Filter Method를 이용한 간경변 발생 의심군 예측 요인의 선택

연세대학교 대학원
의학전산통계학협동과정
의학통계학전공
김 정 숙

Filter Method를 이용한 간경변 발생 의심군 예측 요인의 선택

지도 김 동 기 교수

이 논문을 석사 학위논문으로 제출함

2006년 2월 일

연세대학교 대학원

의학전산통계학과

김 정 숙

김정숙의 석사 학위논문을 인준함

심사위원 _____ 인

심사위원 _____ 인

심사위원 _____ 인

연세대학교 대학원

2005년 12월 일

감사의 글

설레는 마음에 동기들을 만난 날이 바로 엊그제 같은데, 벌써 졸업을 앞두고 논문을 쓰고 석사 학위를 마무리 하려고 합니다. 호기심에 의학통계를 배우려고 했고, 학업에 많은 준비가 되지 않은 상황에도 의학통계의 기본을 잘 이끌어 주신 김동기 교수님께 제일 먼저 깊은 감사의 마음을 전합니다. 많이 부족한 저를 든든히 이끌어 주신 김동건 교수님께 감사의 마음을 전합니다. 유전 통계학의 새로운 내용을 알게 해주신 임길섭 교수님께도 감사의 마음을 전합니다. 학부 과정에서 다하지 못한 통계학의 깊이를 알게 해주신 김현중 교수님, 이학배 교수님께 감사의 마음을 전합니다.

뒤에서 묵묵히 든든한 힘이 되어준 송기준 박사님과 무영오빠에게 감사의 마음을 전합니다. 함께 생활하진 못했지만 스스럼없이 대해주시고 충고 아끼지 않으셨던 성민오빠, 미영언니, 혜리언니, 원열 오빠에게 감사의 마음을 전합니다. 입학하고 바로 학기를 시작하지 못해서 동기임에도 같이 하지 못했던 신영언니와 지금은 사회에 나가있는 수옥언니와 은혜에게 아쉬운 마음과 고마운 마음을 전하며, 앞으로 행복한 일들만 있길 바랍니다. 함께 수업 듣고, 같이 숙제하고 즐거운 시간을 보냈던 예쁜 동생들 소연, 민진, 성은, 혜진과 은희씨에게 고마운 마음을 전합니다. 친언니처럼 때론 친구처럼 대해주시고, 힘들게 공부하던 저에게 많은 도움이 되어주신 명희언니, 은정언니, 현선언니에게 감사의 마음을 전합니다.

학부에서도 도움을 많이 주셨지만, 지금까지도 관심 가져주시고 많은 도움을 주셨던 류제복 교수님, 이승주 교수님, 박효일 교수님께 감사의 마음을 전합니다. 같이 의학통계에 관심을 갖고 있었지만 같은 학교를 다니지 못해 못내 아쉬웠던 후배 금나에게도 감사의 마음을 전하며, 훗날 좋은 논문 쓰길 바랍니다. 대학원 진학에 용기를 주고 응원해 주었던 필순, 자은, 지은, 선희에게 고마운 마음을 전합니다. 학업 정진의 어려움과 고통을 함께 나누었던 태성오빠, 형기오빠 그리고 항상 응원의 대장이 되어서 득이 되는 말만 해주셨던 민정언니, 높은 선배의 자리

에서 후배들을 잘 이끌어 주시려고 노력하시고, 후배 사랑이 돈독하신 서형준 선배님께 감사의 마음을 전합니다. 도움을 주지 못해 아쉽지만 언제나 열심히 공부하는 모습이 보기 좋은 지송 선배와 현주에게 응원을 보냅니다. 비록 알게 된지는 얼마 되지 않았지만 짧은 시간 동안에 친구처럼 가까워졌던 지원, 순형을 포함해 그 외 방송대 식구들에게도 감사의 마음을 전합니다.

자식 둘을 모두 내보내 못내 아쉬워하시던 아빠, 타지에서 고생한다고 항상 마음 쓰신 엄마, 철없지만 언니보다 더 강하고 착실한 동생 정은이에게 고마운 마음을 전하며, 더욱 훌륭한 모습 보여드리도록 노력하겠습니다. 같이 있을 땐 몰랐지만 떨어져 있어서 더욱 소중한 우리 가족들 모두 사랑합니다.

졸업을 앞두니 대학원 생활에 많은 아쉬움이 남습니다. 본분인 학생으로서 학업에 정진을 다하지 못한 것과 많은 경험을 하지 못한 것에 아쉬움이 남습니다. 언제나 이런 후회는 꼭 마지막에 하는 것 같습니다. 도움주신 많은 분들께 감사하며, 그 보답으로 훗날 더욱 더 발전된 모습을 보여드리도록 노력할 것이며, 내내 평탄하고 행복한 날만 계속 되길 바라겠습니다.

2005년 12월

김 정 숙 올림

차 례

그 림 차 례

표 차 례

국 문 요 약

Filter Method를 이용한 간경변 발생 의심군 예측 요인의 선택

2004년 40대 사망원인으로 간 질환이 2위를 차지하면서 통계청은 “남성의 40대 사망률이 여성의 3배 이상인 것에는 여러 가지 원인이 있으나 간 질환이 대표적”이라고 한 바 있다. 실제로 남성이 여성보다 간 질환으로 사망할 확률이 4.4배에 달했다. 건강관리와 조기 건강 진단을 통해 간 질환을 예방할 수 있도록 본 논문에서는 특히 간경변 의심군의 위험인자(Risk Factor)들이 무엇인지 예측해 보고자 한다. 1994년부터 2005년까지 건강 검진 자료에서 총 64,211명을 대상으로 신체계측, 간 기능 검사, 대사 및 전해질, 혈액 검사, 뇨 검사, 간염 검사 등을 결과를 이용한다.

