대사증후군의 예측 모델링을 위한 베이지안 네트워크의 속성 순서 최적화

박한샘° 조성배
연세대학교 컴퓨터과학과
sammy@sclab.yonsei.ac.kr° sbcho@cs.yonsei.ac.kr

An Attribute Ordering Optimization in Bayesian Networks for Prognostic Modeling of the Metabolic Syndrome

Han-Saem Park° and Sung-Bae Cho
Dept. of Computer Science, Yonsei University

요 약

대사증후군은 당뇨병, 고혈압, 복부 비만, 고지혈증 등의 질병이 한 개인에게 동시에 발현하는 것을 말하며, 최근 경제 여건의 향상 및 식생활 습관의 변화와 함께 우리나라에서도 심각한 문제가 되고 있다. 한편 불확실성의 처리를 위해 많이 사용되는 베이지안 네트워크는 사람이 분석 가능한 확률 기반의 모델로 최근 의학 분야에서 질병의 진단이나 예측모델을 구성하기 위한 방법으로 유용하게 사용되고 있다. 베이지안 네트워크의 구조를 학습하는 대표적인 알고리즘인 K2 알고리즘은 속성이 입력되는 순서의 영향을 받으며, 따라서 이 또한 하나의 주제로써 연구되어 왔다. 본 논문에서는 유전자 알고리즘을 이용하여 베이지안 네트워크에 입력되는 속성 순서를 최적화하며 이 과정에서 의학지식을 적용해 효율적인 최적화가 가능하도록 하였다. 제안하는 모델을 통해 1993년의 데이터를 가지고 1995년의 상태를 예측하는 분류 실험을 수행한 결과 속성 순서 최적화 후에 이전보다 향상된 예측율을 보였으며 또한 다층 신경망, k-최근접 이웃 등을 이용한 다른 모델보다 더 높은 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, 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].

베이지안 네트워크(Bayesian network, BN)는 최근 복잡한 도메인에서 불확실성을 해결하기 위한 강력한 데이터 마이닝 방법으로 부각되고 있다[2]. 베이지안 네트워크는 도메인 지식을 적용하기 쉬우며 결과의 분석이 가능하다. 베이지안 네트워크는 또한 분류 문제를 속성 노드와 결과 노드간의 확률 관계로 가정하여 수리 통계학의 여러 가지 장점을 가지며, 이와 같은 장점을 바탕으로 의학 도메인에서 각종 질병의 진단이나 예측 문제를 해결하기 위해 많이 사용되어 왔고 또 좋은 성능을 보여왔다[2-4]. 분류나 예측문제를 해결하기 위해 신경망으로 대표되는 블랙박스 분류기 또한 많이 사용되어 왔다. 이와 비교해 베이지안 네트워크는 입력으로 연속 값이 아닌 범위가 정해진 상태 값을 사용함으로써 정확도 면에서 단점을 갖지만, 도메인 지식의 적용 가능성이나 원인 분석이 가능하다는 특성은 의학지식을 이용한 분석이 가능한 의학도메인에서 큰 장점이 된다.

본 논문은 대사증후군을 예측하는 문제를 다룬다. 이를 위해서 베이지안 네트워크를 이용하여 예측 모델을 구성하며, 네트워크 구조 학습을 위해서는 널리 알려진 Cooper와 Herskovitz의 K2 알고리즘을 사용한다[5]. K2 알고리즘은 입력 속성의 순서에 영향을 받으므로 베이지안 네트워크에 입력되는 속성의 순서를 최적화하는 것 또한 하나의 연구주제이다[6]. 본 논문은 효율적인 속성순서의 최적화를 위해 의학 지식을 적용하고, 유전자 알고리즘을 사용하였다. 이후 구조학습과 파라미터 학습과정을 거쳐 예측모델을 완성하고, 제안하는 모델의 유용성을 보이기 위해 예측 실험을 수행하였으며 그 결과 높은 예측율을 얻었다.

2. 배 경

2.1. 대사증후군의 정의

National Cholesterol Education Panel (NCEP)의 Adult Treatment Panel (ATP) III 보고서에서 제안한 바에 의하면 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. 의학 도메인에서 베이지안 네트워크의 사용

베이지안 네트워크는 데이터 마이닝 방법으로써 여러 가지 장점을 갖는다. 도메인 지식의 적용이 가능하며, 다른 모델보다 분석이 쉽다. 또한 데이터 수가 적더라도 도메인 지식을 데이터와 함께 사용함으로써 그 영향을 덜 받을 수 있다. 이러한 장점을 바탕으로 베이지안 네트워크는 의학분야에서 질병의 진단이나 예측을 위해 많이 적용되어왔다. Antal 등은 난소암의 진단모델을 구성하고 이를 분류하기 위해[3] 베이지안 네트워크를 사용하였다. 그 외에도 베이지안 네트워크는 폐렴이나 유방암, 결핵 등의 진단을 위한 모델을 구성하기 위해 사용되었다[4].

