응용통계연구 (2009)
22(4), 771–780

Random Forests 기법을 이용한 백내장 예측모형 - 일개 대학병원 건강검진 수검자료에서 -

한은정[1] · 송기준[2] · 김동건[3]

[1]국민건강보험공단 건강보험정책연구원, [2]연세대학교 의학전산통계학과,
[3]동덕여자대학교 정보통계학과

(2009년 5월 접수, 2009년 7월 채택)

요 약

백내장 질환은 노령인구가 증가하고 있는 시점에서 사회, 경제적으로 심각한 문제로 부각되고 있는 질병으로 조기 진단이 이루어진다면 발병률을 크게 줄일 수 있는 질병이다. 본 연구에서는 백내장을 조기 진단하기 위한 예측모형을 구축하고자 1994년부터 2001년까지 연세대학병원에서 2회 이상 건강검진을 받고 의사진단을 통해 백내장 여부를 확인할 수 있는 30세 이상 남·녀 3,237명에 대한 건강검진 수검 자료를 활용하여 백내장 발생 위험 예측모형을 개발하였다. 모형개발에는 데이터마이닝 기법인 Random Forests를 사용하였고, 기존의 로지스틱 회귀분석, 판별분석, 의사결정나무 모형(Decision tree), 나이브베이즈(Naive Bayes), 앙상블 모형인 배깅(Bagging)과 아킹(Arcing)을 이용하여 그 성능을 비교 분석하였다. Random Forests를 통해 개발한 백내장 발생 예측모형은 정확도가 67.16%, 민감도가 72.28%였고, 주요 영향요인은 연령, 혈당, 백혈구수치(WBC), 혈소판수치(platelet), 중성지질(triglyceride), BMI였다. 이 결과는 의사의 안과검진 정보 없이 건강검진 수검 자료만으로 백내장 질환 유·무에 관한 정보를 70% 정도 예측할 수 있음을 보여주는 것으로, 백내장의 조기 진단에 많은 기여를 할 것으로 판단된다.

주요용어: Random Forests, 건강검진, 백내장 위험 예측모형, 백내장 질환, 정확도, 민감도.

1. 서론

백내장은 세계적으로 가장 주요한 실명의 원인인데, 백내장으로 인한 실명을 안과적 수술을 통해 막는다고 할지라도 그 수술비용이 매우 높으며 수술이후에도 수정체 후낭의 혼탁과 같은 합병증 발생 위험이 10~50%에 달한다 (Weintraub 등, 2002). 우리나라 65세 이상 노인인구 비율이 2008년에 10.3%를 돌파한 이후 2026년에 20%를 넘는 초고령사회로 진입할 것을 예상하고 있는 가운데 노인성 안질환인 백내장 환자 또한 늘어날 것이며, 이로 인한 의료비는 더 크게 증가할 것으로 내다보고 있다 (통계청, 2008). 실제 2005년부터 2007년까지 백내장으로 지출된 의료비를 살펴보면 2005년 214억원, 2006년 238억원, 2007년 269억원으로 매년 증가하고 있는 것을 알 수 있다 (국민건강보험공단·건강보험심사평가원, 2007).

이와 같이 백내장이 사회·경제적으로 심각한 문제로 부각됨에 따라 해외에서는 미국의 Nurses' Health Study, 유럽의 POLA Study, Blue Mountains Eye Study, Beaver Dam Eye Study, India-US Case-Control Study 등과 같은 안질환과 관련된 위험요인 예측 및 예방을 위한 연구가 활발히 진행되고 있는

[3]교신저자: (136-714) 서울시 성북구 하월곡 2동 23-1, 동덕여자대학교 정보학부 정보통계학과, 부교수.
Email: dongg@dongduk.ac.kr

것에 반해, 국내의 경우 신경환 등 (신경환 등, 1992a)의 연구 외에 백내장 발병 전의 예후 및 위험요인에 대한 연구는 거의 없는 실정이다. 이에 본 연구에서는 일개 대학병원의 건강검진 수검자료를 근거로 백내장의 발생 자체를 줄일 수 있는 백내장 발생 위험 예측모형을 구축하고자 하였다.

