베이지안 네트워크를 이용한 대사증후군의 예측 모델링

박한샘[1,O] 조성배[1] 이홍규[2]
[1]연세대학교 컴퓨터과학과
[2]서울대학교 의과대학 내과학교실
sammy@sclab.yonsei.ac.kr[O] sbcho@cs.yonsei.ac.kr hkleemd@plaza.snu.ac.kr

Prognostic Modeling of Metabolic Syndrome Using Bayesian Networks

Han-Saem Park[1,O] Sung-Bae Cho[1] and Hong Kyu Lee[2]
[1]Dept. of Computer Science, Yonsei University
[2]Dept. of Internal Medicine, Seoul National University College of Medicine

요 약

대사증후군은 당뇨병, 고혈압, 복부 비만, 고지혈증 등의 질병이 한 개인에게 동시에 발현하는 것을 말한다. 미국에서는 25%이상의 성인이 대사성 증후군인 것으로 알려져 있으며, 경제 여건의 향상 및 식생활 습관의 변화와 함께 최근 우리나라에서도 심각한 문제가 되고 있다. 한편 불확실성의 처리를 위해 많이 사용되고 있는 베이지안 네트워크는 사람이 분석 가능한 확률 기반의 모델로 최근 의학 분야에서 지식 발견, 데이터 마이닝을 위한 도구로 유용하게 사용되고 있다. 본 논문에서는 대사증후군을 예측하는 문제를 다루며, 베이지안 네트워크와 의학 지식을 이용한 대사증후군의 예측 모델을 제안한다. 제안하는 모델을 통해 1993년의 데이터를 가지고 1995년의 상태를 예측하는 분류 실험을 수행하였으며, 실험 결과 다층 신경망, *k*-최근접 이웃 등의 분류기보다 높은 81.5%의 예측율을 보였다.

1. 서 론

대사증후군은 당뇨병, 고혈압, 복부 비만, 고지혈증 등이 한 개인에게 동시에 발현하는 것을 말한다. 대사증후군에 해당하는 환자의 경우 관상동맥질환, 심근경색, 뇌졸중 등의 심혈관계 질환의 발생 비율이 일반인보다 3배정도 높은 것으로 알려져 있으며 이로 인한 사망률은 3~5배가 되는 것으로 알려져 있다[1]. 미국에서는 20세 이상 성인의 약25%가, 50세 이상에서는 약 45%가 대사증후군을 가지는 것으로 알려져 있으며[2], 최근 식생활 습관의 변화와 함께 우리나라에서도 비율이 크게 늘어, 성인 남자의 약 20%가 대사 증후군에 이환되어 있다고 알려져 있다. 이처럼 대사증후군이 사회적인 문제가 됨에 따라, 이를 규명하기 위한 많은 연구가 국내외에서 이루어지고 있다[1-3].

베이지안 네트워크(Bayesian network, BN)는 최근 복잡한 도메인에서 불확실성을 해결하기 위한 강력한 데이터 마이닝 방법으로 부각되고 있다. 베이지안 네트워크는 결합 확률 분포를 이용하는 모델로 도메인 지식을 쉽게 반영할 수 있는 장점을 가지며, 방향성 비순환 그래프(directed acyclic graph, DAG)의 형태를 취한다. 이 그래프에서 노드는 변수를, 노드간의 연결은 확률적인 의존관계를 의미한다[4]. 베이지안 네트워크는 분류 문제를 속성 노드와 결과 노드간의 확률 관계로 가정하며, 이로 인해 수리통계학의 여러가지 장점을 갖는다[5].

베이지안 네트워크는 의학 도메인에서 질병의 진단이나 예측 문제를 해결하기 위해 많이 사용되고 있으며, 또한 좋은 성능을 보여왔다. Antal은 난소암의 임상 모델을 만들기 위해, 그리고 분류를 위해 각각 베이지안 네트워크를 적용하는 연구를 하였다[5, 6]. 또한 Aronsky는 폐렴의 진단을 위해[7], Burnside, Wang은 각각 유방암의 진단을 위해 베이지안 네트워크를 적용하여 성공적인 결과를 얻었다[8, 9]. 그 외에 환자를 돌보는 목적으로 혹은 결핵의 진단을 위한 역학 모델구축을 위해 베이지안 네트워크가 사용된 연구[10, 11] 등 의학 도메인에서 베이지안 네트워크가 적용된 다양한 연구가 진행되어 왔다.

