한국지능시스템학회 논문지 2009, Vol. 19, No. 6, pp. 796-801

유전자 알고리즘과 나이브 베이지언 기법을 이용한 의료 노모그램 생성 방법

A Clinical Nomogram Construction Method Using Genetic Algorithm and Naïve Bayesian Technique

이건명* · 김원재** · 윤석중**
Keon Myung Lee, Won Jae Kim, Seok Jung Yun

*충북대학교 전자정보대학 전자계산학과, PT-ERC
**충북대학교 의과대학
E-mail: kmlee@cbnu.ac.kr

요 약

복잡한 진단이나 예측 모델은 계산이 복잡하고 추론 과정을 해석하기 어렵기 때문에 임상현장에서 널리 사용되지 않고 있다. 의료 종사자들은 이러한 복잡한 모델 대신에, 복잡한 함수를 컴퓨터 등을 사용하지 않고도 쉽게 계산할 수 있도록 수치 관계를 그래픽으로 표현한 노모그램을 사용해 왔다. 의료분야에서 질병의 진단과 질병예후의 예측은 매우 주요한 관심사이다. 노모그램은 증상검사결과치료이력질병의 진단 결과 등의 속성을 포함한 임상 데이터들로부터 만들어진다. 노모그램을 만들 때는 가용한 여러 가지 속성 중에서 효과적인 것들을 찾아야 하고, 경우에 따라서는 속성에 대한 파라미터를 함께 결정해야 한다. 이 논문에서는 효과적인 속성과 파라미터를 선택하기 위해 유전자 알고리즘을 사용하고, 노모그램을 생성하기 위해 나이브 베이지언 기법을 사용하는 방법을 제안한다. 또한 제안한 방법을 실제 임상 데이터에 적용한 결과를 보인다.

Abstract

In medical practice, the diagnosis or prediction models requiring complicated computations are not widely recognized due to difficulty in interpreting the course of reasoning and the complexity of computations. Medical personnel have used the nomograms which are a graphical representation for numerical relationships that enables to easily compute a complicated function without help of computation machines. It has been widely paid attention in diagnosing diseases or predicting the progress of diseases. A nomogram is constructed from a set of clinical data which contain various attributes such as symptoms, lab experiment results, therapy history, progress of diseases or identification of diseases. It is of importance to select effective ones from available attributes, sometimes along with parameters accompanying the attributes. This paper introduces a nomogram construction method that uses a naïve Bayesian technique to construct a nomogram as well as a genetic algorithm to select effective attributes and parameters. The proposed method has been applied to the construction of a nomogram for a real clinical data set.

Key Words : 노모그램, 유전자 알고리즘, 나이브 베이지언 학습, 임상 데이터 분석, 기계학습

1. 서 론

질병의 양태와 치료의 효과가 개인별로 차이가 크게 나타날 수 있기 때문에 의료분야에서는 종종 진단 의사결정이 어려운 경우가 있다. 이러한 의사결정을 지원하기 위해 누적된 임상 사례들로부터 구축된 계산적 모델을 이용하려는 노력이 있어왔다.[1] 임상 의사결정 모델을 개발하는 것은 데이터들로부터 모델을 만드는 일이기 때문에 기계학습의 문제로 볼 수 있다.

접수일자 : 2009년 11월 5일
완료일자 : 2009년 12월 5일
본 논문은 2009년도 정부(교육과학기술부)의 재원으로 PT-ERC를 통해 한국연구재단의 지원을 받았습니다.

여러 의료 분야에서 진단 모델을 자동으로 구축하기 위해 기계학습 기법을 적용하는 시도들이 있었다.[1] 이렇게 개발된 모델들은 진단의 정확도 측면에서 어느 정도 향상된 성능을 보여 주었음에도 불구하고 몇 가지 이유 때문에 임상현장에서는 이러한 기술이 폭넓게 사용되고 있지는 않고 있다. 주관적인 의견이나 지식을 정형화하여 기호화된 형태로 기계학습 모델에 쉽게 반영하기 어려운 경우가 있다. 임상에서는 환자별로 선택적인 검사로 이루어지고, 환자에 대한 진단 요소와 추적기간 등의 차이가 있기 때문에, 임상 데이터에는 속성값이 없는 경우가 많이 있다. 대개의 기계학습 방법은 속성값이 비어있는 데이터가 많은 경우 학습결과에 민감한 영향을 받기 때문에, 기계학습 기법이 만족스러운 진단모델을 생성하지 못하는 경우도 있다. 한편, 임상의 진단예측에서는 판단결과에 대한 충분한 설명을 제공하는 것이 필요로 한다. 기계학습 모델은 학습 데이터에 대한 정확도에 초점을

맞추어 개발되기 때문에, 설명기능이 불충분해서 의료분야에서 기계학습 기반의 모델이 활용되지 못하는 측면도 있다.[2] 한편, 기계학습 모델이 지원 도구일 뿐 의사의 역할을 침해하는 것이 아님에도 불구하고, 의료 종사자들이 사용하기를 꺼리는 것도 기계학습 모델이 의료분야에 적극적으로 활용되지 못하는 이유이기도 하다.[1] 그럼에도 불구하고 증거 기반 의료(evidence-based medicine)에 대한 요구가 갈수록 커지고 있기 때문에, 의료 의사결정 지원을 위해 계산적 모델을 개발하려는 관심이 커지고 있다.