백내장 예측모형 구축을 위해 사용된 건강검진 수검자료는 검진자료의 특성상 혈액검사, 대사 및 전해질 검사, 뇨 검사 등 많은 검진항목을 포함하고 있어, 로지스틱모형, 판별모형과 같은 전통적인 통계모형을 사용할 경우 모형의 예측력이 떨어질 수 있으며, 특히 표본의 수가 적을 때 그 문제점은 더욱 커진다 (Bureau 등, 2005; Heidema 등, 2006). 이에 본 연구에서는 백내장과 검진항목간의 연관성을 찾는데 예측력을 높이고자 유전통계학 분야에서 많이 사용되고 있는 Random Forests를 모형구축에 사용하였다. Random Forests는 표본의 수가 적고 모형에 적용되는 독립변수의 수가 많을 때, 독립변수와 특정 질병 간의 연관성을 찾는데 예측력이 탁월한 동시에 과적합을 방지할 수 있는 모형으로 알려져 있다 (Breiman, 2001; Strobl 등, 2007; Lunetta 등, 2004). 또한 Random Forests는 Out-Of-Bag 표본을 제공하는데, 이는 검증용 자료로 활용되어 오류율에 대한 정직한 추정량을 구할 수 있는 동시에 모형에 선택된 독립변수의 중요도를 파악하여 모형에 대한 설명력을 향상시킬 뿐만 아니라 백내장의 주요 영향요인도 파악할 수 있다. 따라서 본 연구에서는 첫째, Random Forests를 통해 백내장 예측모형을 구축하여, 기존의 통계모형과 그 성능을 평가하였으며, 둘째, Random Forests를 통해 백내장질환의 영향요인을 추정하였고, 기존의 통계모형이 추정한 영향요인과 비교분석하였다.

2. 방법

2.1. 연구대상 및 자료수집

본 연구는 1994년 5월 30일부터 2001년 12월 31일까지 연세대학교 의과대학병원에서 2회 이상 건강검진을 받아 의사진단을 통해 백내장 유질환 판정여부를 확인할 수 있는 30세 이상 남녀 3,785명을 대상으로 하였다. 인구사회학적 및 건강검진 수검자료가 결측인 548명을 제외한 3,237명을 최종 분석대상자로 정하였는데, 이때 백내장을 진단받은 자는 544명(16.8%)이었다.

2.2. 분석 변수

본 연구의 종속변수는 백내장 진단 여부로 정하였고, 백내장으로 진단받은 군을 백내장 발생군, 진단 받지 않은 군을 백내장 정상군으로 정하였다. 독립변수는 크게 인구학적 변수, 건강검진 수검 자료와 의사가 백내장 진단시 확인하는 안질환 증상변수로 나누었다. 인구학적 변수는 성, 연령으로 정하였고, 건강검진 수검 자료의 BMI(kg/m^2), 혈액검사, 대사 및 전해질 검사, 뇨 검사, 간 기능 검사, 혈청지질 검사 등의 25개 항목을 모형개발에 활용하였고, 그 세부 항목은 표 2.1과 같다. 의사의 백내장 진단변수를 망막-고혈압성 변화, 망막-당뇨병성 변화, 망막 출혈, 황반부변성의증, 매질혼탁의증, 매질혼탁/백내장검사 요함, 망막병증 정밀검사 요함, 녹내장 정밀검사 요함, 안과 정밀검사 요함 등 9개로 정하였다.

2.3. 분석방법

1) 모형개발 및 평가

백내장 발생 예측모형 구축을 위해 건강검진 수검자료를 학습용자료 75%, 평가용자료 25%로 분할하였다. 모형개발을 위해 Random Forests (Breiman, 2001)를 활용하였고 이 기법으로부터 구축된 모형 결과를 단계적 변수추출방법을 활용한 로지스틱 회귀분석, 판별분석(LDA, QDA), 나이브베이즈, 의사결정나무모형, 배깅, 아킹 등의 모형과 비교하였다. 개발된 모형은 정분류율과 민감도 및 특이도를 통해

