기계학습을 이용한 혈액검사 소견 생성

김유진[O1] 현종환[1] 고병수[1] 최호진[1]
[1]한국과학기술원 전산학부
117kyjin@kaist.ac.kr, hyeon0145@kaist.ac.kr, kobiso@kaist.ac.kr, hojinc@kaist.ac.kr

Generating Clinical Opinions for Blood Test Samples Using Machine Learning Techniques

YouJin Kim[O1] Jonghwan Hyeon[1] ByungSoo Ko[1] Ho-Jin Choi[1]
[1]School of Computing, KAIST

요약

사람들의 건강에 대한 관심이 증가함에 따라, 임상 의사결정지원 시스템(clinical decision support system)에 대한 필요성이 증가해 왔다. 기존의 의사결정지원 시스템은 규칙 베이스와 추론 엔진을 바탕으로 전문가의 의사결정을 보조했다. 하지만 이 시스템을 위한 규칙 베이스를 구축하는 과정은 전문가에게 부담이 될 수 있다. 따라서 본 논문에서는 이미 전문가가 생성한 데이터에 기계학습 기법을 적용해 의사결정지원 시스템을 구축하려 한다. 본 논문에서는 씨젠의료재단에서 제공한 익명화된 환자의 혈액검사 데이터에 기계학습 기법을 적용해 전문가 소견을 생성하는 기계학습 모델을 구축한다. 의사결정 트리(decision tree), 랜덤 포레스트(random forest), 심층 신경망(deep neural network)을 사용해 성능비교를 하였으며, 그 결과 심층 신경망(deep neural network)의 성능이 가장 높게 나타났다.

1. 서론

한국인의 평균수명은 2013년 기준 81.94세, 2014년 기준 82.40세로 지속적으로 늘어나고 있으며[1], 그에 따라 건강한 삶에 대한 사람들의 관심이 높아지고 있다. 혈액검사는 혈당 등 건강상태에 대한 정보를 제공하며, 질병진단을 위해 사용된다. 검사결과의 판독에는 의료 전문인이 필수적이며, 사람들의 건강에 대한 관심이 증가함에 따라 의료 전문인에 대한 수요가 점점 증가하고 있다.

그러나 증가하는 혈액검사 판독 수요에 비해, 제한된 의료 전문인의 공급으로 판독 작업에 병목 현상이 발생하고 있는 상황이다. 따라서 이러한 부담을 완화하기 위해 임상 의사결정지원 시스템(clinical decision support system)에 대한 필요성이 증가해 왔다.

의사결정지원 시스템은 컴퓨터 시스템에 특정분야의 전문적인 지식을 기억시켜 의사결정을 내리는데 이용하는 시스템[2]으로, 일반인들이 전문가의 부재에도 전문지식에 쉽게 접근할 수 있게 한다.

이러한 의사결정 지원 시스템은 통상적으로 규칙 베이스에 기반해 생성된다. 본 논문에서는 이러한 규칙 베이스를 구축하지 않고, 데이터로부터 패턴을 파악하는 기계학습 기반으로 의사결정 지원 시스템을 구축하려 한다. 왜냐하면, 규칙 베이스를 전문가가 직접 구축하는 과정은 번거롭고, 이미 전문가가 생성한 데이터가 매우 많이 존재하기 때문이다.

본 논문의 연구 목표는 특정 환자의 혈액검사 데이터로부터 전문가 소견을 생성하는 의사결정지원 시스템을 기계학습 기술을 이용하여 구축하는 것이다. 이를 위해, 씨젠의료재단에서 제공한 익명화된 혈액검사 데이터에 지도학습(superviesed learning) 기반의 의사결정 트리(decision tree), 랜덤 포레스트(random forest), 심층 신경망(deep neural network) 기법을 적용해 혈액검사 소견 생성 모델을 구축하였고, 다중-레이블 분류 문제를 위해 정의된 평가척도를 통해 각각의 성능을 비교하였다.

본 논문은 다음과 같이 구성되어있다. 2장에서는 의료 도메인의 의사결정지원 시스템에 대한 연구 사례를 소개하고, 3장에서는 혈액검사 데이터로부터 소견문장을 생성하는 기계학습 모델 생성 방법에 대해 설명한다. 4장에서는 사용한 데이터와 생성한 기계학습 모델의 실험 결과를 보이고, 5장에서는 본 논문의 결론을 논한다.

