Decision Tree와 SNP정보를 이용한 간경화 환자의 감수성 예측

김동회[0*], 엄상용[*] , 조성원[**] , 함기백[**], 김진[*]
[*] 한림대학교 컴퓨터공학과
[**] 아주대학교 간소화기 질환센터
{kdh[0*], Suhmn, jinkim[*]}@hallym.ac.kr , sung_woncho[**]@hotmail.com ,kibaek[**]@ajou.ac.kr

Pridict of Liver cirrhosis susceptibility using Decision tree with SNP

Dong Hoi, Kim[0*], Saangyong Uhmn[*], Sung Won Cho[**] ,Ki Baek, Ham[**] , Jin Kim[*]
[*]Dept of Computer Engineering, Hallym University
[**]Ajou university Medical Center Genomic Research Center for Gastroenteroloy

요 약

본 논문에서는 SNP데이터를 이용하여 간경화에 대한 감수성을 예측하기 위해 의사결정 트리를 이용하였다. 데이터는 간경화 환자와 정상환자 총 116명의 데이터를 사용하였으며, Feature 값으로는 간질환과 밀접한 연관성을 갖는 28개의 SNP데이터를 사용하였다. 실험방법은 각각의 SNP에 대하여 의사결정트리로 분류율을 측정한 후 가장 높은 분류율을 가지는 SNP부터 조합해 나가는 방식으로 C4.5 의사결정트리를 이용 leave-one-out cross validation으로 간경화와 정상을 구분하는 정확도를 측정하였다. 실험결과 간 질환 관련 SNP중 IL1RN-S130S, IRNGR2-Q64R, IL-10(-592), IL1B_S35S 4개의 SNP조합에서 65.52%의 정확도를 얻을 수 있었다.

1. 서 론

인간 유전자의 30억개 염기서열 중 개인간 또는 집단간 유전자 차이를 나타내는 하나 혹은 수십개 정도의 염기변이를 Single nucleotide polymorphism(SNP)라 한다. SNP는 인간 염기서열의 0.1%에 해당하는 약 300만개가 존재하는 것으로 알려져 있으나,아직 모든 SNP의 정확한 숫자나 위치는 알려져 있지 않다[1]. 이 SNP의 차이는 사람의 피부색이나 혈액형뿐만 아니라 암, 당뇨, 치매등과 같은 질환 발병에도 관여하는 것으로 알려져 있다. 따라서 질병과 유전자의 연관성 연구가 현재 활발히 진행되고 있다.

기계 학습 기술들은 많은 패턴인식 문제들에 사용되어왔다. 실제로, 의사결정 트리, Neural networks, Support Vector Machine, case based reasoning등과 같은 기계학습방법들은 많은 서로 다른 분야에 폭넓게 적용되었다. 이들 방법들은 공통적으로 일정의 샘플들이 데이터로 주어지면 특성 값들과 분류의 연관성을 가지고 각 샘플의 특성 값에 따라서 분류를 위한 규칙이나 모델을 찾는 것이다. 지금까지의 대부분의 의료정보학 연구는 로지스틱 회귀분석, 판별분석과 같은 전형적인 통계학적 분석방법을 이용해 왔으나 최근 기계학습을 이용한 질환예측 및 진단에 대한 연구가 활발히 진행되고 있다.

간경화란 간세포의 파괴되어 그 부분이 결체조직으로 바뀌면서 간이 굳어지게 된 상태로 원상회복이 어렵고 여러 가지 합병증을 유발하는 질환으로 원인으로는 B형 간염 바이러스 감염이 장기간 지속되어 생기는 경우가 가장 많다.간염이란 간세포 조직의 염증을 의미한다. 간염바이러스는 A,B,C,D,E형(HAV,HBV,HCV,HDC,HEV)이 있으며 이중 B형 간염 바이러스는 우리나라 만성 간 질환의 가장 흔한 원인으로 한국 성인의 7%정도는 B형 간염 바이러스 보유자이다[2]. 간염이 6개월 이상 진행된 것을 만성 간염이라고 하며, HBV(Hepatitis B Virus)로 인한 만성간염환자는 세계적으로 3억5천여명정도로 추정된다[3]. HBV의 보균자는 만성간염, 간경화, 간부전증, 간암으로 발전할 위험성을 가지고 있다. 간경화나 간암은 임상적으로는 거의 증상이 나타나기 않기 때문에 정기적 검사만이 최선이다. 따라서 간경화를 조기에 예측할 수 있는 예측모델이 필요하다.