표 2.1. 백내장 예측모형에 사용된 건강검진 항목

검사	독립변수	
신체 계측	BMI	kg/m^2
혈액검사	WBC	백혈구 수치
	Platelet	혈소판 수치
대사 및 전해질 검사	Na	나트륨
	K	칼륨
	CI	염소
	Ca	칼슘
	P	인
	Co2	체내 이산화탄소량
	Creatinine	신장에 쌓인 노폐물 수치
뇨 검사	Glucose	혈당
	BUN(blood urea nitrogen)	혈중 요소질소
	Ph	소변 속의 산성도 측정
	Uro-bilirubin	소변 속의 bilirubin
	Uric acid	요산
간 기능 검사	Protein	간의 단백질
	Albumin	알부민
	Bilirubin	담즙속의 적황색 색소
	AST(aspartic acid transaminase)	간세포 혈청 효소
	ALT(alanine transaminase)	간세포 혈청 효소
	ALP(alkaline phosphatase)	간세포 혈청 효소
	γ-GTP	간세포 혈청 효소
혈청지질	Cholesterol	콜레스테롤
	Triglycerides	중성지질
	HDL	고농축 지질단백질

그 성능을 평가하였고, 분석모형 중 변수의 중요도에 대한 정보를 제공해주는 Random Forests와 의사결정나무모형에서 제공되는 변수의 중요도와 로지스틱 회귀분석과 판별분석의 변수선택에 의해서 결정된 주요 위험요인들을 비교분석하였다.

모형개발에 사용한 Random Forests는 다수의 붓스트랩 표본에서 생성된 나무모형을 다수결원칙으로 결합하여 안정성을 향상시키는 배깅 방법과 유사하지만, 각 붓스트랩 표본에서 무작위로 선택된 소수의 독립변수만을 사용한다는 점에서 배깅 방법과 다르다. 일반적으로 결합모형에 포함된 개별 분류자의 성능과 모형에 포함된 개별 분류자간의 상관관계사이에는 상충관계가 존재하는데, 개별분류자의 성능이 높을수록, 그리고 모형내 개별분류자간의 상관관계가 낮을수록 결합모형의 성능은 향상된다. Random Forests는 결합모형의 이러한 상충관계를 조정하기 위하여 붓스트랩 표본에 대하여 소수의 독립변수를 무작위 추출하여 나무모형을 적합하되, 가지치기 없이 최대한 적합시킨다. Random Forests의 예측성능은 배깅이나 부스팅과 같은 적중률이 탁월한 예측모형에 뒤떨어지지 않으며, 또한 최신기법의 공통적인 약점인 과적합을 피할 수 있는 것으로 알려져 있다. 또한 Tibshirani (1996)과 Wolpert와 Macready (1999)가 제안한 Out-Of-Bag 표본을 적극 활용하는데, 모형 적합시 붓스트랩 표본추출과정에서 배제되는 31.8% 정도의 Out-Of-Bag 자료를 검증용 자료로 활용하여 오류율에 대한 정직한 추정량을 구하는 동시에 선택된 독립변수의 중요도를 파악하여 모형에 대한 설명력도 향상된다.

모형개발을 위한 툴은 기초통계분석, 로지스틱분석 및 판별분석에는 SAS 9.1을 사용하였고, Random

Forests, 의사결정나무모형, 나이브베이즈, 배깅 그리고 아킹에는 R을 사용하였다.

2) 비용-민감도 달성

백내장 정상군의 수가 발생군의 수에 비해 약 5배 정도 높은 비율을 차지한다. 이와 같은 자료의 불균형은 민감도를 떨어뜨리는 경향이 있고, 만약 오분류 비용을 무시하고 단순히 오분류 확률만을 최소화 하는 방법으로 분류를 시행한다면 분류 결정시 문제가 발생할 수 있다. 본 연구의 목적이 백내장에 걸린 대상자를 더 정확하게 분류해내는 것이므로, 분류의 민감도를 높이기 위해 오분류 확률과 더불어 오분류 비용을 고려하는 것이 바람직하다. 따라서 자료의 불균형 문제의 해결과 오분류 비용의 적용을 위해 Elkan (2001)에 의해 소개된 이론을 적용하였다. 이 방법은 백내장 발생군을 정상군으로 분류하는 비용을 정상군을 위험군으로 분류하는 비용보다 더 높게 정의하여 분석용자료에서 발생군과 정상군의 비율을 변화시키는 것이다. 이에 분류 비용을 1, 3, 5, 7배에 따라 분석을 실시하였으며, 비용을 5배로 주었을 때 민감도, 특이도, 정확도가 적정한 수준을 나타내어 모든 모형을 비용 5에서 비교설명하였다.