2. 관련 연구

의사결정지원 시스템은 컴퓨터 시스템에 특정분야의 전문적인 지식을 기억시켜 의사결정을 내리는데 도움을 주는 시스템[2]이다. 이러한 시스템은 지식 베이스와 추론 엔진으로 구성되어 전문가의 다양한 의사결정 과정을 보조한다.

의료 도메인에서 다양한 의사결정지원 시스템에 대한 연구 사례들이 있다. Polat는 [4]에서 당뇨병 진단 시스템을 제안하였으며, Liu는 [5]에서 알츠하이머 질병 여부와 진행 단계를 판단할 수 있는 심층 신경망 구조를 제안하였다.

Elshami은 나이브 베이즈(naïve Bayes), 의사결정 트리, 신경망(neural network)을 통해 지중해빈혈(thalassemia)을 진단하였으며[6], Symen은 혈액검사 데이터를 전방향 역전파 신경망(feedforward backpropagation neural network) 모델에 학습시켜, 철-결핍 빈혈을 진단하였다[7].

3. 방법

본 논문의 연구 목표는 특정 환자의 혈액검사 데이터로부터 전문가 소견을 생성하는 의사결정지원 시스템을 기계학습 기술을 이용하여 구축하는 것이다.

3.1 혈액검사 데이터

본 논문에서는 씨젠의료재단으로부터 제공받은 혈액검사 데이터를 사용하였다. 해당 데이터는 685개의 검사, 232개의 소견문장으로 구성되어있다. 데이터는 환자 별로 정리되어 있으며, 각 환자가 실시한 검사의 종류와 개수는 모두 다르다. 또한 환자는 검사결과에 따라 다수의 소견문장을 가질 수 있다.

그림 1은 혈액검사 데이터를 전문가에게 보여주는 기존 씨젠의료재단의 의사결정지원 시스템의 스크린샷 이다. 왼쪽엔 검사 데이터, 오른쪽엔 규칙 기반으로 생성된 소견문장을 보여주고 있다.

그림 2는 혈액검사 데이터에 대한 간단한 예시이다. 한 환자의 데이터는 해당 환자의 나이, 성별 그리고 검사결과로 구성되어 있다. 검사결과는 검사 코드와 검사 값으로 이루어져 있고, 검사 값은 분류형(nominal type) 또는 실수형(numeric type) 값을 가진다.

그림 1 씨젠의료재단의 기존 의사결정지원 시스템

환자정보		소견문장
나이	40	• B형간염항원과 항체가 음성입니다. 비면역상태이거나 항체의 역가가 감소한 상태입니다 • 종양표지자검사는 정상입니다. 종양표지자는 종양 등에서 생성되어 혈액이나 체액으로 분비되는 여러 종류의 물질입니다
성별	남	
검사결과		
21101	Negative	
21102	Negative	
21429	2.05	

그림 2 혈액검사 데이터 예시

3,2 기계학습 모델 설정

각각의 환자마다 혈액검사 결과에 따른 다수의 소견문장을 가지므로, 한 입력사례에 대해 하나 이상의 레이블을 갖는 다중-레이블 분류(multi-label classification) 문제이다[3]. 따라서 이를 다룰 수 있는 적절한 기계학습 기법을 선택하는 것이 중요하다. 본 논문에서는 의사결정 트리, 랜덤 포레스트, 그리고 심층 신경망 기계학습 기법을 사용한다.

의사결정 트리의 경우, 지니 불순도(Gini impurity)를 기준으로 사용해 모델을 구축하였고, 랜덤 포레스트의 경우, 총 200개의 의사결정 트리로 구성이 되도록 하였다.

표 1 심층 신경망 모델의 구조

계층	뉴런 수	활성화 함수
입력	1,041	ReLU
1	1,024	ReLU
2	512	ReLU
3	256	Sigmoid
출력	232	-

심층 신경망의 경우, 표 1과 같은 구조를 갖도록 설정한 뒤, 100회 학습을 진행했다. 마지막 계층의 경우 활성화 함수로 시그모이드(sigmoid)를 사용해, 각 소견문장이 발현될 확률을 학습할 수 있도록 했다.

모든 기계학습 모델의 초모수(hyper-parameter)는 5-fold 교차 검증을 통해 선택되었다.

4. 실험 및 결과

4.1 학습 데이터

본 논문에서는 씨젠의료재단으로부터 제공받은 총 14,479명의 익명화된 혈액검사 데이터를 사용해 의사결정 트리, 랜덤 포레스트, 심층 신경망 기계학습 모델을 구축했다. 해당 데이터의 예시는 그림 2에 나타나 있다.