간경변 의심군의 위험인자 선택에 있어서는 Data mining 기법인 LDA, QDA, 로지스틱 회귀모형, CART와 같은 예측 모형을 사용했고, 또 다른 방법으로 위험인자의 선택에 앞서서 Filter Method인 FINCO를 사용해 차수를 줄인 후 걸러진 변수로 다시 데이터 마이닝 기법을 적용해 보았다. Filter Method를 사용하기 전에 예측 모형으로 선택된 위험인자에는 알부민, platelet(혈소판), B형간염S항원, 총단백, 총콜레스테롤, LDH(젖산탈수소효소), γ-GT(감마 글루타미리 트랜스)가 있었으며, Filtering을 통해 예측 모형으로 선택된 위험인자에는 Alk.Phos(알칼리 포스파타제), B형간염S항원, MCH(적혈구 1개당 혈색소 양), 몸무게, 나이가 있었다.

핵심 되는 말: 위험인자, LDA, QDA, 로지스틱 회귀모형, CART, Filter Method, FINCO

제 1장 서 론

2004년 통계청에서 발표한 사망 원인 통계 결과를 보면 암, 뇌혈관질환, 심장질환, 당뇨병, 자살에 이어 간 질환이 총 사망자 245,771명 중 9,272명(3.77%) 사망으로 6위를 차지했다. 간 질환의 종류로는 지방간, 간경변, 간암, 간염 등으로 주된 원인은 장기간의 다량 음주로 인한 영양장애, 약물 남용, 독소 중독, 바이러스, 비만, 노화 등 다양하다. 간 질환에 대한 용어를 살펴보면, 간세포가 바이러스, 알코올, 약물이나 독물, 자가 면역, 대사 장애 등에 의해 손상을 입고 망가져서 염증이 발생된 상태를 간염이라고 하고, 간염이 6개월 이내에 회복되는 경우를 급성 간염이라고 한다. 간염 바이러스가 몸 안에 침입하면 간에 지하당을 만들고 번식한 후 면역 세포와 전쟁을 벌이다 간세포가 파괴되고 간 기능이 손상되는데 이러한 간염이 6개월 이상 낫지 않고 진행되면 만성 간염이 된다. 계속해서 염증세포가 침윤되어 광범위한 섬유화에 의해 두꺼운 섬유질이 형성되고, 살아남은 간세포들에 의해 재생결절이 형성되면서 간의 정상적인 구조가 소실되고 일그러지고 굳어지면서 자갈밭처럼 울퉁불퉁하게 변하는데 이를 간경변(Liver Cirrhosis)이라고 하고 형태학적으로는 원래의 정상 간으로 돌아갈 수 없게 된다. 간 질환은 2003년 연령별 사망원인 분포에서 30대에는 3.89%였지만 40대에는 8.54%로 현저히 높아짐을 알 수 있다. 특히, 본 논문은 건강검진 자료를 통해 간경변 의심군의 위험인자에 대해 연령 외의 또 다른 위험인자들을 예측해 보도록 한다. 예측된 위험인자를 통해 보다 더 빠른 시간에 간경변임을 밝혀낼 수 있길 바란다.

본 논문은 1994년부터 2005년까지 건강 검진(Screening test) 자료를 이용해 간경변으로 의심되는 위험인자를 추출해 내는데 중점을 둔다. 이로써, 예전에 간경변을 판정하는데 사용했던 검사 항목에 비해 더 추가될 항목이 있는지 살펴본다.

위험인자를 선택하는 방법으로는 LDA(Linear discriminant analysis), QDA(Quadratic discriminant analysis), 로지스틱 회귀모형(Logistic Regression),

CART(Classification and Regression Tree), Filter로 불리는 FINCO(Forward and Inconsistency, Edgar Acuna, 2003)를 사용하였으며, 이 방법들로 선택된 변수들을 비교해 보고자 한다. 최근 Edgar Acuna가 개발한 FINCO 방법의 알고리즘을 소개하고, 본 건강검진 자료를 통한 결과로 이 방법의 특징을 알아본다.

제 2장 위험인자 예측 방법

본 논문에서는 건강검진(Screening test) 자료를 통해 간경변 의심군(Screening Positives of Liver Cirrhosis)에 대한 위험인자(Risk Factor)를 예측하고자 한다. 예측 모형을 통해 위험인자를 추출하는 방법과 Filtering을 통해 변수를 한 번 거른 후 추출하는 방법이 있다. 예측 모형으로는 판별모형인 LDA와 QDA, 로지스틱 회귀모형, CART를 사용하였고, Filtering 방법으로는 FINCO(Forward and Inconsistency)를 사용하였다.

2.1 판별모형

판별분석은 어떤 새로운 대상의 특징을 파악하여 미리 정의되어 있는 분류자(Classifier)에 따라 어느 한 범주에 할당하거나 나누는 방법이다. 분류자는 보통 이산형(예로, '0'과 '1' 또는 '예'와 '아니오' 등)이지만, 범주형(예로, 종교에서 '기독교', '불교', '천주교' 등)으로 주어지기도 한다. 본 논문에서는 분류자가 간경변에 대해 이산형(간경변 의심군=1, 간경변 비의심군=0)이다.

2.1.1 LDA

다변량 정규분포를 따를 때, 두 집단의 평균과 공분산은 모르지만, 두 집단의 공분산이 같은 경우로, 선형식의 형태로 집단을 분류한다. $\overline{X_i}$를 각 집단의 표본 평균 벡터라고 하고, S_i를 각 집단의 표본 공분산 행렬이라고 할 때, 합동 표본 분산(pooled sample variance)는 다음과 같다.

식 1. 합동 표본 분산(pooled sample variance)

$$S_p = \frac{n_1 - 1}{n_1 + n_2 - 2} S_1 + \frac{n_2 - 1}{n_1 + n_2 - 2} S_2$$

X_0라는 새로운 관찰치에 대한 판별 함수는 다음과 같다.