의료 의사결정 모델이 계산적으로 복잡하면, 임상현장에서 적용하기 어렵다. 임상현장에서 복잡한 수치 계산을 하지 않고 손쉽게 사용할 수 있는 의사결정모델이 노모그램(nomogram)이다. 노모그램은 함수 모델의 그래픽 표현 방법으로, 이를 이용하면 복잡한 계산을 하지 않고도 주어진 입력에 대한 함수값을 쉽게 구할 수 있다. 계산의 편의성 때문에, 노모그램은 진단 및 예후 예측 등을 위해 활용되고 있다. 임상 데이터들로부터 노모그램을 만들 때는 목표클래스를 효과적으로 예측할 수 있는 속성을 선택하여야 하고, 이들 속성을 이용하여 적합한 노모그램을 모델링해야 한다.

이 논문은 유전자 알고리즘과 나이브 베이지언(naïve Bayesian) 기법을 이용하여 노모그램을 작성하는 방법을 제안하다. 제안한 방법에서는 유전자알고리즘을 이용하여 적합한 예측 속성 및 이들의 파라미터 값을 탐색하도록 하고, 나이브 베이지언 기법을 이용하여 선택된 속성과 파라미터에 기반한 노모그램을 구성한다. 2절에서는 노모그램에 대한 소개와 노모그램에 관련된 기존 연구에 대해서 살펴 본다. 3절에서는 나이브 베이지언 기반의 노모그램 구축 방법에 대해 기술한다. 4절에서는 노모그램의 속성 및 파라미터의 선택을 위해 유전자 알고리즘을 이용하는 방법에 대해서 설명한다. 5절에서는 유전자알고리즘과 나이브 베이지언 기법을 이용하여 노모그램을 생성하는 전체 과정을 기술한다. 6절에서는 제안된 방법의 실제 적용 예를 살펴본 다음, 7절에서 결론을 맺는다.

2. 노모그램

노모그램은 복잡한 함수나 계산 모델을 가시화하기 위해서 사용되는데, 특히 의료분야에서는 질병의 진단, 전이, 재발, 생존예측 등을 위한 모델을 표현하는 데 활용되고 있다.[3-5] 노모그램은 (그림 1)과 같이 그래프로 표현되는데, 입력으로 사용되는 속성마다 하나의 선이 할당되고, 해당 속성이 가질 수 있는 값들이 선 위에 표시된다. 속성값의 위치에 대응하는 점수는 맨 위에 표시된 점수값(point score) 선의 위치값이 된다.

특정 모델에 대한 노모그램을 이용하여 임의의 환자에 대해 평가를 할 때는 다음과 같이 한다. 먼저 환자 데이터의 각 속성의 값에 대한 점수를 구하여, 이들 점수의 합을 계산한다. 점수합(score sum) 선에서 앞에서 구한 점수의 합 위치를 찾고, 이 위치에 대응하는 맨 아래 줄의 확률 선의 위치값을 읽으면, 이것이 노모그램에 의해 계산된 확률값이다. 예를 들어, (그림 1)의 노모그램에서 속성 A의 값이 a_2, 속성 B의 값이 b_3, 속성 C의 값이 c_1이라고 하면, 각각의 점수값은 0.4, 0.19, -0.2가 되고, 점수합은 0.39가 된다. 이 점수합에 해당하는 노모그램의 확률값은 약 67%가 된다. 의료 분야의 노모그램에서 속성은 임상에서 사용되는 환자에 대한 데이터가 되는데, 예를 들면 성별, 나이, 특정 증상의 유무, 검사 결과값 등이다. 노모그램을 통해 계산한 확률값은 해당 환자가 노모그램이 나타내는 대상 클래스에 속할 확률이다. 노모그램이 생존모델을 나타낸다면, 노모그램을 통해 얻어진 값은 해당 환자의 생존 예측확률이 된다.

