2.2.3 CHAID (Chi-squared Automatic Interaction Detection)

CHAID 알고리즘은 범주형 목표 변수를 기준으로 분류하는 알고리즘이며, 수치형으로 나타나는 속성은 범주형으로 변환 후 입력해야 한다. 이 알고리즘은 최적의 예측 변수를 선택하는데 있어 불순도 측도는 카이제곱 통계량을 사용하며, 설정한 임계점에 도달하면 트리 모형 성장을 멈춘다.

2.3 의사 결정 트리 형성 과정

결정 트리는 분류와 예측을 수행하는 분석 방법이며, 나무 형태의 구조에 의한 추론 규칙에 의해 표현되어 분석 과정을 직관적으로 이해할 수 있도록 구성된다. 분석 결과는 조건에 따라 집단을 나누는 'if-then' 규칙으로 표현된다.

전체 데이터 셋은 트리 형성 목적에 따라 학습용 데이터 셋과 검증용 데이터 셋으로 분할한다. 학습용 데이터 셋을 첫 번째 속성에서 마지막 속성까지 각 노드의 분기 기준 값에 따라 'if-then' 규칙에 대입하여 가장 작은 엔트로피를 갖는 속성과 분기 기준을 찾는다. 분기 기준 값을 기준으로 대소 비교하여 자식 노드를 나누고, 이 과정을 반복하여 트리를 형성한다. 검증용 데이터 셋을 통해 의사 결정 트리의 타당성을 검토하고, 해당 결정 트리를 분류 및 예측 모형으로서 사용한다. [4][5]

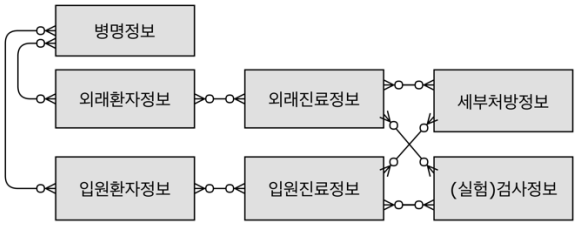
3. 최적화된 처방을 위한 의사 결정 트리 생성

본 논문에서는 실제 의료원으로부터 제공받은 환자의 고유 식별 정보가 제거된 내과계 데이터를 샘플 데이터로 사용하여, 다음과 같은 환자의 질병과 최근 검사에 따른 다음 처방을 예측하는 모델을 제안한다. 특정 질병에 대해 모든 진료 기록과 검사 결과, 처방 결과를 이용해 검사 결과에 대한 처방을 추출하고 이를 이용해 결정 트리 모델을 만든다. 학습을 위한 데이터와 학습용 데이터를 8:2 로 나누어 결정 트리의 정확도(신뢰도)를 검증한다

3.1 데이터 전처리

데이터 마이닝에서 분석에 필요한 데이터를 정제하는 과정은 필수이다. 다양한 릴레이션으로 구성된 데이터베이스에서 목적에 따라 필요한 테이블과 열을 골라내고 값을 정형화하는 작업을 수행한다.

어떤 질병을 갖고 있는 환자의 직전 검사를 바탕으로 다음 진료 시 어떤 처방을 받는 것이 좋을 지에 대해 예측하여 추천을 해주는 본 논문의 목표에 따라 환자 정보, 질병 정보, 입원 및 외래 진료 정보, 검사 결과 정보를 추린다. 다음 [그림 3]은 본 논문에서 다루게 될 데이터의 엔티티 관계를 간략히 나타낸다.



[그림 3] 연구에 사용할 데이터의 간략화된 엔티티 관계

3.2 결정 트리 학습용 테이블

전처리 과정에서 ~~추려낸~~생성한 데이터를 이용해 결정 트리 학습을 위한 테이블을 생성한다. 각 행과 열의 요소는 처방코드이며, 약물이나 검사 등을 모두 포함한다. 학습용 테이블은 처방 결과를 나타내는 pres 컬럼과 검사 결과를 나타내는 검사명 컬럼들로 이루어져 있다.