식 2. LDA의 판별 함수

$$(\overline{X_1} - \overline{X_2})' S_p^{-1} X_0 - \frac{1}{2} (\overline{X_1} - \overline{X_2})' S_p^{-1} (\overline{X_1} + \overline{X_2})$$

판별함수의 값이 $\log\left[\frac{\pi_2}{\pi_1}\right]$보다 크면 집단 1에, 작으면 집단 2에 분류한다.

2.1.2 QDA

다변량 정규분포를 따를 때, 두 집단의 평균과 분산을 모르고, 두 집단의 분산이 같지 않은 경우이다. 이 경우에는 LDA와는 다르게 선형식이 아닌 곡선의 형태로 집단을 분류한다. X_0라는 새로운 관찰치에 대한 판별 함수는 다음과 같다.

식 3. QDA의 판별 함수

$$-\frac{1}{2} X_0' (S_1^{-1} - S_2^{-1}) X_0 + (\overline{X_1}' S_1^{-1} - \overline{X_2}' S_2^{-1}) X_0 - k$$

$$k = \frac{1}{2}\log\left(\frac{|S_1|}{|S_2|}\right) + \frac{1}{2}\left(\overline{X_1}' S_1^{-1} \overline{X_1} - \overline{X_2}' S_2^{-1} \overline{X_2}\right)$$

판별 함수가 $log\left(\frac{\pi_2}{\pi_1}\right)$보다 크면 집단 1로, 작으면 집단 2로 분류한다.

2.2 로지스틱 회귀모형

회귀분석은 오차항이 정규성과 등분산성을 가정하는데, 종속 변수가 이산형 값을 갖는 경우 위의 가정이 만족하지 않기 때문에, 회귀분석이 불가능하다. 그래서 종속 변수에 로짓 변환을 해주게 되는데, n개의 독립변수를 $x_1, x_2, \cdots, x_n$이라 할 때, 로지스틱 회귀모형은 다음과 같다.

식 4. 로지스틱 회귀 모형

$$logit(p_x) = log\left(\frac{p_x}{1-p_x}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

2.3 CART

CART는 반응 변수를 결정하는데 매우 중요한 독립변수를 선택하는 비모수적인 방법이다. 반응 변수가 연속형이면, CART는 회귀 나무를 형성하고, 반응 변수가 이산형이면, 판별 나무를 형성한다. 의사결정 나무는 순환적 분할(recursive partitioning) 방식을 이용하여 나무를 구축하는 기법으로, 나무의 가장 상단에 위치하는 뿌리마디(root node), 속성의 분리기준을 포함하는 내부마디(internal node), 마디와 마디를 이어주는 가지(link), 그리고 최종 분류를 의미하는 잎(leaf)으로 구성된다. 나무를 형성하는 단계는, 먼저 의사결정나무를 최대한 크게 형성하고, Deviance, Entropy, Gini Index를 이용해 적절하게 가지치기를 한다. 본 논

문은 Deviance가 가장 작을 때 가지 치는 방법을 사용했다.

제 3장 Filter Methods

Filter Method는 앞에 소개된 판별모형, 로지스틱 회귀모형, CART와는 달리 종속변수를 사용하지 않고 독립변수의 부분 집합(subset) 중 최상의 부분 집합을 선택한다. 즉 데이터의 분석에 앞서 독립변수들을 평가하는데 자료의 특징만을 사용하여 차수를 줄이도록 변수를 걸러낸다는 뜻이다.

Feature selection method는 독립변수의 부분 집합이 생성되는 생성 함수(Generation Function)와 독립변수의 개수를 결정하기 위해서 사용되는 평가함수(Evaluation Function)에 달려있다. 생성 방법으로는 Complete, Heuristic, Random이 있고, 평가함수로는 거리 측도(Distance Measure), 정보 측도(Information Measure), 종속성 측도(Dependence Measure), 일치성 측도(Consistency Measure), 오분류율(Misclassification Error Rate)이 있다.

3.1 Feature Selection Method

3.1.1 Feature Selection 소개

Feature selection은 차원 축소를 위한 효율적인 기법이다. 실질적인 분석 단계 전에 Pre-Processing 단계에서 데이터 클리닝(Data Cleaning)을 포함해 Feature Selection을 하게 된다. 특히 Feature Selection에서는 특정의 변수를 찾아내기 위해서 변수를 생성하는 생성함수와 생성된 변수를 평가하는 평가함수를 사용한다. 생성함수의 종류로는 Random(랜덤하게 변수를 생성함), Complete(모든 변수를 고려하여 유의하지 않은 변수를 하나씩 제거함), Heuristic(변수 하나씩 고

려하여 유의한 변수를 하나씩 추가함)이 있고, 평가함수로는 거리 측도, 정보 측도, 종속성 측도, 일치성 측도가 있다.

3.1.2 Feature Selection 단계

Feature Selection 방법에는 기본적으로 4단계로 진행된다. 첫째, 평가에 필요한 변수 집합(candidate subset of features)을 생성하는 생성 단계, 둘째, 선택된 변수 집합을 평가하는 평가 단계, 셋째, Feature Selection이 언제 멈출 것인지 결정하는 정지 기준(stopping criterion)을 설정하는 단계, 넷째, 최종 선택된 변수 집합이 타당한지 확인하는 검증 단계이다. 이 단계를 그림으로 표현하면 아래와 같다.