본 논문에서는 SNP와 의사결정 트리를 간경화에 걸릴 수 있는 감수성(Susceptibility)을 예측하기 위한 모델로 사용하였으며, 예측율을 높이기 위한 방법으로 최적의 특징값들을 만들기 위해 각각의 SNP에 대한 분류율을 측정한 후 높은 분류율을 보이는 SNP부터 조합해 나가는 방법으로 예측율을 향상시켰다.

본 논문의 2장에서는 보다 구체적인 배경연구에 대하여 설명하며, 3장에서는 Decision tree를 이용한 감수성 예측실험 및 결과에 대하여 논한다. 마지막으로 결론 및 향후 연구과제에 대하여 논한다.

2. 배경연구

본 연구는 보건복지부 보건의료기반 진흥사업(01-PJ6-01GN14-0007)의 지원에 의해 이루어진 것임

2.1 SNP

단일염기다형성(SNP:Single Nucleotide Polymorphism)은 개인과 개인 혹은 집단간의 DNA에 존재하는 한 염기쌍의 차이를 말한다. 이 SNP는 약 30억개의 인간 유전자의 유전자 코드A(adenine),T(thymine),C(cytosine),또는 G(guanine)중 1000개의 염기당 1개꼴로 하나의 염기가 다른 염기로 치환되어 나타나는 변이부분으로 이로 인하여 유전적 다양성이 발생한다. 그림 1은 SNP에 대한 예를 나타낸다.

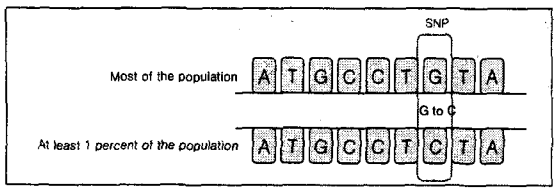

그림 1. SNP 예

SNP의 변이는 사람의 피부색이나 혈액형뿐만 아니라 암, 당뇨, 치매등과 같은 질환에 연관되어 질환에 대한 감수성 및 약물 반응성의 차이를 가져온다는 것으로 알려져 있다. SNP가 임상의학에 있어서 응용될 수 있는 분야는 첫째 변이가 있는 유전자가 질병의 원인이 되는 유전자일 경우 해당 유전자를 타겟으로 하여 진단 및 치료에 이용할 수 있다. 둘째 변이가 있는 유전자가 개개인의 질병 발생에 있어 유전자적 배경(genetic background)이 되는 경우 개개인의 질병에 대한 감수성(susceptibility)에 관여하는 SNP는 질병의 위험인자가 됨으로 질병 예방측면에서 이용되게 된다. 마지막으로 약물 또는 치료 반응에 관여하는 SNP의 경우 개인의 유전자 다형성에 따른 맞춤치료의 가능성을 열어 현재 매우 각광을 받고 있는 분야이다.

2.2 의사결정 트리

인간 유전체 분석은 방대한 양의 데이터로부터 생물학적 또는 의학적 의미를 만들어 내는 일로, 많은 기계 학습 기술들이 이 분야에 적용되고 있다. 의사결정트리는 데이터 마이닝의 분류 작업에 주로 사용되는 기법으로 다음의 몇 가지 중요한 이유로 다양한 분야에서 효율적으로 사용될 수 있다. 첫째 분류나 예측 근거를 제공하기 때문에 사용자가 이해하기 쉽다[4,5]. 둘째 의사결정 트리는 모델 구축시 분류에 영향을 미치지 않는 속성들을 자동으로 제외시킴으로 데이터 선정이 용이하다. 셋째 다른 방법들과 비교하였을 때 모델 구축에 소요되는 시간이 상대적으로 빠르다[6,7]. 넷째 어떠한 속성들이 각각의 분류에 결정적 영향을 미치는지 파악하기 쉽다. 마지막으로 의사결정트리의 정확도는 다른 방법들과 동등하거나 뛰어나다[8,9]. 이러한 특성들로 최근 의사결정 트리 분류기는 레이더 신호 분류, 문자인식, 원격 센싱 , 의료진단, 전문가 시스템 , 발음인식 등과 같은 다양한 분야에 성공적으로 사용되고 있다.

