보건정보통계학회지 제40권 제2호
ISSN 2287-3708(Print) ISSN 2287-3716(Online)
Journal of Health Informatics and Statistics (JHIS)
2015; 40(2): 13-21

SPAN을 이용한 간경변증 발생 위험군 분류 평가

유영애, 송기준†

연세대학교 의과대학 의학통계학과

The Classification of Risk Group for Liver Cirrhosis Using SPAN

Young Ae Yu, Kijun Song†

Department of Biostatistics, Yonsei University College of Medicine

Abstract

Objectives: The statistical predictive methods have been used to find the risk factors related with diseases and to generate predictive probabilities of those diseases. Logistic regression is the most commonly used method for predicting the probability of diseases in the medical fields. Also, data-driven methods, such as CART have been used to identify subjects at increased risk of diseases. However, both of regression and tree models have their specific limitations in spite of their advantages. Recently, an alternative approach called by search partition analysis (SPAN) is suggested, which is based on direct non-hierarchical search algorithm to identify subgroups at risk. SPAN searches subgroups among different Boolean combinations of risk factors.

Methods: SPAN was compared against the performance of the other 3 methods; logistic regression, polychotomous regression and quick unbiased efficient statistical trees. We applied these methods to the real clinical data composed of 4,093 individuals who received the screening test in first and then visited Yonsei University Medical Center for check-up liver cirrhosis between May 1994 and September 2005. The performance of SPAN and that of any other methods were compared and the measures of performance were sensitivity, specificity, and accuracy.

Results: In the results using SPAN, the findings identified by the risk factors for liver cirrhosis were HbsAg, AntiHCV, Family history, platelet and α-FP. And we found that the sensitivity using SPAN were much higher than those of other methods in various data sets.

Conclusions: In conclusion, as long as it works, the performance of SPAN should make sense in the context of medical diagnosis and prognosis. Also, It was known that SPAN had an advantage that its decision rules are usually more interpretable than those of other methods.

Keywords: SPAN, Polychotomous regression, QUEST, Classification, Liver cirrhosis

[Submitted: 2015년 02월 03일, Revised: 2015년 04월 30일, Accepted: 2015년 06월 03일]

† Corresponding Author: Kijun Song, PhD
Department of Biostatistics, Yonsei University College of Medicine, 50 Yonsei-ro, Seodaemun-gu, Seoul 120-749, Korea. Tel: +82-2-2228-2491
E-mail: biostat@yuhs.ac

1. 서 론

임상의학분야에서 질병 발생의 예측이나 위험요인을 분석하기 위해 통계학적 방법을 이용해 왔다. 대표적으로 회귀분석방법 중에서 모수적 방법인 로지스틱 회귀분석이 널리 사용되어 왔다 [1]. 최근에는 회귀분석방법 중에서 비모수적 방법인 multivariate adaptive regression splines (MARS)와 이를 응용한 다항수준 회귀분석(polychotomous regression)도 사용되고 있다. 이런 회귀분석방법의 장점은 질병발생에 대한 각 변수들의 중요도를 판단할 수 있고 아울러 예측이 용이하다는 것이다 [2]. 회귀분석방법들과 함께 데이터마이닝의 한 분야로 나무모형을 이용한 분류(classification) 분석방법 또한 빈번하게 이용되고 있는데, 나무모형을 이용한 분석 방법은 자료의 분포에 대한 특별한 가정이 필요 없고, 결과 해석이 쉽다는 장점을 가지고 있다 [3]. 이런 나무모형 중에 가장 대표적인 방법으로 classification and regression tree (CART)가 있다. CART는 반복적인 탐색을 통해 최상의 분리점을 찾는 장점이 있다. 하지만 독립변수의 수와 그 범주가 많아지면 계산의 양이 많아져 시간이 오래 걸린다는 단점이 있다 [4]. 그래서 이를 보완하여 변수 선택은 통계학적 유의성 검정을 사용하고, 분리점 선택은 탐색적 방법을 이용한 quick unbiased efficient statistical trees (QUEST)도 사용되고 있다. 하지만 CART나 QUEST 같은 경우에는 위계적으로(hierarchically) 자료를 탐색하기 때문에 예상치 못한 결과가 종종 발생되어 해석을 난해하게 한다는 단점이 있다 [5]. 이런 단점을 보완하고자 비위계적(non-hierarchical) 분류 분석방법인 search partition analysis (SPAN)가 최근에 제안되었다. SPAN은 모든 가능한 변수의 조합 중에서 최상의 조합을 찾아내는 것으로, 종속변수와의 관계가 명백한 변수들의 조합만을 찾기 때문에 임상의학적으로 의미 있는 결론을 얻을 수 있으며, 이에 대한 해석과 적용이 쉽다는 장점을 가지고 있다 [6,7]. 본 연구에서는 SPAN의 질병 발생 위험군 분류에 대한 유용성을 로지스틱 회귀분석, 다항수준 회귀분석, QUEST의 분석결과와 비교하여 평가하였다. 이 연구를 위해 1994년부터 2005년까지 연세의료원 건강검진센터에서 건강검진을 받은 검진자 중 병원에 내원하여 간경변증 발생 여부에 대한 진단을 받은 4,093명의 검진자료를 이용하는데, 자료를 둘로 나누어 하나는 훈련용 자료(training data)로 모형 설정을 위해 쓰고, 나머지 하나는 검증용 자료(test data)로 만들어진 모형의 검증을 위해 사용하였다.

2. 연구 방법

1) Search Partition Analysis (SPAN)

SPAN은 자료를 두 개의 집단으로 분류하는 알고리즘의 일종이다. 만약 두 개의 집단을 S와 S'이라고 하면 두 개로 나누어진 분류의 결과는 각 속성들의 부울 결합(boolean combinations)을 이용하여 표현된다 [8,9]. 여기서 말하는 속성이란 독립변수로부터 얻은 특징으로, 예를 들면 '나이가 40세 이상이다', '간경변증에 대한 가족력이 있다' 등의 것으로 질병에 대한 위험인자로 생각할 수 있는 것들이다. 속성들의 부울 결합은 자료를 두 개의 공간으로 나누게 되는데, 이때의 목표는 S와 S' 각각을 가장 동질적으로 만드는 분류 결과 A와 A'를 찾는 것이다. 속성은 이분형 변수의 형태로만 가능하기 때문에 독립변수가 이분형 변수일 경우에는 문제가 되지 않지만, 다항형이나 연속형 변수일 때에는 연구자가 기존에 알려져 있는 정보를 이용하거나 탐색적 자료 분석을 통해 이변형의 속성으로 정의해야 한다. 속성을 정의한 후에는 S를 잘 반영할 수 있는 m개의 속성을 선택하여 다음과 같이 속성들의 집합을 정의한다.

$$T_m = \{X_1, X_2, \cdots X_m\}$$