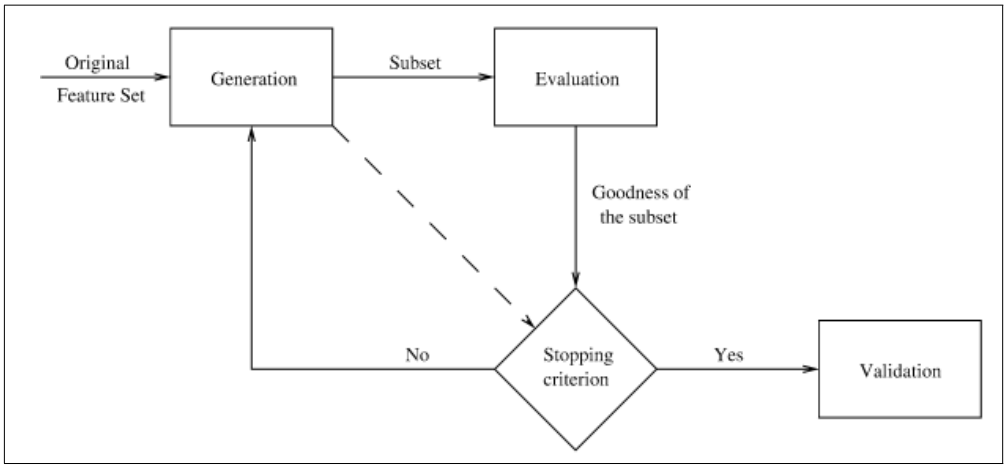

그림 1. Feature Selection 단계

N을 총 변수의 개수라고 하면, 2^N개의 변수 집합을 모두 비교해 가장 좋은 변수 집합을 찾아내는 방법이 가장 좋을 것이다. 그러나 이 모든 변수 집합을 비교하기에는 비용면, 시간면에서 경제적이지 못하다. 따라서 생성 단계에는 3가지 방법이 있는데, 먼저 집합을 공집합으로 설정해 놓은 후에 가장 유의한 변수를

하나씩 추가하는 방법인 Heuristic, 전체 집합으로 설정해 놓은 후에 가장 유의하지 않아 보이는 변수를 하나씩 제거하는 방법으로 Complete, 임의로 변수 집합을 선택하는 Random 방법이 있다.

평가 단계에서는 평가함수가 생성 함수에 의해 만들어진 변수 집합의 goodness를 측정한다. 반복을 통해 가장 좋은 goodness를 가진 변수 집합과 현재의 변수집합을 비교해 더 좋은 변수 집합으로 교체한다. 생성 함수가 달라지면 평가 함수가 같더라도 변수 집합은 달라질 수 있다.

Feature selection 단계에서 정지 기준이 없으면 계산이 무한히 길어질 수 있는데, 앞서 설명한 생성 함수와 평가 함수에 의해 정지 기준이 설정된다. 미리 선택될 변수의 개수를 제한하거나, 반복 횟수를 제한하는 것이 정지 기준이 될 수 있다. 또한 평가 함수의 적절한 값을 Threshold로 사용하여 정지 기준을 설정할 수 있다.

Feature selection의 마지막 단계인 검증 단계는 이전에 정의된 Feature selection의 여러 가지 방법을 통해 나온 결과와 비교하거나, 실제 데이터 셋을 사용해 방법들 간의 결과를 비교함으로써 변수 집합의 타당도를 시험하게 된다.

3.2 FINCO 소개

FINCO는 2003년 Edgar Acuna에 의해 소개된 Feature selection 방법이다. 차원 축소를 통해 계산 시간을 단축하고, 모형을 보다 쉽게 해석하려는 시도로, 변수 중에서 가장 유의한 변수 부분 집합을 선택한다. FINCO는 생성함수로 Sequential Forward Generation을 사용하고, 평가 함수로는 Inconsistency를 사용한다.

3.2.1 Sequential Forward Generation

Sequential Forward 방법은 판별 모형에서 변수를 선택하는 방법에서와 같

이, 모든 독립변수 중에서 가장 유의한 독립변수 순으로 유의한 독립변수가 나오지 않을 때까지 더해주는 방식이다.

3.2.2 Consistency Measure

일치성 측도는 보다 새롭고 가장 최근에 주목을 받는 방법으로, 단조 증가·감소(monotonic)하고, 유의한 독립변수 집합을 선택하는데 걸리는 시간이 짧고, 중복되거나 관계없는 변수를 제거하고, noise를 제거하는 장점이 있다. 일치성 측도는 불일치율(Inconsistency rate)을 계산해서 불일치율이 0%이면 일치성 측도는 100%이 된다.

N을 총 독립변수(feature)의 수라고 하면, 총 독립변수의 집합은 2^N개가 된다. 이 중 선택된 n개의 독립변수의 집합을 $\{S_0, S_1, ..., S_n\}$ ($S_0 \supset S_1 \supset ... \supset S_n$)라고 하면, 불일치율 U는 $U(S_0) \leq U(S_1) \leq ... \leq U(S_n)$가 된다.

D를 전체 데이터, S는 선택된 feature의 집합, N는 D안에 있는 개개인의 수라고 하면, 불일치율을 구하는 식은 다음과 같다.

식 5. 불일치율

$$Inconsistency(S,P) = \frac{\sum_{i=1}^{M} |P_i| - |M_i|}{N}$$

$|P_i|$는 S에 대해서 i번째 feature 값의 조합이 발생할 횟수, $|M_i|$는 i번째 feature 값을 가지는 개개인들이 제일 많이 속해있는 class의 횟수, M은 S에 있는 feature 값의 모든 조합 수가 된다.

불일치율을 계산하는 과정은 다음과 같다. 먼저 S_i (i=0,…,n)에 있는 개개인

이 갖는 변수 값들을 패턴(pattern)이라고 하고, 각 패턴이 있는 마지막 값을 class label로 둔다. 두 패턴을 비교하여 모두 같은 값을 갖는데 class label이 다르면 불일치한(Inconsistent) 것이다. 이렇게 한 독립변수의 집합 안에서 나타나는 불일치 횟수에서 가장 많이 나온 class label의 횟수를 뺀 것을 Inconsistency count라고 한다. 예로, m(m≤M)개의 matching 패턴이 있었고, class label이 1인 횟수 C_1, class label이 2인 횟수 C_2, class label이 3인 횟수 C_3라고 하면, m=C_1 +C_2+C_3이 된다. C_2가 가장 큰 값이었다면, Inconsistency count는 m-C_2가 된다. 한 독립변수 집합에서 모든 가능한 패턴에 대한 Inconsistency count의 합을 총 individual의 수로 나눈 것이 불일치율(Inconsistent rate)이 된다.