의사결정트리의 예는 그림2에서 보는 것과 같다. 의사결정트리 분류기는 입력 데이터에 대하여 클래스 예측에 성공적으로 적용될 수 있는 체계적인 규칙의 집합이다. 이들 규칙은 데이터를 둘 혹은 그 이상의 그룹으로 분할케 되며, 각 분할들은 동질의 데이터로 구성되는 자식노드를 가진다. 각 분할 노드들은 그림 2에서 사각형으로 표현되는 부분은 클래스를 구분하기 위한 최소한의 조건을 가지는 마디(Node)를 의미한다. 최상위에 위치하는 마디를 루트 마디(Root Node)라한다. 더 이상 분할이 일어나지 않는 시점의 마디가 종단 마디(Leaf Node) 가 된다. 이 때 가지 루트마디로부터 이어지는 단말마디까지의 노드의 집합으로 표현 되는 경로가 각 클래스를 구분하기 위한 최소한의 조건을 의미한다. 그리고 단말에 원으로 표현되는 부분은 해당 조건 경로를 가질 때의 클래스를 의미한다.

위에서 언급한 특성과 더불어 의사 결정 트리는 특징 값들로 수치 데이터와 더불어 기호데이터도 허용됨으로서 A,T,G,C 염기로 표현되는 SNP 데이터를 표현하기에도 적합하다. 본 논문에서는 의사결정 트리 알고리즘 중 가장 잘 알려져 있고 가장 폭넓게 사용되고 있는 C.4.5 Release 8을 사용하였다.

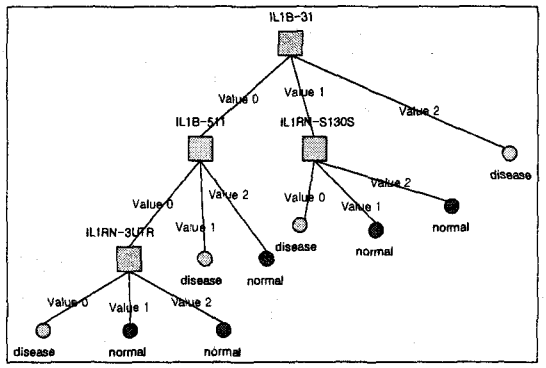

그림 2. 의사결정 트리 예

3. 실험 방법 및 결과

3.1 데이터

본 논문에서 실험을 위해 표 1에서 보여 지는 것과 같이 28개의 SNP 정보를 가지는 116명의 데이터를 사용하였다. 이 데이터는 아주대학교병원 간, 소화기 질환 유전체 연구센터에서 발취하였다.[11]. SNP는 간 질환과 밀접한 연관을 가질 것으로 연구되고 있는 후보유전자들을 선택하였다.

표 1. 28개의 SNP와 환자수

No	SNPs	Sequence	환자수
1	CCR5(-2459)	G/A	간경화:58 정상:58
2	RANTES(-403)	G/A	
3	MCP1(-2518)	G/A	
4	CCRR2-V64I	G/A	
5	CXCR1-S276T	C/G	
6	CXCR4-I138I	G/A	

7	IL1B-31	C/T	
8	IL1B-511	C/T	
9	ILIRN-S130S	C/T	
10	IL1RN-3UTR	C/G	
11	MBP G54D	G/A	
12	IRF1(-410)	G/A	
13	IFNGR2-Q64R	G/A	
14	IRF1(-388)	C/T	
15	IL-10(-592)	A/C	
16	IL-10(-1082)	G/A	
17	IFNGR1(-56)	C/T	
18	IFNGR1(+95)	C/T	
19	IFNG(+874)	A/T	
20	TNF-238	G/A	
21	TNF-308	G/A	
22	IL18-S35S	C/A	
23	MMP3-E45K	G/A	
24	MMP3-D96D	C/T	
25	MMP3-A362A	C/T	
26	MMP9-R279Q	G/A	
27	MMP9-Q688R	G/A	
28	MMP9-G607G	C/A	

표2는 실험데이터 표현의 예로 SNP의 표현에서 두 문자가 일치하는 형태를 Home-zygosity , 두 문자가 다른 형태를 Hetero-zygosity라 한다.