3. 유전자 알고리즘과 의학지식을 이용한 베이지안 네트워크의 속성순서 최적화

그림 1은 제안하는 방법의 흐름도이다. 진행과정은 크게 전처리, 속성 선택, BN 학습, 예측과정의 네 부분으로 나뉠 수 있다. 데이터 전처리와 속성 선택 과정에 의학지식을 적용해 보다 신뢰도 높은 모델이 구성되도록 하였으며, 속성 순서 최적화 과정에도 의학지식의 도움을 받은 후 유전자 알고리즘을 적용하여 효율적인 최적화가 가능하도록 하였다.

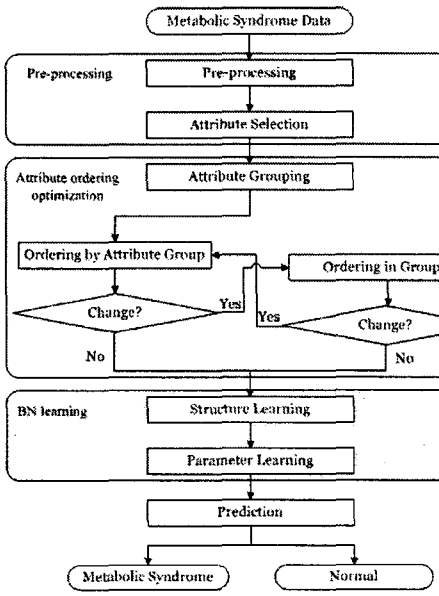

그림 1. 제안하는 방법의 흐름도

3.1. 의학지식을 이용한 데이터 전처리와 속성 선택

데이터가 일부 속성만을 제외하고 연속 값을 갖기 때문에 베이지안 네트워크의 입력으로 사용하기 위해 전처리 과정을 거쳤다. 본 논문에서는 이를 위해 대사증후군의 관련 연구를 참고하고[1], 의학 전문가의 도움을 받아 적절한 범위를 결정하였다.

몇 가지 중요 속성이 어떤 기준으로 나뉘어 졌는지를 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 속성의 상태를 나누기 위해 본 논문에서 사용된 기준은 대부분 실제로 임상에서 사용되고 있는 의학적으로 유의한 수치이다[1].

3.2. 속성 순서의 최적화

베이지안 네트워크의 구조 학습을 위해서는 K2 알고리즘을 사용하였는데, K2 알고리즘은 속성이 입력되는 순서에 영향을 받으므로 보다 정교한 구조를 얻기 위해 이 순서를 최적화하는 연구가 진행되어 왔다. Larranaga 등은 이 최적화 문제를 TSP (Traveling Salesman Problem)로 생각하고, 유전자 알고리즘의 여러 연산자를 적용하는 연구를 수행하였다[6]. 본 논문 또한 속성 순서를 최적화하기위해 유전자 알고리즘을 사용하였으며, 의학지식을 함께 적용하여 GA의 탐색공간을 줄여 보다 효율적인 수행이 가능하도록 하였다.

1) 속성 그룹 순서의 최적화

그림 1에서 보이듯, 이 부분은 속성 순서 최적화에 앞서 수행된다. 속성순서간의 인과관계는 사실 아주 복잡해서 간단히 단순화해서 표현할 수 없다. 하지만 일반적으로 비만관련 속성이 먼저 나타나고 대사이상 관련 속성에 영향을 미친다고 알려져 있다[1]. 본 논문은 레이블을 포함한 열 두 개의 속성을 위의 사실을 바탕으로 비만관련 속성(허리둘레, 체질량 지수, 허리 엉덩이 둘레비), 대사이상 관련 속성(공복혈당, 경구 당부하 검사 2시간째 혈당, 중성지방, HDL 콜레스테롤), 그리고 개별 속성을 그룹으로 하는 나머지 다섯 개의 속성 그룹으로 각각 묶었다.

이렇게 묶여진 7개의 속성 그룹으로 최적화를 수행한다. 그 후에는 각 그룹 내부의 속성 순서를 최적화하고, 속성 그룹 순서의 최적화와 그룹 내 속성순서의 최적화를 더 이상 변화가 없을 때까지 번갈아 수행한다.