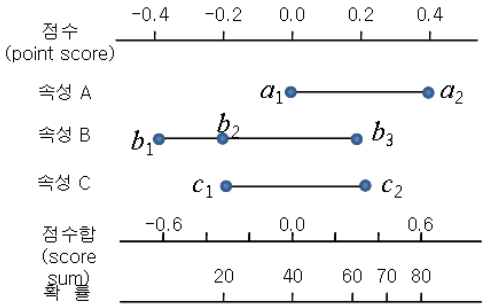

그림 1. 노모그램의 예

Fig. 1 An example of nomogram

노모그램을 구축하기 위해서 사용되는 대표적인 방법으로 Cox proportional hazard 모델을 이용하는 것과 나이브 베이지언 기법을 이용하는 것이 있다. Coxproportional hazard 모델은 통계학적 생존분석 모델의 하나로, 모델링 대상의 위험도은 요인함수의 곱셈으로 표현할 수 있다는 가정을 기반으로 한 것이다.[6] Cox proportional hazard 모델에서는 누적 생존함수 $S(t)$를 나타낼 때, 시간에만 영향을 받는 베이스라인(baseline) 생존 함수 $S_0(t)$와 속성값의 선형결합을 지수로 하는 지수 함수 $e^{(c_1a_1+c_2a_2+\cdots+c_ma_m)}$의 곱으로 다음과 같이 표현한다.

$$S(t) = S_0(t)e^{(c_1a_1+c_2a_2+\cdots+c_ma_m)} \quad (1)$$

Cox proportional hazard 모델의 $S(t)$를 특정 클래스에 속할 확률에 대응시켜서 계수값들 $c_1, c_2, \ldots, c_m$을 결정하고, 이들 값을 사용하여 확률을 얻는 방법이 노모그램 구성에 사용되기도 한다. Cox proportional hazard 모델을 사용하여 노모그램을 작성할 때 범주속성값은 이산수치값을 변환하고, 연속속성값은 이산수치값으로 바꾸어주는 전처리를 한 다음 계수결정을 위한 회귀분석 (regression)을 하게 된다.

Mozina 등[3,4]은 나이브 베이지언 기법을 이용하여 노모그램 구성할 수 있다는 것을 보이고, 이를 임상 분야에 적용할 수 있다는 것을 제시했다. Jakulin 등[5]은 SVM(Support Vector Machine)을 가시화하기 위한 노모그램 구성 방법을 제안했다. SVM이 분류기로서는 매우 효과적인 방법이기는 하지만, SVM을 가시화한 노모그램이 Cox proportional Hazard 모델이나 나이브 베이지언 기반 모델에 의해 만든 노모그램보다 복잡하기 때문에 임상 분야에서는 적용된 결과를 찾아보기 어렵다.

논문에서는 편의상 다음의 표기법을 사용하여 내용을 기술한다.

$A = \{a_1, a_2, \ldots, a_m\}$: 임상 데이터를 표현하는데 사용되는 속성의 집합

$C = \{0,1\}$: 클래스의 라벨로서, 0은 다른 클래스를, 1이면 목표 클래스를 나타냄

$D = \{d_1, d_2, \ldots, d_n\}$: 환자 데이터의 집합

$d_i = (v_{i1}, v_{i2}, ..., v_{im}, c_i)$: i번째 데이터.
v_{ij}는 속성 a_j의 값을 나타내고
c_i는 클래스를 나타냄.
N_j: j-번째 생성된 노모그램
$E_{N_j}(d_i)$: 데이터 d_i가 목표 클래스에 속하는 정도를 노모그램 N_j로 평가한 값
$(d_{(1)}, d_{(2)}, ..., d_{(n)})$: 데이터 집합 D를 $E_{N_j}(d_i)$가 증가하는 순서로 정렬한 서열
$D_P = \{d_k \mid c_k = c,\ d_k \in D\}$: 목표 클래스 c에 속하는 D의 부분집합
$D_N = \{d_k \mid c_k \neq c,\ d_k \in D\}$: 목표 클래스 c에 속하지 않는 D의 부분집합
$n_P = |D_P|$: D_P의 데이터 개수
$n_N = |D_N|$: D_N의 데이터 개수
$p_j^P(v) = \frac{\{d_i \mid a_i^j = v,\ d_i \in D_P\}}{n_P}$: D_P에서 j 번째 속성이 값 v를 가지는 데이터의 상대 빈도수
$p_j^N(v) = \frac{\{d_i \mid a_i^j = v,\ d_i \in D_N\}}{n_N}$: D_N에서 j 번째 속성이 값 v를 가지는 데이터의 상대 빈도수

3. 나이브 베이지언 기법을 이용한 노모그램 구축

나이브 베이지언 기법을 이용하여 노모그램을 구축하는 방법은 나이브 베이지언 분류기(naïve Bayesian classifier) 모델의 특성을 이용한다.[3,4] 나이브 베이지언 분류기 모델은 데이터의 속성값들이나 사건들은 서로 독립이라는 가정하에 베이즈 정리(Bayesian theorem)를 적용하여 특정 클래스에 대한 확률을 결정한다. 속성 독립의 가정을 사용하면, 어떤 개체 $X = (a_1, a_2, ..., a_m)$가 클래스 c에 소속될 확률인 사후 (posterior) 확률 $P(c|X)$은 다음과 같이 계산된다.