3. 연구결과

1) 백내장 발생의 주요 영향요인

백내장 발생군과 백내장 정상군 간의 인구학적 변수의 차이를 살펴보면, 백내장 발생군의 연령이 평균 64.0(±10.1)세로 백내장 정상군보다 10세 정도 연령이 높아 유의한 차이를 나타내었고, 성별은 남자의 비율이 각각 48.9%와 45.2%로 비슷한 분포를 나타내었다. 건강검진 수검자료에서 백내장 발생군이 백내장 정상군에 비해 혈소판($p = 0.02$), Ca($p = 0.02$), 알부민($p \leq 0.01$)의 평균 수치가 통계적으로 유의하게 낮았고, Na($p = 0.03$), K($p = 0.02$), 혈당($p \leq 0.01$), BUN($p \leq 0.01$), Uro-bilirubin($p = 0.02$), ALP($p \leq 0.01$), 콜레스테롤($p \leq 0.01$)의 평균 수치는 높게 나타났다. 의사 백내장 진단변수의 경우, 백내장 발생군에서 망막-고혈압성 변화, 망막-당뇨병성 변화, 매질환탁의증 등의 증상이 발생한 비율과 매질혼탁 및 백내장 검사를 요하는 대상자의 비율이 정상군 보다 높게 나타났고, 이와 반대로 녹내장 검사를 요하는 사람의 분포는 백내장 정상군보다 8.8% 낮게 나타났다 (표 3.1).

2) 모형개발 결과

백내장 발생 예측모형은 Random Forests를 통해 개발하였고, 그 성능을 기존의 데이터마이닝 기법과 비교분석하였다. 표 3.2의 학습용자료의 분석결과를 살펴보면 Random Forests와 배깅은 99%이상의 매우 높은 정확도를 보여주고 있고, 민감도와 특이도도 100%의 높은 예측력을 나타낸다. 기존의 데이터마이닝 모형 중에서는 의사결정나무모형이 정확도 76.69%로 높게 나타났고, 반면 나이브베이즈는 51.17%로 가장 낮은 예측력을 보였다. 평가용자료에서는 Random Forests와 LDA에서 각각 67.89%, 66.05%로 정확도가 가장 높았고, 나이브베이즈는 정확도가 45.76%로 가장 낮게 나타났다. Random Forests는 정확도(67.89%), 민감도(71.33%), 특이도(63.63%) 모두에서 다른 모형보다 높은 성능을 나타냈다.

그림 3.1은 Random Forests에서 제공하는 Out-Of-Bag 자료의 오류율을 나타낸다. 그림을 살펴보면 학습용자료와 평가용자료 그리고 Out-Of-Bag의 오류율은 모두 나무숲이 커질수록 일정한 값으로 수렴하며, 특히, Out-Of-Bag 자료의 오류율이 학습용자료와 평가용자료의 오류율보다 더 안정적으로 수렴하는 것을 볼 수 있다. 이때 Out-Of-Bag 자료의 오류율 32.18%는 Random Forests의 평가용자료의 32.11%와 유사하며, 이는 Out-Of-Bag 자료를 통해 평가용자료의 역할을 대신할 수 있음을 나타낸다.

표 3.1. 백내장 발생 위험요인 분석(단변량 분석)