베이지안 네트워크는 역시 분류나 예측 문제를 위해 많이 사용되고 있는 신경망 등의 블랙박스 분류기와 비교해 도메인 지식을 적용하기 쉬우며, 결과의 분석이 가능하다는 장점을 갖는다[9]. 입력으로 연속 값이 아닌 범위가 정해진 상태 값을 사용함으로써 정확도 면에서 단점을 가질 수 있고, 노드 수가 많아지면 실험 시간이 오래 걸리는 등 단점도 존재한다[9]. 하지만 의학 지식을 적용하여 분석 등이 가능한 의학 도메인에서 도메인 지식의 적용 가능성이나, 원인 분석이 가능하다는 특성은 큰 장점이 된다. 본 논문은 베이지안 네트워크와 의학 지식을 이용한 대사증후군의 예측 모델을 제안하며 의학 지식을 예측 모델에 적용하여 실험한 결과 높은 예측율을 얻었다.

2. 배 경

2.1. 대사증후군의 정의

고혈압, 비만, 고중성지방혈증, 저HDL콜레스테롤혈증 등의 심혈관계 위험인자가 함께 나타나는 대사증후군의 개념은1980년대 후반에 정립되었으나, 진단 기준이 불분명하여 표준화된 연구에 어려움을 겪어 왔다[1, 3]. 1998년 WHO에서 대사증후군이라는 명칭과 함께 처음으로 진단 기준을 제안하였다[1]. 그러나 WHO의 진단 기준은 모호하여 임상에서 측정하기가 어려운 측면이 있어 National Cholesterol Education Panel (NCEP)의 Adult Treatment Panel (ATP) III 보고서에서 새로운 진단 기준을 제안하였다. NCEP-ATP III는 다음의 5가지 중 3가지 이상을 만족하는 경우를 대사증후군으로 정의하였다[1].

1) 복부비만: 남성은 허리둘레 >102 cm (>40 in), 여성은 >88 cm (>35 in),
2) 고중성지방혈증 (hypertriglyerides): =150 mg/dL,
3) 저HDL콜레스테롤혈증 (low high-density lipoprotein cholesterol): 남성은 <40 mg/dL, 여성은 <50 mg/dL,
4) 고혈압: =130/=85 mmHg,
5) 고혈당: 공복혈당 =110 mg/dL

서양인들을 위한 이 진단 기준이 아시아인들에게는 적합하지 않다는 주장이 제기되어, 1)의 기준을 남자는 102 cm에서 90 cm로 여자는 90 cm에서 80 cm으로 바꾼 기준이 최근 사용되고 있다[2]. 본 논문에서는 위와 같이 수정된 진단기준인 Modified NCEP-ATP III를 바탕으로 대사증후군을 판별하였다.

2.2. 대상

실험에 사용된 데이터는 원래 지역 사회를 기반으로 한 역학 연구를 위해 조사되었다. 1993년에 경기도 연천군에 거주하는 2520명 중 당뇨병이 없던 2293명을 대상으로 1차 조사를 하였으며, 1995년에 2차 조사 시에는 그 중 1193명을 조사하였다[12]. 각 샘플은 연령, 성별 등의 외형적인 정보부터 혈당, 체질량 지수 등과 같은 신체 지수까지 여러 속성을 포함하고 있으며, 본 논문에서는 이 속성을 바탕으로 Modified NCEP-ATP III의 기준을 적용하여 대사증후군을 판별하였다. 두 차례의 조사에 모두 참가한 1193명 가운데 결측값을 포함한 일부를 제외한 1135명의 샘플을 대상으로 하였으며, 다른 속성과 중복되거나 모든 샘플이 동일한 값을 가져 분류에 영향을 주지 않는 일부 속성을 제외한 18개의 속성을 사용하였다. 표 1은 실험을 위해 사용된 속성을 보여준다.