$$P(c \mid X) = \frac{P(a_1, a_2, ..., a_m \mid c)P(c)}{P(X)} = \frac{P(c)\prod_i P(a_i \mid c)}{P(X)} \quad (2)$$

노모그램은 어떤 개체가 특정 클래스에 얼마 만큼 적합한지 평가하기 위한 모델로 볼 수도 있다. c 를 노모그램이 대상으로 하는 목표 클래스라고 하고, $\bar{c}$를 c가 아닌 클래스라고 할 때, $P(\bar{c} \mid X)$는 객체 X 가 클래스 c에 속하지 않을 확률을 나타내다. 이 두 확률에 대한 승산비(odds ratio) *Odds*는 식(2)를 이용하여 다음과 같이 표현될 수 있다.

$$Odds = \frac{P(c \mid X)}{P(\bar{c} \mid X)} = \frac{P(c)\prod_i P(a_i \mid c)}{P(\bar{c})\prod_i P(a_i \mid \bar{c})} \quad (3)$$

$\log it$은 *Odds*의 로그(logarithm)를 취한 것으로 정의된다. $P(c|X)$에 대한 $\log it$는 다음과 같이 표현될 수 있다.

$$\log it P(c \mid X) = \log it P(c) + \sum_i \log \frac{P(a_i \mid c)}{P(a_i \mid \bar{c})}$$
$$= \log it P(c) + \sum_i \log OR(a_i) \quad (4)$$

위의 식은 $P(c|X)$의 $\log it$값이 각 속성값의 $\log OR(a_i)$들의 합으로 표현될 수 있다는 것을 보인다. 최종 확률값이 각 속성값의 평가값의 합으로 표현될 수 있기 때문에, 이 성질은 노모그램을 해석하는 방법과 유사하다. 따라서 나이브 베이지언 분류기의 위 성질을 이용하여 다음과 같은 과정을 통해서 노모그램을 생성할 수 있다.

1. 주어진 데이터 D에 대해서 목표 클래스 c와 상대 클래스 $\bar{c}$에서 각 속성 a_i의 모든 속성값 v_j의 상대 빈도 $p_i^P(v_j)$와 $p_i^N(v_j)$를 구한다.
2. 각 속성a_i의 각 속성값v_j에 대한 $\log OR$를 계산하다.

$$\log OR(v_j) = \log \frac{p_i^P(v_j)}{p_i^N(v_j)}$$

3. $\log OR(v_j)$를 대응하는 속성 a_i에 대한 점수로 간주하여, 데이터 d_i에 대한 평가 함수 $E(d_i)$를 $\sum_j \log OR(v_{ij})$로 정의하다.

$$E(d_i) = \sum_{a_j \in SAT} \log OR(v_{ij}), \quad (5)$$

여기에서 $\log OR(v_{ij}) = \log_{10} \frac{p_j^P(v_{ij})}{p_j^N(v_{ij})}$

4. 모든 가능한 속성값의 조합에 대해서 최대값 max와 최소값 min 을 찾는다.
5. 구간 [min,max]에 대해서 다음 식을 이용하여 데이터 d_i 가 목표 클래스 c 에 속할 확률값 $p(c|d_i)$을 결정한다.

$$P(c \mid X) = \left[1 + e^{-\log it P(c) - E(d_i)}\right]^{-1} \quad (6)$$

6. 각 속성에 대한 log*OR*과 구간 [min,max]에 대한 확률값을 이용하여 노모그램에 대한 그래프를 작성한다.

임상 데이터에 대해서 노모그램을 작성할 때, 위의 절차는 모든 속성이 데이터의 클래스에 의미있게 영향을 준다는 것을 가정하고 있다. 그렇지만, 모든 속성이 의미있는 영향을 주는 것은 아니다. 어떤 속성은 클래스와 관련성이 거의 없을 수 있으며, 또한 잡음으로 작용하여 클래스 식별을 더 어렵게 만들 수 있다. 따라서 노모그램을 구성할 때는 사용할 속성들을 선택하는 것이 매우 중요하다. 노모그램은 일반적으로 범주값을 다루기 때문에, 수치 속성은 범주 속성으로 변환하는 것이 필요하다. 따라서, 노모그램을 구성할 때는 수치 속성값에 대해 효과적인 범주 값들을 결정 하는 것도 필요하다.