변수			백내장 발생군		백내장 정상군		p-value
			N	%	N	%	
전체			544	16.8	2,693	83.2	
인구학적 변수		연령	64.0± 10.1		54.3± 12.3		<0.01
		성별					
		남자	266	48.9	1,218	45.2	0.12
		여자	278	51.1	1,475	54.8	
건강검진 수검자료	신체 계측	BMI	29.8± 52.0		30.6± 53.4		0.75
	혈액검사	WBC	6.7± 1.8		6.6± 2.8		0.34
		Platelet	233.0± 65.1		240.0± 59.1		0.02
	대사 및 전해질 검사	Na	141.5± 2.3		141.2± 7.0		0.03
		K	4.21± 0.37		4.17± 0.39		0.02
		CI	104.1± 3.1		103.8± 5.7		0.06
		Ca	9.3± 0.4		9.4± 1.7		0.02
		P	3.5± 0.5		3.6± 0.6		0.25
		Co2	25.2± 2.4		25.1± 2.6		0.44
		Creatinine	1.0± 0.3		1.0± 0.2		0.06
	뇨 검사	Glucose	114.6± 47.3		98.5± 30.8		<0.01
		BUN	15.6± 4.7		14.5± 4.2		<0.01
		Ph	5.7± 1.0		5.6± 1.0		0.46
		Uro-bilirubin	0.2± 0.6		0.1± 0.2		0.02
		Uric acid	4.8± 1.4		4.8± 1.4		0.98
	간기능 검사	Protein	7.4± 0.4		7.4± 0.6		0.41
		Albumin	4.5± 0.3		4.6± 0.4		<0.01
		Bilirubin	0.8± 0.4		0.8± 0.4		0.28
		AST	21.8± 13.2		21.3± 15.3		0.40
		ALT	22.2± 16.2		22.9± 22.6		0.36
		ALP	79.6± 25.9		73.5± 29.8		<0.01
		γ-GTP	33.5± 52.1		29.7± 37.9		0.11
	혈청지질	Cholesterol	205.6± 37.2		199.6± 38.3		<0.01
		Triglycerides	157.3±118.6		146.9±100.8		0.06
		HDL	49.0± 12.3		50.1± 13.0		0.07
의사 백내장 진단변수		망막-고혈압성 변화	193	35.5	792	29.4	<0.01
		망막-당뇨병성 변화	6	1.1	7	0.3	<0.01
		망막 출혈	2	0.4	8	0.3	0.79
		황반부변성의증	1	0.2	7	0.3	0.74
		매질혼탁의증	36	6.6	28	1.0	<0.01
		매질혼탁/백내장 검사요함	38	7.0	48	1.8	<0.01
		망막병증 검사요함	63	11.6	268	10.0	0.25
		녹내장 검사요함	10	1.8	286	10.6	<0.01
		안과 정밀검사 요함	47	8.6	223	8.3	0.78

3) 중요 변수선택

백내장 발생 예측모형을 추정하는 데에 총 34개의 독립변수를 사용하였다. 모형개발에 사용된 판별분석 방법 중 로지스틱 회귀분석과 판별분석, 의사결정나무, Random Forests 등 4개의 모형이 백내장 질환

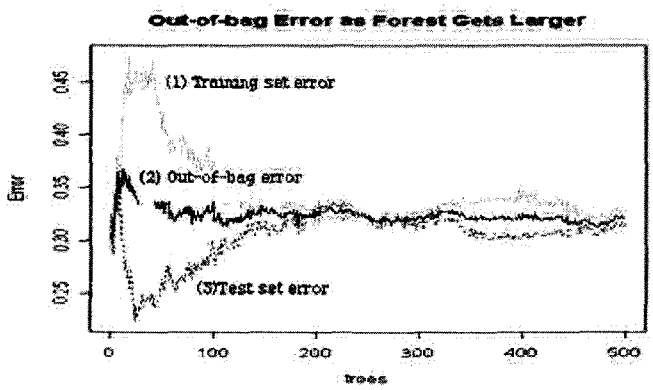

그림 **3.1.** Out-Of-Bag의 오류율

표 **3.2.** 비용 5에서의 모형의 결과표

자료	모형	민감도	특이도	정확도
학습용	Logistic Regression	74.11	73.38	73.74
	LDA	74.11	72.66	73.37
	QDA	58.88	85.85	72.75
	NaiveBayes	1.77	97.84	51.17
	Decision Tree	78.68	74.82	76.69
	Bagging	100.00	99.76	99.88
	Arcing	81.28	64.93	73.12
	Random Forests	100.00	100.00	100.00
평가용	Logistic Regression	66.00	64.46	65.31
	LDA	68.00	63.63	66.05
	QDA	52.00	71.90	60.89
	NaiveBayes	2.66	99.17	45.76
	Decision Tree	62.00	66.11	63.84
	Bagging	67.33	63.63	65.68
	Arcing	69.56	58.64	64.20
	Random Forests	71.33	63.63	67.89

에 영향을 미치는 주요 요인에 대한 정보를 제공해주었는데, 각 모형에서 백내장의 주요 위험요인을 추정한 결과는 표 3.3과 같다. 또한 이 위험요인들을 Random Forests에 적용하여 성능을 분석한 결과는 표 3.4와 같다.