표 2. 실험 데이터의 표현

	SNP 1	SNP 2	SNP 3	Class
Patient 1	AA	TT	CA	간경화
Patient 2	AA	CC	AA	정상
Patient 3	GA	CT	AA	간경화

3.2 실험 방법

본 논문에서는 실험을 위해 C4.5 Release 8[12] 의사결정 트리 분류기를 사용하였다. 분석에서는 SNP 데이터를 가지는 116개의 Case에 대해 C4.5 의사결정 트리 분류기를 이용 간경화에 대한 감수성을 예측하였다. 또한 부분적인 특징 값들이 분류에 더 효율적일 경우가 있기 때문에 각 SNP에 대한 분류율을 기준으로 높은 분류율을 보이는 SNP순으로 조합해 나가는 방식으로 예측율을 측정하였다. 실험은 leave-one-out cross validation[13]을 사용하였다. 전체 116명의 Case(간경화 58, 정상 58)에 대해서 의사 결정 트리는 하나의 case를 제외하고 나머지에 대한 트리와 그에 따른 규칙들을 생성한다. 나머지 하나의 case는 생성된 트리를 운행함으로서 클래스를 결정짓는다. 전체 116번의 실험을 통해 각 case의 클래스가 정확히 측정 되었는가에 따라 감수성 예측율이 결정된다. leave-one-out cross validation을 사용한 이유는 간경화 환자에 대한 다량의 Case를 얻는 것이 매우 어렵기 때문이다.

3.3 실험 결과

본 연구의 실험 결과는 정확도,민감도,특이도로 표현된다. 결과에 대한 계산은 표3와 같은 방법으로 산출된다. 정확도는 주어진 데이터에 대하여 얼마나 정확하게 클래스를 구분할 수 있는 가로 질환과 정상을 구분할 수 있는 확률을 뜻한다. 민감도는 질환 환자를 정확히 구분해 낼 수 있는가에 대한 확률이며 특이도는 정상을 정확히 구분해 낼 수 있는가에 대한 확률이다..

기계학습에 있어서 전체 특징값보다 일부 일부 특징값들의 조합이 데이터에 대해 보다 정확한 예측을 하기도 한다. 28개의 SNP에 대해서 가능한 특징 값의 조합 가능 가지수는 2^{28}개이다. 따라서 모든 가능한 특징 값의 조합을 계산한다는 것은 불가능하다. 따라서 본 연구에서는 분류율을 높일 수 있는 방법으로 특징값 축소를 위해 28개 각각의 SNP에 대하여 분류율을 측정하고 분류율이 높은 순서부터 SNP를 조합하여 예측율을 측정하였다.Feature size는 조합된 SNP의 개수 이며, 17개를 제외한 나머지 SNP들은 분류율이 0으로 조합에서 제외시켰다.. 표 4에서 보는 것과 같이 IL1RN-S130S, IRNGR2-Q64R, IL-10(-592), IL1B_S35S 4개의 SNP조합에서 65.52%의 가장 높은 정확도를 보였다.

표 3.정확도 민감도 특이도의 계산

		Liver disease	
		+	-
Test result	+	True Positive(TP)	False Positive(FP)
	-	False Negative(FN)	True Negative(TN)

민감도 = TP/(TP+FN)
특이도 = TN/(TN+FP)
정확도 = (TP+TN)/(TP+TN+FP+FN)

표 4. 실험 결과

Feature Size	정확도
ALL	48.28%
1	63.79%
2	63.79%
3	56.03%
4	65.52%
5	64.66%
6	63.79%
7	63.79%
8	59.48%
9	51.72%
10	52.59%
11	52.59%
12	43.1%
13	43.97%
14	50%
15	50.86%
16	50%
17	50%

그림 3은 조합의 개수에 따른 정확도,민감도,특이도를 나

타낸 그림이며 조합된 Feature size가 4일때 가장 높은 정확도와 민감도를 보였다.