4. 유전자 알고리즘 기반의 속성 및 파라미터 선택

노모그램에 사용할 속성을 선택하고, 수치 속성의 범주화(categorization)를 위해, 제안한 방법에서는 유전자 알고리즘에 기반한 방법을 사용한다. 유전자 알고리즘은 자연선택과 유전학의 메커니즘에 기초한 확률적 탐색방법 중의 하나이다. 유전자 알고리즘에서 염색체는 해결하려고 하는 문제에 대한 후보해를 코딩해 놓은 것이다. 생물학적 집단의 세대교체에서처럼 부모 염색체들에 대해 확률적으로 연산자를 적용하여 새로운 염색체 집단을 만들어진다. 진화론에서 환경에 적응하는 개체는 후손을 만들 가능성이 높고, 그렇지 않는 것은 소멸될 가능성이 높다고 하는 것처럼, 유전자 알고리즘은 적합도가 높은 염색체들은 세대교체가 계속되는 동안 생존할 가능성이 높도록 하고, 그렇지 못한 염색체는 도태될 가능성이 높아지도록 세대교체를 수행한다. 이와 같이 세대교체를 확률적으로 진행하도록 함으로써, 유전자 알고리즘은 세대교체에 따라 적합도가 높은 집단을 점진적으로 찾아가게 된다.

어떤 문제를 해결하기 위해 유전자 알고리즘을 사용하기 위해서는 다음과 같은 유전자 알고리즘의 구성요소를 개발하는 것이 필요하다. 우선 후보해를 염색체로 코딩하는 방법이 필요하고, 기존 염색체들로부터 새로운 염색체들을 만들기 위해 사용되는 유전연산자가 필요하다. 또한 염색체에 의해 표현되는 후보해가 문제에 대한 해로서 얼마나 좋은지 측정하는 적합도 평가 방법이 필요하다. 한편, 유전자 알고리즘은 집단의 세대교체를 이용하기 때문에, 초기 모집단을 생성하는 방법도 제공되어야 한다. 임상 노모그램을 만들 때, 사용될 속성의 선택과 파라미터의 결정을 위해 제안한 유전자 알고리즘은 다음과 같은 코딩방법, 유전 연산자, 적합도 평가함수, 그리고 초기 모집단 구성 방법을 사용한다.

4.1 모집단 표현방법

유전자 알고리즘을 통해서 구하고자 하는 해는 사용할 속성과 속성의 파라미터들이다. 따라서 염색체는 이들 정보를 표현해야 하는데, 제안된 방법에서는 비트열(bit string)을 사용한다. 속성별로 한 개의 비트를 할당하여 속성의 사용여부를 나타내고, 연속수치 속성인 경우에는 몇 개의 구간으로 이산화(discretization)하기 위해 사용할 경계값들을 비트열로 표현한다. 이들 경계값이 수치 속성의 파라미터에 해당한다. 하나의 연속수치를 2개 구간으로 구별하는 경우에는 하나의 경계값을 코딩하면 되고, 3개이상이라면 이에 대응하는 경계값들을 코딩한다. 경계값의 코딩은 속성의 범위를 고려하여 필요한 정밀도로 코딩한다.

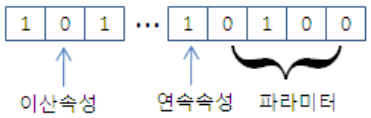

그림 2. 염색체 코딩
Fig. 2 Chromosome encoding

(그림 2)에서 보는 바와 같은 범주를 나타내는 이산속성의 경우에서는 한 비트가 할당되고, 연속속성인 경우에는 해당 속성의 사용여부를 나타내는 비트와 이에 연속해서 경계값들을 나타내는 비트열이 할당된다.

4.2 유전연산자

기존 염색체들로부터 새로운 염색체를 만들어내기 위한 유전연산자로는 비트별 교차(crossover)연산자[7]와 돌연변이(mutation) 연산자[7]를 사용한다. 비트별 연산자를 적용할 때는 해당 비트가 속성을 나타내는지 파라미터값을 나타내는지 구별하지 않고 연산자를 적용한다.

4.3 적합도 함수

염색체가 주어지면, 염색체가 표현하고 있는 선택된 속성들과 연속수치속성의 이산화를 위한 경계값을 사용하여 3절에서 설명한 나이브 베이지언 기법을 통해서 노모그램을 작성한다. 즉, 이 과정을 통해서 염색체가 나타내는 후보해인 후보 노모그램이 생성되는데, 유전자 알고리즘의 진화메커니즘을 통해서 최선의 노모그램을 선택하기 위해서는, 염색체(즉, 노모그램)가 해로서 얼마나 적합한지 평가하는 평가함수가 필요하다. 제안한 방법에서는 유전자 알고리즘의 적용과정에서 만들어지는 노모그램의 적합성을 평가하기 위해서 다음과 같은 절차를 사용한다.