4.2 데이터 전처리

혈액검사 데이터로부터 기계학습 모델을 생성하기 전, 데이터 전처리를 수행하였다. 왜냐하면, 제공받은 데이터가 기계학습에 적합하지 않은 불규칙한 형태를 가지고 있었기 때문이다. 예를 들면, Non-Reactive 1.0과 같이 분류형과 실수형이 혼재된 검사 결과 값이 존재하였다. 따라서 이러한 불규칙한 값들을 규칙적인 형태로 표현할 필요가 있다. 또한, 분류형 데이터를 One-hot 벡터로 표현해 기계학습에 적용이 가능하도록 하였다. 이러한 전처리 과정을 통해, 685개의 혈액검사와 환자의 나이 및 성별을 1,041 크기의 벡터로 표현할 수 있었다.

4.3 평가 척도

본 논문에서 다루는 혈액검사 데이터는 한 데이터마다 다수의 소견문장을 갖는 다중-레이블 분류 문제이다. 따라서 기존의 정확도(accuracy), 정밀도(precision), 재현도(recall), F_1-measure의 평가 척도 정의를 그대로 사용하는 것은 적합하지 않다. 왜냐하면, 다중-레이블 분류 문제에선 부분 일치(partially correct)라는 새로운 관점이 존재하기 때문이다. [8]

그러므로 본 논문에서는 위의 평가 척도에서 부분 일치를 고려하도록 Godbole가 제안한 정의를 사용한다 [9].

n개의 데이터 $(\mathbf{x}_i, \mathbf{Y}_i), 1 \le i \le n$로 이루어진 데이터 집합을 T, 다중-레이블 분류기를 h라 하고, 분류기 h의 분류 결과 $h(\mathbf{x}_i)$를 Z_i라고 하자. 이 때, 정확도, 정밀도, 재현도, F_1-measure는 다음과 같이 정의된다.

정확도(A): 실제(Y_i) 및 예측한(Z_i) 레이블 중 예측에 성공한 레이블 비율의 평균

$$Accuracy, A = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|}$$

정밀도(P): 예측한(Z_i) 레이블 중 예측에 성공한 레이블 비율의 평균

$$Precision, P = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i \cap Z_i|}{|Z_i|}$$

재현도(R): 실제(Y_i) 레이블 중 예측에 성공한 레이블 비율의 평균

$$Recall, R = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i \cap Z_i|}{|Y_i|}$$

F_1-measure(F): 정밀도와 재현도의 조화 평균

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2|Y_i \cap Z_i|}{|Y_i| + |Z_i|}$$

위와 같이 정의된 평가 척도는 기존과 같이 높을수록 모델의 성능이 좋음을 의미한다. 또한, 본 논문에서는 Hamming loss를 사용해 예측 오류율과 미발견 오류율 평가하였다.

$$HammingLoss, HL = \frac{1}{kn}\sum_{i=1}^{n}\sum_{l=1}^{k}[I(l \in Z_i \wedge l \notin Y_i) + I(l \notin Z_i \wedge l \in Y_i)]$$

여기서, I는 지시 함수이고, k는 데이터 집합 T가 소유한 모든 레이블의 개수이다. Hamming loss는 작을수록 모델의 성능이 좋음을 의미한다.

4.4 실험 결과

본 논문에서는 의사결정 트리(DT), 랜덤 포레스트(RF), 심층 신경망(DNN) 기계학습 모델을 사용해 실험을 진행하고 결과를 평가했다. 모든 평가는 5-fold 교차 검증을 통해 진행되었고, 평가 결과는 표 2에 요약되어 있다.

평가 결과에 따르면, 심층 신경망이 모든 평가 척도에 대해 우위에 있음을 볼 수 있다. 의사결정 트리와 랜덤 포레스트의 경우 다중-레이블 분류 문제를 지원하는 기계학습 기법인데도 불구하고 그렇지 않은 심층 신경망 기법보다 보다 낮은 성능을 보였다.

표 3은 가장 성능이 높은 심층 신경망 모델이 생성한 소견 문장의 예시를 보여준다.