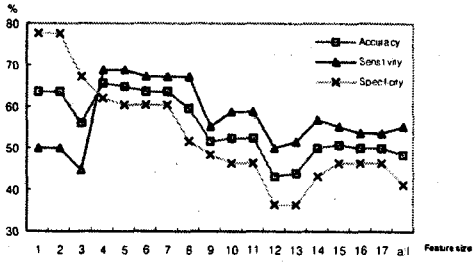

그림 3. Feature size별 민감도,특이도,정확도

의사결정 트리로부터 생성된 의사결정 규칙은 항상 모든 SNP들을 포함하지 않는다. 이는 예측율을 최대화 시킬 수 있는 SNP의 조합들을 의사결정 트리로 부터 구성하여 클래스를 구분하게 된다. 따라서 전체의 SNP보다는 클래스 구분을 최대화 할수 있는 SNP조합을 사용했을때 더 낳은 예측율을 보일수 있다. 여기서 우리는 분류율이 높은 SNP 순으로 조합해 나가는 방식으로 예측율을 높일 수 있었다.

4. 결론 및 향후 연구

본 논문에서는 SNP 데이터에 의사결정트리를 적용하여 간경화 진단의 예측율을 측정하는 문제에 대하여 논의하였다. 실험 결과에서 볼 수 있듯이 의사결정트리는 전체SNP 데이터를 입력으로 하여 48.28%의 간경화 감수성을 보였다. 그러나 단일 SNP에 따른 분류율을 측정하여 분류율이 높은 SNP들을 우선적으로 조합해 나가는 방법으로 65.52%로 간경화 감수성 예측율을 높일 수 있었다. 결과적으로 SNP를 이용한 간경화 감수성 예측에 특징값 선택 기법을 이용 특징값을 최적으로 하여 의사결정 트리 분류기를 적용함으로서 효율적으로 활용될 수 있을 것이다.

향후 더 많은 간 질환 관련 SNP들을 이용해 특징 값을 보다 더 최적화 하고 간 질환과 관련성이 높은 흡연, 음주량, 스트레스 지수와 같은 삶의 질 데이터와 성별, 나이와 같은 개인 정보를 특징 값으로 사용하여 간경화에 대한 감수성 예측율을 향상시키고자 한다.

5. 참고문헌

[1] Anthony J.Brookes " Review The essence of SNPs" GENE pp178 1999.
[2] 한철주, "만성B형간염의 자연경과와 예후", 소화기연관학회 춘계학술대회,pp.354 2005.
[3] Anna S.F.Lok and Brian J.McMahon "Chronic Hepatitis B" AASLD PRACTICE GUIDELINES.
[4] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Classification and Regression Trees. Wadsworth, Belmont, 1984.
[5] M. Mehta, R. Agrawal, and J. Rissanen, SLIQ: A fast scalable classifier for data mining, In Proc. Of the Fifth Int'l Conf. on Extending Database Technology (EDBT), Avignon, France, March 1996.
[6] J. Shafer, R. Agrawal, and M. Mehta, SPRINT: A scalable parallel classifier for data mining, VLDB 1996.
[7] J. Gehrke, R. Ramakrishnan, and V. Ganti, Rainforest A framework for fast decision tree construction of large datasets, VLDB 1996.
[8] S. K. Murthy, On growing better decision trees for data. PhD thesis, Dept. of Computer Science, Johns Hopkins University, Baltimore, Maryland, 1995.
[9] Tjen-Sien Lim, Wei-Yin Loh, and Yu-Shan Shih, An empirical comparison of decision trees and other classification methods, TR-979, Dept. of Stat., Univ. of Wisconsin, Madison, June 1997.
[10] J.R Quinlan, "C4.5:Programs for Machine Learning" Morgan Kaufmann Publishers, San Francisco, CA 1993.
[11] http://www.agcg.re.kr
[12] http://www.rulequest.com/Personal/
[13] B. Efron, Bootstrap methods: Another loot at the jackknife, The annals of Statistics, 7(1):1-26, January 1979.