표 3.3과 3.4를 살펴보면, 로지스틱 회귀분석의 경우 백내장 발생의 주요 위험요인을 6개로 추정하고 있는데, 그 중 의사진단변수인 매질혼탁, 백내장 검사요망, 녹내장 검사요망 등이 4개를 차지하고 있었다. 이 변수들을 통하여 백내장 발생군 여부를 예측할 확률은 67.1%로 4개의 모형 중 예측력이 가장 높았다. 판별모형은 백내장의 위험요인으로 추정한 9개의 변수 중 매질혼탁, 백내장 검사요망, 녹내장 검사요망, 고혈압성 변화 등 4개의 변수를 의사진단변수로 포함하고 있었다. 반면, 의사결정나무모형과 Random Forests모형은 건강검진 수검 자료를 백내장 발생의 중요 변수로 추정하고 있는데, 의사결정나무는 연령, 혈당(glucose), 중성지질(triglyceride), 혈소판수치(platelet), 혈중 요소질소(BUN), 알부민(Albumin), ALT, 콜레스테롤(cholesterol) 등 8개의 변수를 백내장 발생 위험요인으로 추정하였

표 3.3. 백내장 발생 위험 요인

모형	위험 요인
Logistic	연령, 혈당(glucose), 매질혼탁, 백내장 검사요함, 녹내장 검사요함(6개)
Discriminant	연령, 혈당(glucose), 매질혼탁, 백내장 검사요함, 녹내장 검사요함, 고혈압성 변화, 크레아티닌(creatinine), 인(P), 유로빌리루빈(Uro-bilirubin)(9개)
Decision Tree	연령, 혈당(glucose), 중성지질(triglyceride), 혈소판수치(platelet), 혈중요소질소(BUN), 알부민(Albumin), ALT, 콜레스테롤(cholesterol)(8개)
Random Forests[1)]	연령, 혈당(glucose), 백혈구 수치(WBC), 혈소판수치(platelet), 중성지질(triglyceride), BMI(6개)

주 1) GINI계수의 감소 정도를 통해 중요 변수 추정

표 3.4. 개별 모형에서 추정한 백내장 질환의 위험 요인을 이용한 Random Forest의 모형결과

자료	모형	민감도	특이도	정확도
학습용	Logistic Regression	82.23	64.98	73.37
	Discriminant	83.50	77.93	80.64
	Decision Tree	100.00	100.00	100.00
	Random Forest[1)]	100.00	100.00	100.00
평가용	Logistic Regression	72.67	60.33	67.16
	Discriminant	69.33	61.15	65.68
	Decision Tree	68.00	65.29	66.79
	Random Forest[1)]	68.00	61.98	65.31

주 1) GINI계수의 감소 정도를 통해 중요 변수 추정

으며, 이를 통한 백내장 발생 정확도는 66.79%를 보였다. 한편 Random Forests는 백내장 발생 위험요인으로 연령, 글루코스(glucose), 백혈구 수치(WBC), 혈소판수치(platelet), 중성지질(triglyceride), BMI를 꼽았고, 이에 대한 정확도는 65.31%로 나타났다.

4. 토의 및 결론

본 연구에서는 건강검진 수검 자료를 활용하여 백내장을 조기진단하기 위한 백내장 발생 예측모형을 구축하였다. Random Forests에 의해 도출된 백내장 발생 예측모형은 정확도가 67.16%, 민감도가 72.28%로 다른 모형에 비해 높았다. Random Forests가 추정한 백내장의 주요 위험 요인은 연령, 혈당, 백혈구 수치(WBC), 혈소판 수치(platelet), 중성지질(triglyceride), BMI였다. 이는 위 예측모형을 통해 의사의 안검진 정보 없이 건강검진 수검 자료만으로 백내장 질환 유·무를 67.16% 예측 할 수 있음을 나타낸다.