표 1. 실험에 사용한 속성들

속성	
나이	간수치1(GOT)
성별	간수치2(GPT)
공복혈당	콜레스테롤
경구 당부하 검사 2시간째 혈당	중성 지방
신장	HDL 콜레스테롤
체중	체질량 지수
허리 둘레	엉덩이-허리 둘레 비
엉덩이 둘레	수축기 혈압
고혈압 여부	이완기 혈압

3. 대사증후군의 예측 모델

그림 1은 예측 모델을 구성하고 이를 통해 대사증후군을 예측하는 전체적인 흐름을 보여준다. 대사증후군의 도메인 지식을 데이터의 전처리와 베이지안 네트워크의 속성 선택 과정에 적용함으로써 보다 효과적인 예측 모델을 구성할 수 있다. 모든 속성이 대사증후군의 예측에 중요한 역할을 하지는 않으므로, 보다 많은 정보를 담고 있는 속성을 선택하는 과정은 중요한 역할을 한다.

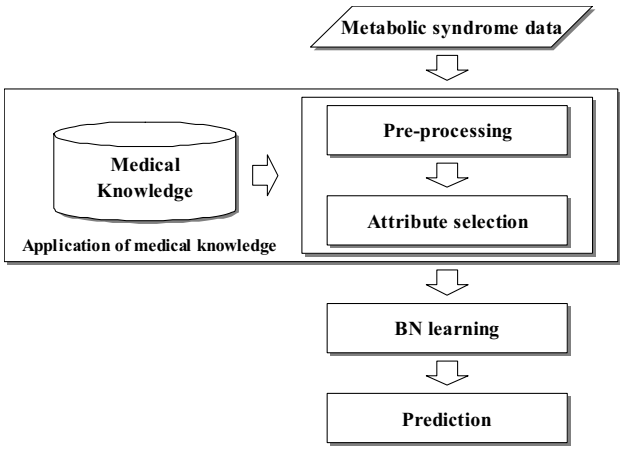

그림 1. 제안하는 방법의 전체적인 흐름도

3.1. 데이터 전처리와 도메인 지식의 적용

데이터가 일부 속성만을 제외하고 연속 값을 갖기 때문에 베이지안 네트워크의 입력으로 사용하기 위해 범위를 정하여 나누었다. 본 논문에서는 이를 위해 대사증후군의 관련 연구를 참고하고[1, 3], 의학 전문가의 도움을 받아 적절한 범위를 결정하여 나누었다. 표 2는 도메인 지식을 적용하여 베이지안 네트워크를 위한 전처리를 마친 후의 속성과 각 속성이 가질 수 있는 상태의 범위를 보여준다.

표 2. 전처리 후의 속성과 각 속성 상태의 범위

속성	상태
나이	young, middle aged, old
성별	male, female
공복혈당	normal, impaired glucose metabolism, diabetes
경구 당부하 검사 2시간째 혈당	normal, impaired glucose metabolism, diabetes
신장	short, medium, tall
체중	light, medium, heavy
허리 둘레	normal, abdominal obesity
엉덩이 둘레	small, large
고혈압 여부	normal, hypertension
간수치1(GOT)	normal, increased
간수치2(GPT)	normal, increased
콜레스테롤	normal, increased
중성 지방	normal, increased
HDL 콜레스테롤	decreased, normal
체질량 지수	low weight, normal, over weight, obesity
엉덩이-허리 둘레 비	normal, abdominal obesity
수축기 혈압	normal, hypertension
이완기 혈압	normal, hypertension

표 2에서 중요한 몇 가지 속성을 살펴보면, 공복혈당의 경우 110보다 작으면 normal 상태가 되고 110 ~ 125의 값을 가지면 impaired glucose metabolism, 126이상이면 diabetes 상태가 된다. 중성 지방의 경우 150 혹은 200을 기준으로 높은 경우 increased 상태(고중성지방혈증)가 되는데 최근에는 150을 기준으로 삼는 추세이다. HDL 콜레스테롤의 경우 다른 콜레스테롤 수치와는 달리 기준보다 낮으면 병적 상태이다. 남자에서는 40미만, 여자에서는 50미만이면 decreased 상태(저HDL콜레스토롤혈증)이다. 각 속성들의 상태를 나누기 위해 본 논문에서 사용된 기준은 대부분 실제로 임상에서 사용되고 있는 의학적으로 유의한 수치이다[3].