3.2.3 FINCO 알고리즘

처음에는 비어 있는 T 안에 각 step이 진행될 때마다 가장 낮은 불일치율(Inconsistency rate)을 가진 feature가 차례차례 더해진다. Sequential forward의 특성으로 T안에 선택된 feature는 절대로 제거될 수 없다. 이 과정은 미리 정의된 Threshold보다 작을 때까지 진행된다. 자세한 FINCO 알고리즘은 아래와 같다.

단계 1 : Input 단계

D=데이터 셋, p=D 안에 있는 변수의 수, S=D 안에 있는 모든 변수의 집합, Threshold=δ

단계 2 : Initialization 단계

k=0, T_k=∅ (T_k는 k번째 단계까지 선택된 변수의 집합)

단계 3 : Inclusion 단계

For k=1 to p

$x^+ = argmin Incons(T_k + x)$, $x \in S - T_k$

여기서, $S - T_k$는 아직 선택되지 않은 변수의 집합이고,

$Incons(T_k + x)$는 T_k에 x를 포함시켰을 때의 Inconsistency rate이고,

따라서, x^+는 T_k에 대해 가장 중요한 변수가 된다.

단계 4 : Termination 단계

$Incons(T_{k+1}) > Incons(T_k)$이거나, $Incons(T_{k+1}) \leq \delta$이면 종료한다.

단계 5: Output 단계

T_k는 최종 선택된 변수의 집합과 불일치율(Inconsistency rate)

제 4장 건강검진 자료를 통한 실증 분석

4.1 건강검진 자료 소개

본 논문은 1994년 5월 30일부터 2005년 8월 18일까지 총 124,121건의 건강검진 자료를 통해 간경변 의심군에 대한 위험 인자를 추출하는데 주 목적을 둔다. 건강 검진 항목에는 기초 정보를 기본으로 신체계측, 안과 검사, 청력 검사, 폐 기능 검사, 혈액 검사, 면역혈청 검사, 대사 및 전해질 검사, 간 기능, 신 기능 검사, 지혈 검사, 간염검사, 갑상선 기능 검사, 소변 검사, 대변 검사, 심전도 검사, 부인과 검사, 내시경 검사, 복부 초음파 검사, 흉부 X-선 검사, 유방 X-선 검사, 위 검사, 영양 상담, 운동능력 검사, 피로도 측정이다. 이 외에 보다 정밀한 검사를 요하는 환자들에 대해서는 선택 검사를 하도록 하였는데, 이 중 간염 활동성 여부를 확인하는 검사로 B형 간염 보균자에만 해당되는 HBeAg와 Anti-HBe가 있었다. 이 두 항목은 간경변 의심군의 위험 인자로 보이지만, 이는 사비를 들여야 하는 선택 항목이므로 극히 소수의 환자들만 검사를 했기 때문에, 이 두 가지 항목은 변수로 제외시켰다.

본 논문의 목적은 간경변 의심군의 위험인자를 알아내기 위함이므로, 위의 건강 검진 항목 중에서 간경변 의심군의 위험 인자로 알려진 이전의 자료들을 바탕으로 성별, 연령, 신체계측, 간 기능 검사, 간염 검사, 소변 검사, 초음파 검사, 내시경 검사 등으로 분석을 실시하였다. 총 124,121건의 건강검진 자료에는 선택 검사를 포함해 최대 7번까지 검사를 받은 사람이 있었다. 이를 가장 최근에 건강검진을 받은 날짜를 기준으로 선택해 64,211명이 최종으로 뽑혔으며, 본 논문은 이를 대상으로 분석을 실시하였다.

종속 변수 즉, 간경변 의심군의 여부를 알 수 있는 정밀한 검사가 없었기 때문에, 종속변수를 몇 가지의 혈액 검사 항목(5항목)과 초음파 검사(5항목)와 내시경 검사(5항목) 결과로 만들었다. 이 종속변수의 설정은 해당분야 전문가들이 제시한 혈액 검사, 초음파 검사, 내시경 검사의 소견으로 이루어졌으며, 본 논문

은 3개의 종속변수를 만들어냈다.

표 1. 건강 검진 기본검사

검사 항목		변수	검사 항목		변수
기초 정보		성별 연령	생화학검사	대사 및 전해질	나트륨 칼륨 염소 이산화탄소 칼슘 인 혈당 당화혈색소 혈중요소질소 크레아티닌 요산
신체 계측		신장 체중 비만도			
혈액검사	적혈구	RBC Hb Hct MCV MCH MCHC		간 기능	총단백 알부민 총빌리루빈 직접빌리루빈 Alk.Phos AST(GOT) ALT(GPT) γ-GT LDH 크레아틴키나제 혈청 철분 총 철결집함능
	백혈구	WBC PMN(다핵구) LYM(임파구) MON(단핵구) EOS(호산구) BAS(호염구)			
	혈소판	Platelet(혈소판)		혈청지질	총콜레스테롤 중성지방 고밀도콜레스테롤 저밀도콜레스테롤
	혈액형	ABO RH		간염검사	B형간염S항원 B형간염C항체 B형간염S항체 C형간염항체 B형간염E항원 B형간염E항체
뇨 검사		비중 산도 단백 요당 케톤체 잠혈 요빌리노겐 빌리루빈 아질산염 백혈구 색 탁도			
대변 검사		잠혈 기생충검사			

4.2 간경변 의심군의 정의

건강 검진 자료에는 간경변 여부에 대한 정확한 검사 결과가 없었기 때문에, 간경변 의심군을 예측하는데 필요한 종속 변수의 정보가 존재하지 않았다. 따라서 간경변 의심군의 종속 변수 설정은 간 기능 검사, 간염검사, 혈소판 수치, 초음파 검사, 내시경 검사 중 몇 가지 결과들의 조합으로 이루어졌다. 이는 해당 분야 전문가들에 의해 제시된 것이다. 간경변 종속 변수는 “간경변 의심군”과 “간경변 비의심군”으로 나뉘어졌으며, 자세한 항목은 아래 표와 같다.