pres	25EW	21AD	24BF	24BG	24CRP	24D7
JHEPA	25					
YCOLON		8	157	152	3.64	32
BBPD		2.1	140		4.35	79
BAV10			63	310	0.37	
BMEGMB			60	255	0.53	
BCUT			74	183		
BXLACI			46	349		
BITO			46	111		
BROXA			58			
BOCN40			80	162	23.02	
BTERP		48.9	60		1.09	49
YSMET		47.8	48	651		
24DAT			88	318		
24DL			98			
24BG			129	207		
24DF			156	215		
BLEVO			164	268		
BXVZB		42.2				
81TB		9	21	74		
BXVZB		0.5				
24DL		67.7				
24DK		0.5				

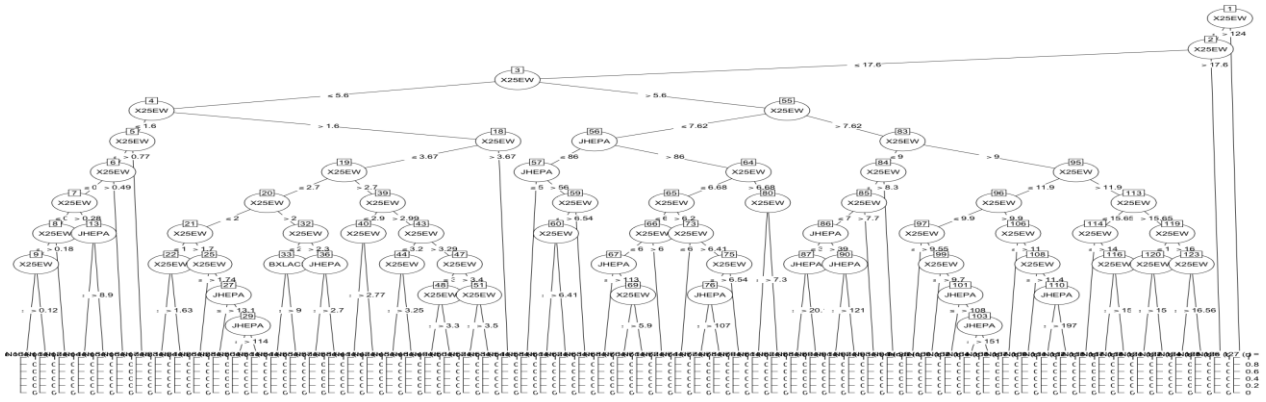
[그림 4] 결정 트리 학습용 테이블(생략, 2000 레코드)

3.3 결정 트리

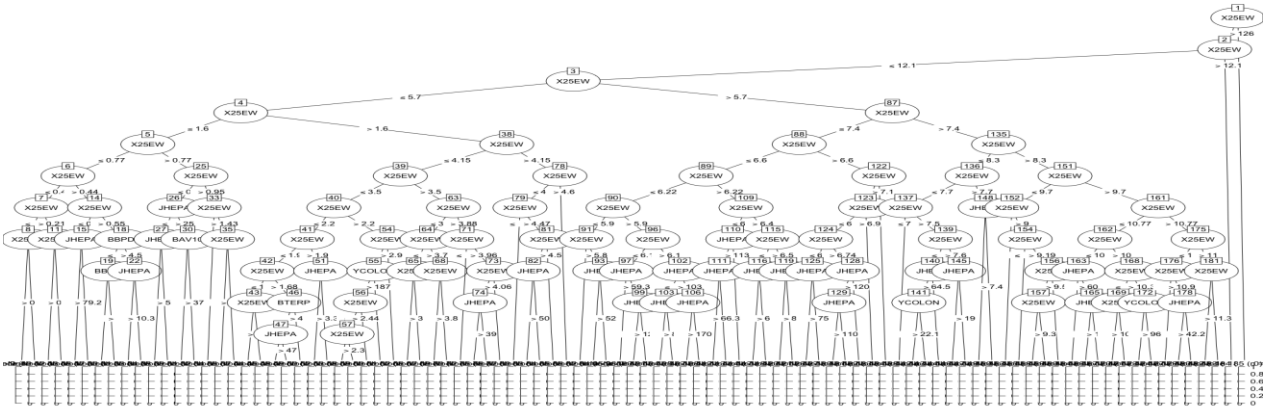
수치로 나타나는 검사 결과에 따라 처방이 도출되어야 하므로 수치형 변수를 입력할 수 있는 C5.0 알고리즘을 택하고, 생성한 학습용 테이블을 이용해 결정 트리를 생성한다. 다음 페이지의 [그림 5-1]과 [그림 5-2]과 같이 결정 트리를 생성하는 데 사용한 레코드의 개수(n)가 증가함에 따라 보다 세밀한 구조의 결정 트리가 생성되며, 이는 분류와 예측이 더 세분화됨을 나타낸다.