1. 주어진 노모그램 N_j의 각 데이터 d_i에 대한 평가값 $E_{N_j}(d_i)$을 계산한다.
2. 전체 데이터 D를 평가함수값 $E_{N_j}(d_i)$이 증가하는 순으로 정렬하고, 정렬된 데이터를 $(d_{(1)}, d_{(2)}, ..., d_{(n)})$와 같이 표현한다.
3. 노모그램 N_j의 적합도 함수 $fitness(N_j)$는 다음과 같이 정의한다.

$$fitness(N_j) = \Delta h \times \left(1 - \frac{DescH}{\Delta h}\right) \tag{7}$$

$p_k = \dfrac{\left|\left\{d_{(j)} \middle| c_{(j)} = c, j = k...n\right\}\right|}{n-k+1}$: 정렬된 데이터 집합 $(d_{(k)}, d_{(k+1)}, ..., d_{(n)})$에서 k 이후의 데이터 중에서 목표 클래스에 속하는 데이터의 비율.

$\Delta h = \max_k p_k - \min_k p_k$: p_k 분포에서 첫 번째 것과 마지막 것의 차이.

$DescH = \sum_{i=1,(p_{i+1}-p_i)<0}^{n-1} (p_{i+1} - p_i)$: $p_i\ (i = 1,...,n)$의 분포에서 인접하는 값에서는 작아지는 값들의 누적 합.

$SAT = \{a_{(1)}, a_{(2)}, ..., a_{(K)}\},\ where\ a_{(k)} \in A$: 현재 염색체에서 사용하고 있는 속성의 집합.

염색체의 평가함수는 $E_{N_j}(d_i)$의 값에 따라 정렬된 데이터 집합 $(d_{(1)}, d_{(2)}, ..., d_{(n)})$에 대한 p_k의 분포에 대해서, 단조증가 (monotonously increasing)하는 성질이 클수록, 또한 분포 p_k의 최대값과 최소값의 차이가 클수록 큰 값을 내는 특성을 갖는다. 이러한 특성은 노모그램에 의해서 평가된 확률값이 클수록 목표 클래스에 속할 확률이 크다는 의미이고,

또한 클래스에 속하는 것과 그렇지 않은 것에 대해서 평가확률값의 차이를 크게 한다는 것을 반영한다. 이러한 성질은 노모그램에서 요구되는 특성이기 때문에, 염색체 평가함수는 노모그램의 성능을 평가할 수 있는 척도가 된다.

4.4 모집단 초기화 방법

유전자 알고리즘의 초기 모집단은 후보해를 나타내는 염색체들로 구성되어야 한다. 제안된 방법에서는 염색체가 비트열로 코딩이 되기 때문에, 초기 염색체들은 속성의 개수, 수치속성별 경계값의 개수 및 경계값 표현에 사용할 비트수를 고려하여 결정된 크기의 무작위 비트열로 초기화된다.

5. 노모그램 생성 방법

다음은 주어진 임상 데이터에 대해서 노모그램을 생성할 때 사용되는 절차를 보인 것이다. 입력으로는 속성값들과 클래스의 정보로 구성된 임상 데이터의 집합과, 연속 속성의 경우 이산화할 때 만들 범주의 개수와 범위정보를 제공해야 한다.

procedure nomogram-construction

1. 유전자 알고리즘의 모집단을 4.4절에서 설명한 방법에 따라 초기화 한다.
2. 주어진 데이터 집합 D 에서 목표 클래스 c 와 다른 클래스 $\overline{c}$ 에 대해 각 속성 a_i별 가능한 값들 v_j에 대한 상대빈도 $p_i^P(v_j)$와 $p_i^N(v_j)$를 계산한다.
3. 각 속성 a_i 별 모든 값들 v_j에 대한 logOR값을 구한다.

$$\log OR(v_j) = \log \frac{p_i^P(v_j)}{p_i^N(v_j)}$$

4. 각 염색체에 대해서 3절에서 설명한 방법에 따라 노모그램을 구성한다.
5. 각 구성된 노모그램 N_j의 적합도를 적합도 함수 $fitness(N_j)$를 이용하여 구하고, 대응하는 염색체의 적합도로 설정한다.
6. 최대 세대교체수 또는 최적 평가값의 변화량 등에 대한 종료조건이 만족되면 단계 8로 간다.
7. 유전연산자를 적용하여 새로운 세대의 염색체 집단을 생성하고, 단계 4로 간다.
8. 모집단으로부터 적합도 값이 가장 큰 노모그램 N을 최종해로 선택한다.

$$N = \arg\max_{N_j} fitness(N_j)$$

6. 적용 사례

제안된 방법의 적용 가능성을 확인하기 위해서, 제안된 방법을 방광암 수술환자의 암재발 확률을 예측하는 노모그램 개발에 적용하였다. 실험에 사용된 대상 데이터는 166명의 방광암 환자에 대한 성별(sex), 나이(age), 방광절제수술 또는 방광수술여부 (operation), 종양의 크기(tumor size), 암종존재 여부 (CIS, carcinoma in situ), 종양의 개수(tumor number), 종양의 종류(tumor type), 암의 병기(grade), 방광치료 여부 (intravesical therapy) 등 10가지 속성을 포함하고 있다. 또한 추적기간 중에 방광암의 재발 여부를 나타내는 클래스 정보를 포함하고 있다.