표 2. 간경변 의심군 설정 기준표

종속변수	검사 항목	하위 검사 항목 결과
간경변 의심군	내시경	식도 정맥류(Esophageal varix)
		식도 정맥류(Isolated Esophageal varix)
		식도 및 위저부 정맥류(Esophageal & fundal varix)
		울혈성 위병증(Congestive gastropathy)
		위 저부 정맥류(Fundal varix)
	초음파	비장확장(Splenomegaly)
		복수(Ascites)
		초기간경화(Early stage liver cirrhosis)
		간경화의심(rule out liver cirrhosis)
		간경화(Liver cirrhosis)
	혈액검사	혈소판(Platelet) < 130 (10^3/uL)
		알부민(Albumin) < 3.3(g/dL)
		총빌리루빈(T.Bil) > 1.2(mg/dL)
		B형간염S항원 HBsAg 양성
		C형간염항체 AntiHCV 양성

위에 제시된 간경변 의심군 기준표를 이용해 몇 가지 조합에 의해 3개의 종

속변수를 만들었다. 종속변수 조합에 앞서 건강검진 자료의 구조는 아래 그림과 같고, 간경변 의심군 조합은 아래 표와 같다.

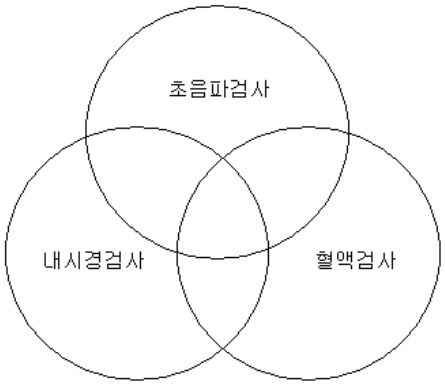

그림 2. 종속변수 조합을 위한 건강 검진 자료의 구조

표 3. 설정된 종속변수

종속변수1	초음파검사∪내시경검사∪혈액검사
종속변수2	초음파검사∪내시경검사
종속변수3	초음파검사∪(내시경검사∩혈액검사)

4.3 분석자료

본 논문은 1994년 5월 30일부터 2005년 8월 18일까지 총 124,121건의 자료 중에는 1번에서 최고 7번까지 건강 검진을 받은 사람이 있었다. 이 중에는 기본 검사 외에 보다 세밀한 선택 검사를 받은 횟수도 포함하고 있다. 건강 검진을 많이 받은 사람일수록 건강 관리에 심혈을 기울이고 있으므로 건강 검진 결과는 좋았다. 본 논문은 가장 최근에 받은 자료만을 선정해 64,211명으로 간경변 의심군

에 대한 위험인자를 추출해 내도록 한다.

간경변 의심군을 정의하는 종속변수 3개에 대해서는 이미 설정 기준과 설정 기준의 조합을 설명했다. 이에 대해 각각의 종속변수에 대한 빈도수는 다음과 같다.

표 4. 간경변 의심군 종속변수의 도수표

Screening		Frequency	Percent	Total
종속변수1	의심군	10,683	16.64 %	64,211
	정상군	53,528	83.36 %	
종속변수2	의심군	288	0.45 %	
	정상군	63,923	99.55 %	
종속변수3	의심군	230	0.36 %	
	정상군	63,981	99.64 %	

간경변 의심군을 설정할 때, 가장 큰 범위로 설정된 종속변수1은 간경변 의심군이 10,683명 간경변 비의심군이 53,528명으로, 종속변수2는 간경변 의심군이 288명 간경변 비의심군이 63,923명으로, 종속변수3은 간경변 의심군이 230명 간경변 비의심군이 63,981명으로 나타났다. 간경변 의심군과 간경변 비의심군의 비율 차이가 큰 것을 감안해 사전확률을 동일하게 설정해주기로 한다. 따라서 가장 큰 범위로 설정된 종속변수1의 의심군과 정상군의 수를 동일시하여 본 논문의 분석이 시작된다.

간경변 의심군의 위험인자 추출을 위해서 기본적으로 성별, 연령, 신체계측을 포함하여 간 기능 검사, 간염 검사, 소변 검사, 초음파 검사, 내시경 검사 등 54개의 독립변수를 사용하여 본 논문의 분석이 시작된다. 54개의 독립 변수 중에는 (rbc, hb, hct), (mch, mcv, mchc), (gran, lymph), (bioAst, bioAlt)의 상관관계가 높아 다중공선성을 우려하여 rbc, mch, lymph, bioAst만 사용하기로 하고 나머지 5개 변수는 제외하였다. 따라서 독립변수 49개로 간경변 의심군에 대한 위험인자를 추출하도록 한다. 종속변수 3개에 대해 각각 영향을 미치는지 알아보기

위해 일변량 분석을 실시하였고 결과는 다음과 같다.