Random Forests가 추정한 백내장의 관련 인자 중 연령, 혈당, BMI는 이미 잘 알려진 백내장 위험 요인이다. 신경환 (신경환 등, 1992b) 등은 백내장의 발병은 연령에 따라 필수적으로 동반되는 질환으로 말하고 있고, Delcourt (Delcourt 등, 2000) 등의 코호트 연구에서 80세 이상에서의 발생률이 62.4%에 이른다고 보고하였다. Hennis (Hennis 등, 2004) 등도 연령, 성, 경제력, 당뇨질환의 과거력을 백내장의 주요 위험 요인으로 보고하고 있는데, 특히 당뇨질환의 영향력을 강조하고 있다. 당뇨질환을 예방하고 증진시키는 것이 백내장의 발생을 줄일 수 있는 최우선 과제라고 말하고 있다. BMI 또한 많은 연구에서 백내장 발생의 주요 요인으로 주장하고 있다. Blue Mountains Eye Study (Panchapakesan 등, 2003)는 BMI가 30이상인 사람에서 피질형과 후낭하형 백내장에 걸릴 확률이 높다고 보고하고 있고, Shihpai Eye Study (Kuang 등, 2005)에서는 BMI가 핵형과 피질형의 유의한 위험 요인으로 조사되었

다.

백내장 발생 예측모형에 사용한 Random Forests는 다른 모형에 비해 가장 높은 성능을 보였다. 이는 Random Forests가 50개에서 200여개의 붓스트랩 샘플로부터 다양한 나무분류자를 생성하고, 나무 확장시 독립변수를 제거하지 않고 무작위로 선택하여 모형을 구축하기 때문에 분류의 정확도가 높아진 것으로 볼 수 있다 (Prasad 등, 2006). 모형에 포함된 독립변수를 모두 모형구축에 활용하기 때문에 Random Forests는 독립변수의 수가 많고, 그 독립변수 각각이 가지고 있는 정보가 적은 데이터를 분석할 때 탁월한 성능을 보여준다. 이런 점 때문에 복잡한 인과관계와 많은 유전요인이 작용하여 발생하는 질병의 주요 위험유전인자를 찾는 유전통계학 분야에 많이 활용되고 있다. 또한 본 연구에서 Random Forests이 제공하는 Out-Of-Bag 자료는 검증용 자료의 기능을 대체 할 수 있음을 확인하였다. Random Forest의 개선에 대한 연구도 진행되고 있는데, 소수의 선택된 독립변수를 이용함으로 발생하는 개별분류자의 성능저하문제를 개선하는 방법이나, 단순 다수결원칙이 아닌 각 개별분류자의 성능을 반영한 가중 다수결원칙(Weighted voting)으로 결합하는 방법 등이 제시되었다 (Robnik-Sikonja, 2004).

본 연구는 건강검진 수검자료를 이용하여 다양한 통계방법을 통해 백내장 발생 예측모형을 개발 및 평가하고, 각 모형에서 추정된 위험요인을 비교분석 하였다는데 의의가 있다고 할 수 있다. 백내장발생 예측모형을 통해 백내장질환의 조기진단이라는 예방적 측면에 대한 기틀을 마련할 수 있을 것으로 보이며, 이를 바탕으로 백내장 발생 모니터링 시스템을 구축한다면 백내장 발생률 또한 줄일 수 있을 것으로 기대된다.

참고문헌

국민건강보험공단·건강보험심사평가원 (2007). 2006 건강보험통계연보.

신경환, 김재찬, 김원식, 안병헌, 이진학, 노세현, 송준경, 이용환 (1992a). 한국 백내장 역학 조사회에 의한 노인성 백내장의 재반 위험 인자에 관한 연구 조사, <대한안과학회지>, **33**, 127–134.

신경환, 홍내선, 안상기, 김재찬, 이진학, 안병헌, 김만수, 노세현, 송준경 (1992b). 노인성 백내장의 위험인자 및 환경요소에 대한 역학적 연구: 인구를 기초로 한 역학 조사, <대한안과학회지>, **33**, 834–843.

통계청 (2008). <2008 고령자 통계>, 통계청, 서울.

Bauer, E. and Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants, *Machine Learning*, **36**, 105–139.

Breiman, L. (2001). Random forest, *Machine Learning*, **45**, 5–32.

Bureau, A., Dupuis, J., Falls, K., Lunetta, K. L., Hayward, B., Keith, T. P. and Van Eerdewegh, P. (2005). Identifying SNPs predictive of phenotype using random forests, *Genetic Epidemiology*, **28**, 171–182.

Delcourt, C., Cristol, J. P., Tessier, F., Léger, C. L., Michel, F. and Papoz, L. (2000). Risk factors for cortical, nuclear, and posterior subcapsular cataracts: The POLA study, *American Journal of Epidemiology*, **151**, 497–504.