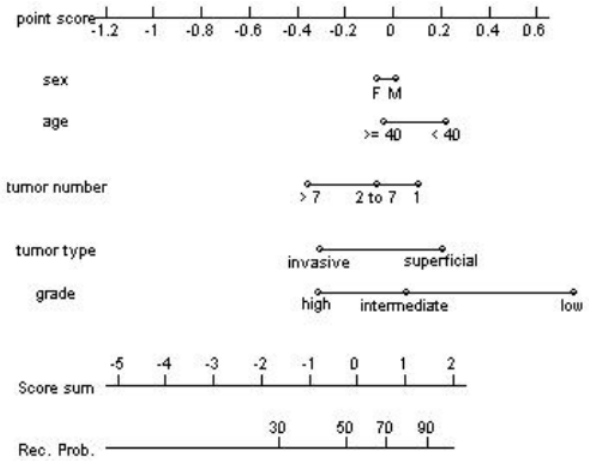

그림 3. 방광암의 재발에 대한 노모그램
Fig.3 A nomogram for bladder cancer recurrence

염색체는 10개 속성 각각이 노모그램의 입력으로 사용될지 여부를 나타내기 위해 10비트를 비트열로 코딩하였다. 초기 모집단은 무작위로 염색체의 비트값을 설정하도록 하는 방법으로 구성하였다. 염색체에 대한 평가는, 염색체에서 비트값이 1인 속성을 입력속성으로 선택하고, 주어진 임상데이터에 대해서 3절에 기술한 방법으로 노모그램을 구축하고, (식 7)을 사용하여 평가값을 결정하였다. (그림 3)은 166명의 임상데이터에 대해서 실험을 통해 얻어진 평가값이 가장 양호한 노모그램을 보인 것이다. 전체 10개의 속성 중에서 (그림 3)에 나타난 5개 속성을 사용하는 노모그램이 가장 양호한 결과를 도출한다는 것을 보여준다. (그림 3)과 같은 노모그램을 방광암 재발 예측에 활용할 때는 다음과 같이 재발확률을 계산한다. 예를 들면, 40세 미만(age < 40)의 남성(sex = male)이, 종양의 개수가 1개(tumor number = 1)였고, 표재성 방광암(tumor type = superficial)이었고, 암의 병기가 낮았다(grade = low)면, 노모그램 구축에 사용되었던 데이터의 관점에서 볼 때 이 환자에 대한 속성별 point score값은 각각 0.04, 0.21, 0.1, 0.22, 0.65이다. 따라서 score sum은 0.04+0.21+0.1+0.22+0.65 = 1.22이고, 이에 대한 암이 재발할 확률(Rec. Prob.)은 87%라고 판정한다.

(그림 4)는 정렬된 데이터 서열 $(d_{(1)}, d_{(2)}, \ldots, d_{(166)})$에 대해서 p_k에 의해 만들어지는 분포를 보인 것으로, 노모그램의 품질을 평가하는데 사용된다. 노모그램의 품질은 p_k분포에 평가함수 $fitness(N_j)$를 적용하여 결정된다. (그림 4)와 같은 p_k 분포는 다음과 같은 과정으로 구성되었다. (그림 3)의 노모그램을 실험에 사용된 166명의 환자데이터에 대해서 적용하여 재발 확률값을 계산한 다음, 이 확률값이 증가하는 순서로 데이터를 정렬하여 데이터 서열 $(d_{(1)}, d_{(2)}, \ldots, d_{(166)})$을 만든다. p_k분포의 x축은 이 데이터 서열에 대응된다. 즉, $d_{(k)}$는 x축에서 k번째 위치에 해당한다. $d_{(k)}$에 대응하는 y축의 값은 p_k 값을 나타내는데, p_k 는

$d_{(k)}$부터 $d_{(n)}$까지의 데이터 중에서 실제로 암이 재발한 환자의 확률을 나타낸다. 이상적인 노모그램은 노모그램에 의한 확률값이 클수록 해당 환자에 대한 재발확률이 커지는 것이다. 이러한 성질은 p_k의 분포가 단조증가할 때 만족된다. 그렇지만 실제의 임상데이터에서는 개인별 다양성이 크고, 가용한 임상데이터의 크기가 제한적이기때문에, 절대적인 단조증가의 특성을 기대하기 어렵다. (그림 4)의 분포에서도단조증가하지 않는 부분들이 여러 군데 나타난다. 그렇지만, 베이지언 기법 노모그램 구축 방법을 이용하여 유전자알고리즘을 통해서 노모그램의입력으로 사용할 속성을 선택하는 실험에서, 최선의 것으로 찾아진 노모그램이 (그림 4)의 것이다. (식 7)의 평가함수는 (그림 4)와 같은 p_k의 분포에 대해서 단조증가하는 특성이 클수록, 또한 가능하면 분포에서의 최대값과 최소값의 차이를 클수록 큰 값을 출력한다.