표 2 기계학습 모델에 따른 성능 비교

모델	A	P	R	F	HL
DT	0.712	0.816	0.810	0.810	0.023
RF	0.760	0.902	0.804	0.846	0.016
DNN	0.821	0.904	0.879	0.888	0.012

표 3 심층 신경망 모델에 의해 생성된 소견 문장

실제 소견 문장	생성 소견 문장
B형간염항체 양성으로 B형간염 면역상태입니다.	B형간염항체 양성으로 B형간염 면역상태입니다.
간 기능 검사의 결과는 정상입니다.	간 기능 검사의 결과는 정상입니다.
공복 혈당이 정상입니다. 당뇨 검사의 결과가 정상입니다.	공복 혈당이 정상입니다. 당뇨 검사의 결과가 정상입니다.
	백혈구수(WBC)가 다소 감소되었습니다. 백혈구가 심하게 감소하면 면역기능이 손상됩니다. 일내변화를 보이므로 주기적으로 검사하면 도움이 됩니다.
빈혈이 없습니다.	빈혈이 없습니다.
신장기능 검사 결과는 정상입니다.	신장기능 검사 결과는 정상입니다.
지질검사의 결과 정상 입니다.	지질검사의 결과 정상 입니다.
혈구질환 관련 검사의 결과 정상입니다.	

5. 결론

본 논문에서는 기계학습을 이용하여 혈액검사 데이터로부터 전문가 소견을 생성하는 방법을 제안하였다. 의사결정지원 시스템을 구축하기 위해 규칙 베이스를 사용하던 기존의 방법과 달리, 기계학습을 통해 의미 있는 수준의 시스템을 구축할 수 있음을 볼 수 있었다.

혈액검사는 다중-레이블 분류 문제이기 때문에, 이를 다룰 수 있는 적절한 기계학습 방법을 선택하는 것이 중요하였다. 따라서 본 논문에서는 의사결정 트리, 랜덤 포레스트, 심층 신경망 기계학습 기법을 통해 혈액검사의 소견 문장을 생성하는 방법에 대해 살펴보았다.

또한, 다중-레이블 분류 문제를 다루는 모델의 성능을 평가하기 위해 기존의 척도가 아닌 다중-레이블 분류 문제를 위해 정의된 평가 척도를 사용하였다. 그 결과, 심층 신경망을 사용한 모델에서 가장 높은 성능을 확인할 수 있었다.

후속 계획으로 다중-레이블 분류 문제를 위한 학습 방법인 BP-MLL을 심층 신경망에 적용해, 모델의 성능을 더욱 향상시킬 예정이다. [10]

6. 사사

본 연구는 산업통상자원부 및 한국산업기술평가관리원의 산업핵심기술개발사업(지식서비스)의 일환으로 수행하였음. [10052955, 현장 전문가의 경험지식 획득 및 활용을 위한 경험지식플랫폼 개발 연구]

7. 참고문헌

[1] 통계청,「생명표, 국가승인통계 제10135호」

[2] Keen, Peter GW. "Decision support systems: a research perspective." Decision Support Systems: Issues and Challenges (New York: Pergamon Press, 1980). 23-44. 1980.

[3] Tsoumakas, Grigorios, and Ioannis Katakis. "Multi-label classification: An overview." Dept. of Informatics, Aristotle University of Thessaloniki, Greece. 2006.

[4] Polat, Kemal, and Salih Güneş. "An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease." Digital Signal Processing 17.4. 702-710. 2007.

[5] Liu, Siqi, et al. "Early diagnosis of Alzheimer's disease with deep learning." Biomedical Imaging (ISBI), 2014 IEEE 11th International Symposium on. IEEE, 2014.

[6] Elshami, Eyad H., and Alaa M. Alhalees. "Automated Diagnosis of Thalassemia Based on DataMining Classifiers." The International Conference on Informatics and Applications (ICIA2012). The Society of Digital Information and Wireless Communication, 2012.

[7] SEYMEN, Res Asst Volkan, Res Asst Gamze DOĞALI ÇETİN, and Devrim AKGÜN. "The Diagnosis of Iron-Deficiency Anemia using Feedforward Backpropagation Neural Network." Hemoglobin (HGB) 12: 16. 2014.

[8] Sorower, Mohammad S. "A literature survey on algorithms for multi-label learning." Oregon State University, Corvallis. 2010.

[9] Godbole, Shantanu, and Sunita Sarawagi. "Discriminative methods for multi-labeled classification." Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 22-30. 2004

[10] Zhang, Min-Ling, and Zhi-Hua Zhou. "Multilabel neural networks with applications to functional genomics and text categorization." Knowledge and Data Engineering, IEEE Transactions on 18.10. 1338-1351. 2006