3.2. 베이지안 네트워크와 의학지식을 이용한 예측 모델

식 (1)은 베이지안 네트워크를 나타내는 것으로 B는 베이지안 네트워크의 구조를, θ_B는 확률 변수를 의미하며 $P<B,\theta_B>$는 네트워크의 모든 결합 확률 분포를 나타낸다.

$$P < B,\theta_B >= P(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} P(x_i \mid Pa(x_i)) \qquad (1)$$

3.1의 과정을 통해 데이터 전처리를 거친 후 도메인 지식을 이용해 예측 모델을 구성하기 위해 사용될 속성을 선택하였다. 앞에서 설명된 대사증후군의 정의에 의하면 기본적으로 대사증후군의 결정을 위해 필요한 속성은 8가지이다. 그러나 그 외에도 나이, 경구 당부하 검사 2시간째 혈당, 고혈압, 체질량 지수는 대사증후군에 영향을 끼치는 것으로 알려져 있다[13]. 이 중 고혈압과 체질량 지수는 다른 속성과 중복되는 정보를 포함하는 속성이다. 따라서 본 논문은 대사증후군의 정의에 사용되는 기본 8속성에 나이, 경구 당부하 검사 2시간째 혈당의 두 속성을 더한 10가지 속성을 이용한 베이지안 네트워크 모델을 제안한다.

의학지식을 적용해 선택된 10가지 속성의 데이터를 이용해 10개의 노드를 갖는 베이지안 네트워크를 학습하였다. 학습에는 K2 알고리즘이 사용되었으며 각 노드의 조건부 확률 테이

블은 학습 데이터의 빈도수를 이용해 계산하였다. 그림 2는 학습 데이터를 이용해 학습된 베이지안 네트워크의 구조를 보여준다.

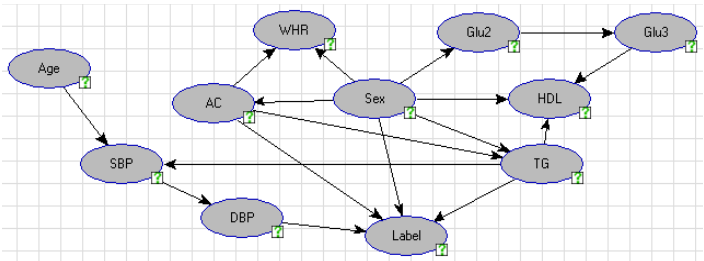

그림 6. 10가지 속성을 이용해 학습된 베이지안 네트워크

4. 실험 및 결과

제안하는 모델의 유용성을 보이기 위해 두 가지의 비교 실험을 수행하였다. 두 실험 모두 10-fold cross validation을 수행하였으며 각 실험은 3번씩 반복하였다.

첫 번째 실험은 속성을 달리한 베이지안 네트워크 모델간의 비교 실험으로 기본 8속성을 이용한 모델, 10속성을 이용하는 제안하는 모델, 중복 정보를 포함한 12속성을 이용한 모델, 그리고 전체 18속성을 이용한 모델을 이용한 경우의 예측율을 비교하였다. 표 3을 보면 10속성을 갖는 제안하는 모델이 가장 높은 예측율을 보이는 것을 확인할 수 있으며 이 차이는 통계적으로 유의한 값이다. (p<0.05)

표 3. 선택속성에 따른 예측율 비교

속성 수	8	10	12	18
예측율(%)	64.01	**81.53**	77.40	76.43

표 4는 예측 모델을 바꿔가며 예측율을 비교한 결과를 정리한 것이다. 신경망(10-20-2)과 k-최근접 이웃(k=3)보다 제안하는 BN 모델이 더 높은 예측율을 보이고 있으며 역시 통계적으로 유의한 차이를 보인다. (p<0.001)

표 4. 예측 모델에 따른 예측율 비교

속성 수	신경망	3-NN	BN 모델
예측율(%)	74.01	65.92	**81.53**

그림 3과 4는 각각 표 3과 4를 그래프로 표현한 것으로 그래프의 x축은 예측 모델을, y축은 예측율을 나타낸다.