표 5. 종속변수1에 대한 독립변수들의 일변량 분석

독립변수	p-value	독립변수	p-value
sex	<.0001	bilirubin	<.0001
age	<.0001	bioCa	0.0031
height	<.0001	bioP	<.0001
weight	<.0001	bioGlucose	0.9340
bmi	<.0001	bioBun	<.0001
wbc	<.0001	bioCreat	<.0001
rbc	<.0001	bioUric	<.0001
mch	<.0001	bioChol	<.0001
lymph	0.0351	bioProtein	<.0001
mono	<.0001	bioAlk	<.0001
eos	0.8960	bioAst	<.0001
baso	0.0383	bioRgt	<.0001
sg	0.7942	bioLdh	<.0001
ph	0.0097	bioTrig	<.0001
protein	<.0001	bioHdl	0.0079
blood	<.0001	bioNa	0.0219
uwbc	0.0002	bioK	<.0001
nitrite	0.0003	bioCl	0.8958
glucose	0.2096	bioCo2	0.1344
ketone	<.0001	hbc	<.0001
urobil	<.0001	ahbs	<.0001

종속변수1에 대한 일변량 분석 결과 eos, sg, glucose, bioGlucose, bioCl, bioCo2 변수들은 유의하지 않은 것으로 나타나고 있다. 따라서 종속변수1에 대한 분석에서는 6개의 독립변수를 제외한다.

표 6. 종속변수2에 대한 독립변수들의 일변량 분석

독립변수	p-value	독립변수	p-value
sex	0.0461	bilirubin	0.0037
age	0.0002	bioCa	0.1622
height	0.5815	bioP	0.1831
weight	0.8132	bioGlucose	0.2035
bmi	0.8623	bioBun	0.3168
wbc	<.0001	bioCreat	0.0042
rbc	<.0001	bioUric	0.2823
mch	0.0026	bioChol	<.0001
plt	<.0001	bioProtein	0.0153
lymph	0.8643	bioAlb	<.0001
mono	0.0026	bioBil	0.1383
eos	0.0671	bioAlk	<.0001
baso	0.0031	bioAst	<.0001
sg	0.4534	bioRgt	<.0001
ph	0.9740	bioLdh	<.0001
protein	0.3497	bioTrig	0.1275
blood	0.4634	bioHdl	0.3550
uwbc	0.3122	bioNa	0.8565
nitrite	0.9201	bioK	0.2850
glucose	0.0028	bioCl	0.8878
ketone	0.1326	bioCo2	0.2884
urobil	<.0001	hcv	0.8341
ahbs	<.0001	hbs	<.0001
		hbc	0.0362

위 표는 간경변 의심군의 종속변수2에 대한 일변량 분석 결과이다. 간경변 의심군의 종속변수를 설정할 때, 종속변수2에 대해서는 혈액검사 소견이 포함되지 않았으므로, 독립변수에 혈액검사 항목(plt, bioAlb, bioBil, hbs, hcv)를 포함시켰다. 위의 결과 height, weight, bmi, lymph, eos, sg, ph, protein, blood, uwbc,

nitrite, ketone, bioCa, bioP, bioGlucose, bioBun, bioUric, bioBil, bioTrig, bioHdl, bioNa, bioK, bioCl, bioCo2, hcv 변수들은 유의하지 않은 것으로 나타나고 있다. 따라서 종속변수2에 대한 분석에서는 25개의 독립변수를 제외한다.

표 7. 종속변수3에 대한 독립변수들의 일변량 분석

독립변수	P-value	독립변수	P-value
sex	0.0050	bilirubin	<.0001
age	0.0254	bioCa	0.0502
height	0.2954	bioP	0.2227
weight	0.3417	bioGlucose	0.1133
bmi	0.6702	bioBun	0.2414
wbc	<.0001	bioCreat	0.0012
rbc	<.0001	bioUric	0.2709
mch	<.0001	bioChol	<.0001
lymph	0.7728	bioProtein	0.0017
mono	<.0001	bioAlk	<.0001
eos	0.0039	bioAst	<.0001
baso	0.0012	bioRgt	<.0001
sg	0.3317	bioLdh	<.0001
ph	0.8834	bioTrig	0.0243
protein	0.1075	bioHdl	0.5195
blood	0.9752	bioNa	0.2066
uwbc	0.4588	bioK	0.6352
nitrite	0.8438	bioCl	0.5518
glucose	0.0012	bioCo2	0.3008
ketone	0.0261	hbc	0.0004
urobil	<.0001	ahbs	<.0001

종속변수3에 대한 일변량 분석 결과 height, weight, bmi, lymph, sg, ph, protein, blood, uwbc, nitrite, bioCa, bioP, bioGlucose, bioBun, bioUric, bioHdl, bioNa, bioK, bioCl, bioCo2 변수들은 유의하지 않은 것으로 나타나고 있다. 따라

서 종속변수3에 대한 분석에서는 20개의 독립변수를 제외한다.

4.4 간경변 의심군의 위험인자 선택

간경변 의심군에 대한 위험인자를 선택함에 있어서, 종속변수1, 2, 3에 대해 각각 판별 모형, 로지스틱 회귀모형, CART와 Filter Method인 FINCO를 사용했다. 먼저 판별모형으로 선택된 위험인자는 아래 표와 같다.

종속변수1에 대해 판별 모형으로 선택된 위험인자로는 간 기능 검사(bioAst, bioProtein, bioLdh), 간염검사(hbc, ahbs, mch), 혈액검사(wbc,rbc), 혈청 지질 검사(bioTrig, bioChol, bioHdl), 대사 및 전해질 검사(bioUric, bioK, bioP, bioBun, bioCa), 뇨검사(urobil, uwbc, protein, blood, ketone), 신체계측(age, , sex, height, weight)이었고, 종속변수2에 대해 판별 모형으로 선택된 위험인자는 간 기능 검사(bioAlb, bioAlk, bioRgt, bioProtein), 간염검사(hbs), 혈액검사(plt), 혈청 지질 검사(bioChol)였고, 종속변수3에 대해 판별모형으로 선택된 위험인자는 간 기능 검사(bioAlk, bioRgt, bioLdh), 간염검사(hbc, ahbs), 혈액검사(rbc, eos, wbc), 혈청지질 검사(bioChol), 인적사항(sex)이었다.