4. 결정 트리 검증

만들어진 의사 결정 트리의 정확도(신뢰도) 검증을 위하여 표본 중 80%의 학습용 데이터셋과 20%의 검증용 데이터셋을 무작위로 산출하여 만들어진 트리에 검증용 데이터셋을 모두 넣어보는 단위검증을 실시하였다. 또한 표본의 크기를 늘려가며 각 50 번씩 수행한 결과의에서의 최댓값, 최솟값, 평균값을 다음 [표 1]에 정리하였다.



[그림 5-1] 시각화된 Decision Tree(n=500)



[그림 5-2] 시각화된 Decision Tree(n=2000)

[표 1] 신뢰도 테스트 결과 요약

N	AVG(%)	MIN(%)	MAX(%)
250	63.435	52.940	71.023
500	64.921	58.221	66.925
750	67.339	64.239	69.155
1000	74.595	66.241	77.450
1250	72.443	65.704	78.821
1500	74.812	68.302	76.591
1750	76.944	71.430	77.240
2000	76.492	72.440	81.43

위 결과를 통해 알 수 있는 이 모델의 신뢰도는 최대 76.492%(2000 개의 표본)의 신뢰도를 얻을 수 있었고, 데이터 셋의 크기가 증가함에 따라 신뢰도가 증가함을 볼 수 있다.

5. 결론 및 향후 계획

본 논문에서는 1 만여명의 환자데이터 중 '당뇨' 키워드를 병명으로 가진 환자를 이용해 진료기록과 처방기록을 토대로 의사 결정 트리를 생성하였다. 신뢰도 검증을 위하여 표본 중 80%의 학습용 데이터셋과 20%의 검증용 데이터셋을 무작위로 산출하여 단위검증을 실시하였다. 의사 결정 트리의 특성상 학습 데이터수가 증가함에 따라 신뢰도가 증가하고 처리할 수 있는 처방이 고도화될 것으로 예상된다. 즉, 더욱 방대한 규모의 데이터를 기반으로 의사 결정 트리를 구축한다면 당뇨 질환뿐만 아니라 수많은 질환에 대해서 더욱 정밀하고 다양한 처방을 제시할 수 있을 것으로 예상된다.

반면, 수집되고 있는 의료 데이터는 그 형식이 표준화되어 있지 않아 빅데이터로서 사용하기에 제약이 존재한다는 한계점을 갖는다. 이를 극복하기 위해 OHDSI 는 국제적으로 방대한 의료 데이터의 포맷을 정형화할 수 있도록 표준화 작업을 수행하기 위한 작업의 일환인 CDM(Common Data Model)을 구축하고 있다[6]. 이러한 모델은 본 연구에서 진행한 것 이상으로 방대하며 각 데이터들이 다른 스토리지에 저장되어 있어 분산처리와 클라우드 개념이 필수로 요구된다. 따라서 향후 계획으로 본 연구의 결과 모델의 가지 정리와 함께 이러한 모델을 신속하게 처리할 수 있는 분산처리 기반 플랫폼에 대한 연구를 진행할 것이다.

* 본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 SW 중심대학지원사업의 연구결과로 수행되었음. (2017-0-00093)

6. 참고 문헌

[1] 장성재, "보건의료 빅데이터 관리시스템 최신 동향", 생물학연구정보센터(BRIC) 동향리포트, 2017-T07

[2] 김승환. "의료 IT 융합 기술 연구 동향." 전자공학회지, 43.2 (2016.2): 18-24.

[3] 송태민. "우리나라 보건복지 빅데이터 동향 및 활용 방안." 과학기술정책, .192 (2013.9): 56-73.

[4] 김진호, 박인식, 김봉옥, 양윤석, 원용관, 김정자. "결정트리 데이터마이닝을 이용한 족부 임상 진단", 대한전자공학회, 전자공학회논문지-CI 48(2), 2011.3, pp.28-37 (10 pages)

[5] 최종후, 서두성. "데이터마이닝 의사결정나무의 응용", 통계청「통계분석연구」제 4 권 제 1 호(99.봄), pp.61-83

[6] George Hripcsak 외 16 명, "Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers", MEDINFO 2015: eHealth-enabled Health I.N. Sarkar et al. (Eds.), pp.574-578