2) 유전자 알고리즘의 적용

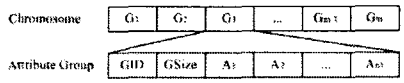

그림 2. 염색체 표현

그림 2는 유전자 표현을 나타낸다. 하나의 염색체는 m개의 속성 그룹을 표현하며, 염색체 내의 하나의 속성 그룹은 속성 그룹 ID, 속성 그룹 내 속성의 수, 그룹에 속한 속성들로 이루어진다. 이렇게 구성된 염색체는 개별 속성으로 이루어진 염색체와 동일한 유전연산이 적용될 수 있다. 초기 염색체의 순서는 임의로 결정되며 개체의 적합도는 예측율을 바탕으로 계산된다.

$$p_{(g,j)} = \frac{n - Rank(f(I_{(g,j)})) + 1}{n(n+1)\,/\,2} \qquad (1)$$

적합도 평가 후에는 다음세대의 개체가 선택된다. 식 (1)에서 $I_{(g,j)}$는 g번째 세대의 j번째 개체를 의미하며, $Rank\,(f(I_{(g,j)}))$는 적합도에 근거한 각 개체의 랭크를 의미한다. n을 개체수라고 하면, 이 식은 각 개체 $I_{(g,j)}$가 선택될 확률을 의미한다.

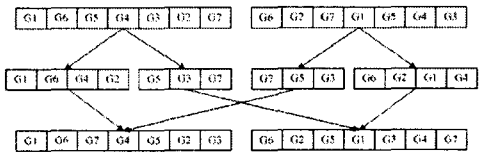

그림 3. 순환 교차 연산의 예

이후 교차와 돌연변이 연산을 거치게 된다. Larranaga 등은 여러 교차 연산자를 TSP문제에 적용해 비교해 본 결과 순환 교차 연산자(Cycle Crossover Operator)가 가장 좋은 결과를 보였다. 순환 교차 연산은 각 개체의 위치에 대응되는 상대 개체의 위치를 따라가며 순환을 이루는 부분을 분리하고 이들을 교환하는 방식으로 수행된다. 그림 3은 순환 교차 연산의 예를 보여준다.

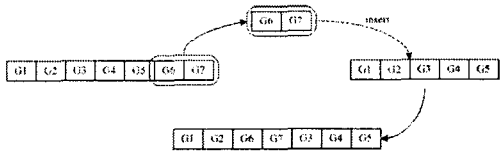

그림 4. 전치 돌연변이 연산의 예

돌연변이 연산은 전치 돌연변이 연산(Displacement mutation operator)이 사용되었는데 이는 앞서 언급한 Larranaga의 연구에서

순환 교차 연산자와 함께 사용되어 가장 좋은 성능을 보였다[6]. 전치 돌연변이 연산은 임의의 부분 문자열을 선택한 후, 이를 제거해서 다른 임의의 위치에 삽입하는 방식으로 동작한다. 그림 4는 전치 돌연변이 연산의 예이다.

3) 구조 및 파라미터 학습

앞에서 최적화된 속성 순서를 바탕으로, 베이지안 네트워크의 학습을 수행하였다. 대표적인 구조학습알고리즘인 K2 알고리즘이 네트워크 구조학습을 위해 사용되었으며[5], 구조가 결정된 후 학습 데이터의 빈도수를 바탕으로 파라미터를 계산한 후 예측 모델을 구성하였다.

4. 실험 및 결과

4.1. 실험 데이터 및 환경

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[7]. 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 불필요한 일부 속성을 제외한 18개의 속성을 사용하였다. 18개의 속성은 나이, 성별, 공복혈당, 경구 당부하 검사 2시간째 혈당, 신장, 체중, 허리둘레, 엉덩이 둘레, 고혈압여부, GOT, GPT, 콜레스테롤, 중성지방, HDL 콜레스테롤, 체질량 지수, 엉덩이-허리 둘레비, 수축기 혈압, 이완기 혈압이다.

제안하는 모델의 유용성을 보이기 위해 신경망(Neural Network, NN)과 k 최근접 이웃(k-nearest neighbors, kNN)의 두 모델과 비교 실험을 수행하였다. 신경망은 11-20-2의 입력-중간-출력 노드를 사용하였으며 kNN의 k는 예비 실험을 통해 높은 성능을 보인 3을 선택하였다. 유전자 알고리즘의 한 세대의 개체 수는 20이 설정되었고 100세대까지 진화시켰다. 0.8의 선택율, 0.02의 돌연변이율, 1.0의 교차율을 사용하였으나, 순환 교차연산의 경우 교차 후 동일한 결과가 나올 수도 있으므로 실질적으로는 1보다 작은 비율로 교차가 이루어진다. 진화 전의 실험에는 10-fold cross validation을 30차례 수행하여 평균값을 사용했으며, 진화시에는 데이터를 3:1:1의 비율로 나눠 학습, 검증, 테스트를 위해 사용하였다.