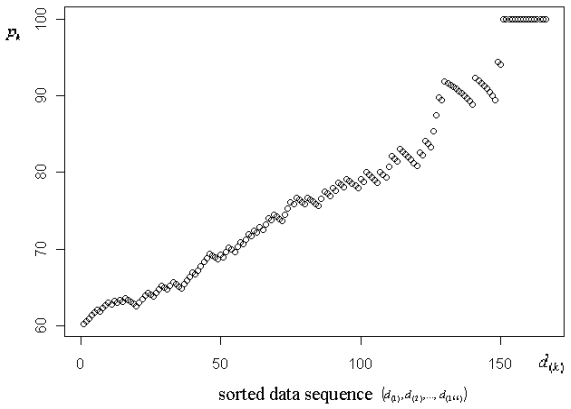

그림 4. 노모그램 품질평가를 위한 p_k의 분포

Fig.4 A distribution of p_k used to evaluate the quality of a nomogram performance

7. 결 론

의료분야에서는 보다 나은 의료 서비스를 제공하기 위해 증거기반의 의료를 제공하는 데 많은 관심을 기울이고 있다. 노모그램은 누적된 실제 임상 데이터를 바탕으로 한 증거기반 의료 도구의 하나이다. 노모그램은 임상 데이터의 누적에 따라 지속적으로 개선되도록 하는 것이 바람직하다. 제안된 방법은 유전자알고리즘과 나이브 베이지언 기법을 이용하여 임상 데이터로부터 노모그램을 생성할 수 있도록 해준다. 제안된 방법은 실제 임상 데이터에 대해 적용되어 노모그램을 작성하였다.

노모그램은 나이브 베이지언 기법 뿐만 아니라, Cox proportional hazard 모델을 이용하여서도 구성할 수 있다. Cox proportional hazard 모델인 경우에는 수치속성을 미리 어떻게 범주화할 것인지 정보를 설계자가 제공하여야 한다. 반면 제안된 방법은 최선의 경계값들을 유전자 알고리즘이 찾도록 하기 때문에, 설계자가 제공할 필요가 없다. 이런 관점에서 향후 유전자알고리즘과 Cox proportional hazard 모델을 결합한 노모그램 구축 방법을 개발할 수 있을 것으로 기대한다.

참 고 문 헌

[1] I. Kononenko, Inductive and Bayesian Learning in Medical Diagnosis, *Applied Artificial Intelligence*, vol.7, pp.317-337, 1993.

[2] V. Pirnat, I. Kononenko, T. Janc, I. Bratko, Medical Estimation of Automatically Induced Decision Rules, *Proc. of 2nd Europ. Conf. on Artifical Intelligence in Medicine*, pp.24-36, 1989.

[3] M. Mozina, J. Demsar, M. Kattan, B. Zupan, Nomograms for Visualization of Naïve Bayesian Classifier, *Proc. of PKDD 2004*, *LNAI 3202*, pp.337-348, 2004.

[4] M. Mozina, J. Demsar, M. Kattan, B. Zupan, Nomograms for Naïve Bayesian Classifiers and How can They Help in Medical Data Analysis, *Proc. of MEDINFO 2004*, pp.1762, 2004.

[5] A. Jakulin, M. Mozina, J. Demsar, M. Kattan, B. Zupan, Nomograms for Visualizing Support Vector Machines, *Proc. of SIGKDD'05*, 2005.

[6] 송경일, 안재억, *SPSS for Windows를 이용한 생존분석*, SPSS 아카데미, 1999.

[7] S. Olariu, A. Y. Zomaya (eds.), *Handbook of Bioinspired Algorithms and Application*, Chapman & Hall/CRC, 2006.

[8] 안재억, 유근영, *의학보건학 통계분석, 한나래, 2006.*

저 자 소 개

이건명 (Keon Myung Lee)
15권 3호 참조

김원재 (Won Jae Kim)
충북대학교 의과대학 교수
관심분야 : 비뇨기학, 종양학

윤석중 (Seok Jung Yun)
충북대학교 의과대학 조교수
관심분야 : 비뇨기학, 종양학