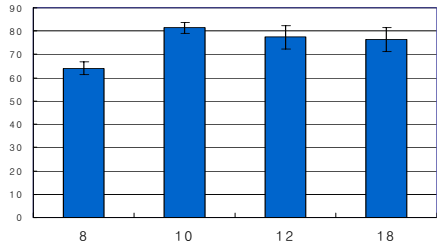

그림 3. 선택 속성에 따른 예측율 비교 그래프

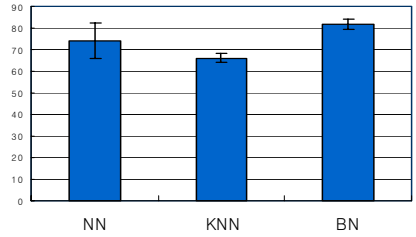

그림 4 예측 모델에 따른 예측율 비교 그래프

일반적으로 신경망 모델은 분석을 할 수 없는 단점이 있지만 BN 모델보다 높은 분류율을 보인다. 하지만 그림 4는 반대의 결과를 보이고 있으며 이는 예측 모델의 구성을 위한 데이터 전처리 및 속성 선택 과정에서 도메인 지식이 적절히 사용되었기 때문이라고 해석될 수 있다.

5. 결론

식습관 및 생활습관의 변화로 최근 우리나라에서도 대사증후군이 심각한 문제가 되고 있다. 대사증후군의 예측을 위해 본 논문은 베이지안 네트워크와 의학 지식을 이용한 예측 모델을 제안한다. 데이터의 전처리와 속성 선택 과정에 의학 지식을 적용하여 예측 모델을 구성하였으며 제안한 모델은 다른 속성을 이용한 모델이나 신경망, k-최근접 이웃보다 더 높은 81.5%의 예측율을 보임을 확인하였다.

감사의 글

본 논문은 보건복지부의 보건의료기술진흥사업의 지원에 의하여 이루어진 것임.

참고 문헌

[1] H. K. Lee, et al., "The prevalence of metabolic syndrome and its relation with chronic complications in Korean type 2 diabetic patients," *Korean Society of Lipidology and Atherosclerosis*, vol. 13, pp. 382-391, 2003.

[2] N. N. Mehta and M. P. Reilly, "Mechanisms of the metabolic syndrome," *Drug Discovery Today: Disease Mechanisms*, vol. 1, no. 2, 2004.

[3] H. K. Lee, et al., "Metabolic syndrome," *The Korean Society of Endocrinology*, vol. 18, pp. 105-116, 2003.

[4] P. Larranaga, et al., "Structure learning of Bayesian networks by genetic algorithms: A performance analysis of control parameters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, pp. 912-926, 1996.

[5] P. Antal, et al., "Using literature and data to learn Bayesian networks as clinical models of ovarian tumors," *Artificial Intelligence in Medicine*, vol. 30. pp. 257-281, 2004.

[6] P. Antal, et al., "Bayesian applications of belief networks and multilayer perceptrons for ovarian tumor classification with rejection," *Artificial Intelligence in Medicine*, vol. 29, pp. 29-60, 2003.

[7] D. Aronsky and P. J. Haug, "Automatic identification of patients eligible for a pneumonia guideline," *Proc/AMIA AnnualSymposium*, pp. 12-16, 2000.

[8] E. Burnside, et al., "A Bayesian network for mammography," *Proc/AMIA Annual Symposium*, pp. 106-110, 2000.

[9] X.-H. Wang, et al., "Computer-assisted diagnosis of breast cancer using a data-driven Bayesian belief network," *International Journal of Medical Informatics*, vol. 54, pp. 115-126, 1999.

[10] B. Sierra, et al., "Using Bayesian networks in the construction of a bi-level multi-classifier. A case study using intensive care unit patients data," *Artificial Intelligence in Medicine*, vol. 22, pp. 233-248, 2001.

[11] L.Getoor, et al., "Understanding tuberculosis epidemiology using structured statistical models,"*Artificial Intelligence in Medicine*, vol. 30, pp. 233-256, 2004.

[12] H. K. Lee, et al., "Prevalence of diabetes and IGT in Yonchon County, South Korea," *Diabetes Care*, vol. 18, pp. 545-548, 1995.