4.2. 실험 결과

첫 번째 실험은 속성을 달리한 베이지안 네트워크 모델간의 비교 실험으로 정의를 위해 필요한 기본 8속성을 이용한 모델, 의학지식 통해 선택한 11속성을 이용하는 제안하는 모델, 전체 18속성을 이용한 모델을 이용한 경우의 예측율을 비교하였다. 표 1을 보면 11속성을 갖는 제안하는 모델이 가장 높은 예측율을 보이는 것을 확인할 수 있으며 이 차이는 통계적으로 유의한 값이다. (p<0.001)

표 1. 선택속성에 따른 예측율 비교

속성 수	8	11	18
예측율(%)	70.74 (±0.0017)	**72.15** (±0.0082)	70.85 (±0.0079)

그림 5는 속성 순서가 최적화되어감에 따라 적합도가 올라가는 과정을 보여준다. 약 60세대를 지나며 평균적합도와 최고적합도가 모두 수렴하는 것을 확인할 수 있다.

마지막으로 다른 예측 모델과의 예측을 비교실험을 수행하였다. 표 2를 보면 신경망과 kNN보다 제안하는 방법의 의해 만들어진 BN 모델이 더 높은 예측율을 보임을 확인할 수 있으

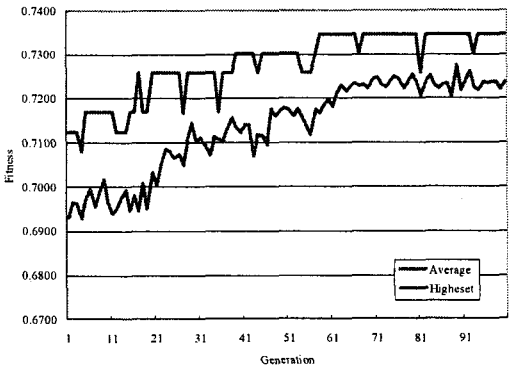

그림 5. 진화과정에서의 적합도 변화 추이

며, 역시 통계적으로 유의한 차이를 보인다. (p<0.001) 일반적으로 BN이 신경망보다 정확도 면에서는 떨어진다고 알려져 있으나, 본 논문에서는 의학지식을 이용한 전처리 및 속성 선택이 효과적으로 이루어진 결과 BN이 더 좋은 결과를 보였다고 해석할 수 있다.

표 2. 예측 모델에 따른 예측율 비교

속성 수	신경망	3-NN	**BN**
예측율(%)	63.19 (±0.0152)	62.56 (±0)	**72.12** (±0)

5. 결론

최근 심각한 문제가 되고 있는 대사증후군의 예측을 위해 본 논문은 베이지안 네트워크를 이용한 예측 모델을 제안하였다. 보다 정확한 예측을 위해 구조 학습 전에 속성 순서를 최적화하였으며, 더 효율적이고 안정적인 모델을 구성하기 위해 의학지식을 적용한 결과 다른 모델보다 정확한 예측을 수행할 수 있었다.

감사의 글

본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었음, IITA-2005-(C1090-0501-0019).

참고 문헌

[1] N. N. Mehta and M. P. Reilly, "Mechanisms of the metabolic syndrome," *Drug Discov Today*, vol. 1, no. 2, pp. 187-194, 2004.

[2] S.-M. Lee and P.A. Abbott, "Bayesian networks for knowledge discovery in large datasetes: Basics for nurse researchers," *J Biomed Inform*, vol. 36, pp. 289-299, 2003.

[3] P. Antal, *et al.*, "Using literature and data to learn Bayesian networks as clinical models of ovarian tumors," *Artif Intell Med*, vol. 39, pp. 257-281, 2004.

[4] 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," *Artif Intell Med*, vol. 22, pp. 233-248, 2001.

[5] C. F. Cooper and E. A. Herskovits, "A Bayesian method for induction of probabilistic networks from data," *Mach Learn*, vol. 9, no. 4, pp. 309-347, 1992.

[6] P. Larranaga, *et al.*, "Learning Bayesian nework structures by searching for the best ordering with genetic algorithms," *IEEE T Syst Man Cy A*, vol. 26, no. 4, 1996.

[7] Y. Park, *et al.*, "Prevalence of diabetes and IGT in Yonchon County, South Korea," *Diabetes Care*, vol. 18, pp. 545-548, 1995.