Elkan, C. (2001). The foundations of cost-sensitive learning, In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence(IJCAI'01)*, 973–978.

Heidema, A. G., Boer, J. M. A., Nagelkerke, N., Mariman, E. C. M., van der A, D. L. and Feskens, E. J. M. (2006). The challenge for genetic epidemiologists: How to analyze large numbers of SNPs in relation to complex disease, *BMC Genetics*, **7**, 23.

Hennis, A., Wu, S. Y., Nemesure, B. and Leske, M. C. (2004). Risk factors for incident cortical and posterior subcapsular lens opacities in the Barbados Eye Studies, *Arch Ophthalmol*, **122**, 525–530.

Kuang, T. M., Tsai, S. Y., Hsu, W. M., Cheng, C. Y., Liu, J. H. and Chou, P. (2005). Body mass index and age-related cataract: The Shihpai Eye Study, *Archives of Ophthalmol*, **123**, 1109–1114

Lunetta, K. L., Hayward, L. B., Segal, J. and Van Eerdewegh, P. (2004). Screening Large-scale association study data: Exploiting interactions using random forests, *BMC Genentics*, **5**, 32.

Panchapakesan, J., Mitchell, P., Tumuluri, K., Rochtchina, E., Foran, S. and Cumming, R, G. (2003). Five year incidence of cataract surgery: The blue mountains eye study, *British Journal of Ophthalmology*, **87**, 168–172.

Prasad, A. M., Iverson, L. R. and Liaw, A. (2006). Newer classification and regression tree techniques: Bagging and random forests for ecological prediction, *Ecosystems*, **9**, 181–199.

Robnik-Sikonja, M. (2004). *Improving Random Forests, Lecture Notes in Computer Science*, Springer, 359–370.

Strobl, C., Boulesteix, A. L., Zeileis, A. and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution, *BMC Bioinformatics*, **8**, 25.

Tibshirani, R. (1996). Bias, Variance and Prediction Error for Classification Rules, Technical Report, Statistics Department, University of Toronto.

Weintraub, J. M., Willett, W. C., Rosner, B., Colditz, G. A., Seddon, J. M. and Hankinson, S, E. (2002). A prospective study of the relationship between body mass index and cataract extraction among US women and men, *International Journal of Obesity*, **26**, 1588–1595.

Wolpert, D. H. and Macready, W. G. (1999). An efficient method to estimate Bagging's generalization error, *Machine Learning*, **35**, 41–55.

A Prediction Model for the Development of Cataract Using Random Forests

Eun-Jeong Han[1] · Kijun Song[2] · Dong-Geon Kim[3]

[1]Research Center, National Health Insurance Corporation;
[2]Department of Biostatistics, Yonsei University;
[3]Department of Statistics and Information Science, Dongduk Women's University

(Received May 2009; accepted July 2009)

Abstract

Cataract is the main cause of blindness and visual impairment, especially, age-related cataract accounts for about half of the 32 million cases of blindness worldwide. As the life expectancy and the expansion of the elderly population are increasing, the cases of cataract increase as well, which causes a serious economic and social problem throughout the country. However, the incidence of cataract can be reduced dramatically through early diagnosis and prevention. In this study, we developed a prediction model of cataracts for early diagnosis using hospital data of 3,237 subjects who received the screening test first and then later visited medical center for cataract check-ups cataract between 1994 and 2005. To develop the prediction model, we used random forests and compared the predictive performance of this model with other common discriminant models such as logistic regression, discriminant model, decision tree, naive Bayes, and two popular ensemble model, bagging and arcing. The accuracy of random forests was 67.16%, sensitivity was 72.28%, and main factors included in this model were age, diabetes, WBC, platelet, triglyceride, BMI and so on. The results showed that it could predict about 70% of cataract existence by screening test without any information from direct eye examination by ophthalmologist. We expect that our model may contribute to diagnose cataract and help preventing cataract in early stages.

Keywords: Random forest, screening test, prediction model of cataracts, accuracy, sensitivity.

[3]Corresponding author: Associate Professor, Department of Statistics and Information Science, Dongduk Women's University, 23-1 Wolgok-Dong, Sungbuk-Gu, Seoul, Korea. E-mail: dongg@dongduk.ac.kr