표 8. 판별모형을 통한 위험인자 선택

종속변수1			종속변수2			종속변수3		
Variable	F Value	p-value	Variable	F Value	p-value	Variable	F Value	p-value
ahbs	523.16	<.0001	plt	109.53	<.0001	bioAlk	113.22	<.0001
sex	350.60	<.0001	bioAlk	92.36	<.0001	bioChol	40.71	<.0001
hbc	255.24	<.0001	bioAlb	34.60	<.0001	bioRgt	37.75	<.0001
urobil	174.29	<.0001	bioRgt	17.09	<.0001	ahbs	15.55	<.0001
bioChol	119.01	<.0001	bioChol	14.57	0.0001	wbc	12.94	0.0003
bioK	86.21	<.0001	hbs	12.85	0.0003	rbc	8.76	0.0031
wbc	56.88	<.0001	bioProtein	3.12	0.0772	sex	9.99	0.0016
bioProtein	54.99	<.0001				bioLdh	7.19	0.0074
height	47.84	<.0001				eos	4.27	0.0389
bioAst	35.57	<.0001				hbc	2.55	0.1102
bioLdh	60.37	<.0001						
bioHdl	47.73	<.0001						
rbc	34.30	0.0101						
mch	72.41	0.0184						
bioUric	18.67	<.0001						
ketone	17.01	<.0001						
weight	7.91	0.0049						
uwbc	6.96	0.0084						
bioP	7.41	0.0065						
bioTrig	6.62	0.0101						
blood	6.02	0.0142						
bioBun	5.45	0.0196						
age	6.02	0.0141						
bioCa	4.08	0.0434						
protein	3.77	0.0522						

세 종속변수에서 빈번히 나타나는 간경변 의심군의 위험인자로는 bioRgt(γ-GT), bioProtein(총단백), bioAlk(알칼리 포스파타제), bioLdh(젖산탈수소효소), hbc(B형간염C항체), ahbs(B형간염S항체), WBC, rbc(적혈구 수), bioChol(총콜레

스테롤), sex가 선택되었다.

표 9. 로지스틱 회귀모형을 통한 종속변수1의 위험인자 선택

Parameter	Estimate	Standard Error	p-value
Intercept	-7.2185	0.8470	<.0001
age	-0.00535	0.00223	0.0162
height	0.0249	0.00392	<.0001
weight	-0.00986	0.00297	0.0009
wbc	-0.1151	0.0135	<.0001
rbc	0.5809	0.0605	<.0001
mch	0.1013	0.0119	<.0001
blood	-0.0993	0.0490	0.0427
uwbc	0.2237	0.0730	0.0022
ketone	0.3416	0.1128	0.0025
urobil	1.2711	0.1147	<.0001
bioP	-0.1686	0.0465	0.0003
bioBun	0.0179	0.00623	0.0041
bioUric	0.0903	0.0201	<.0001
bioChol	-0.00539	0.000713	<.0001
bioProtein	0.2121	0.0552	0.0001
bioAst	0.0183	0.00204	<.0001
bioLdh	-0.00198	0.000197	<.0001
bioTrig	-0.00073	0.000272	0.0070
bioHdl	0.00916	0.00204	<.0001
bioK	-0.5052	0.0664	<.0001
hbc	0.7620	0.0467	<.0001
ahbs	-1.0552	0.0459	<.0001

로지스틱 회귀 모형을 통해 종속변수1의 위험인자를 추정한 결과 신체계측(나이, 신장 몸무게), 뇨검사(blood, uwbc, ketone, urobil), 간염검사(hbc, ahbs)가 간경변 의심군의 위험인자로 선택되었다.

표 10. 로지스틱 회귀모형을 통한 종속변수2의 위험인자 선택

Parameter	Estimate	Standard Error	p-value
Intercept	0.1035	0.9909	0.9168
plt	-0.0105	0.00145	<.0001
bioChol	-0.00660	0.00248	0.0078
bioAlb	-0.4297	0.2046	0.0357
bioAlk	0.00783	0.00173	<.0001
bioLdh	0.00126	0.000564	0.0255
hbs	0.5837	0.1852	0.0016

종속변수2에 대한 로지스틱 회귀 모형에서는 간경변 의심군 위험인자 로 이미 간 질환의 위험인자로 알려진 plt, bioAlb, hbs가 추정되었다. 그 외에도 hbs(B형간염S항원: 양성이면 간 질환 판명), bioLdh(젖산탈수소효소: 젖산탈수소효소 수치는 급성 간염의 경우 급상승함)와 bioAlk(알칼리 포스파타제: 간 기능, 뼈의 이상 유무를 판단함)가 위험인자로 추정되었다.

표 11. 로지스틱 회귀모형을 통한 종속변수3의 위험인자 선택

Parameter	Estimate	Standard Error	p-value
Intercept	-4.5534	1.4981	0.0024
sex	-0.4151	0.2049	0.0428
wbc	-0.1838	0.0581	0.0016
mch	0.1006	0.0389	0.0097
eos	0.0737	0.0306	0.0159
bioChol	-0.0135	0.00275	<.0001
bioAlk	0.0120	0.00164	<.0001
bioLdh	0.00192	0.000582	0.0010
ahbs	-0.7627	0.1879	<.0001

종속변수3에 대한 로지스틱 회귀모형에서 간경변 의심군의 위험인자 추정 결과에서는 sex, wbc(백혈구 수), mch(적혈구 1개당 혈색소 양), eos(호산구), bioChol 등이 선택되었다.

다음으로는 CART에 대해 간경변 의심군의 위험인자를 그림으로 표현해 보았다. 종속변수1의 위험인자로는 아래 그림과 같이 ahbs, hbc, sex, age, urobil, bioAst가 선택되었다. ahbs(B형간염S항체)에 양성 반응을 보이고, 요빌리노겐 수치가 0.55보다 크면 간경변 의심군으로 판명된다.

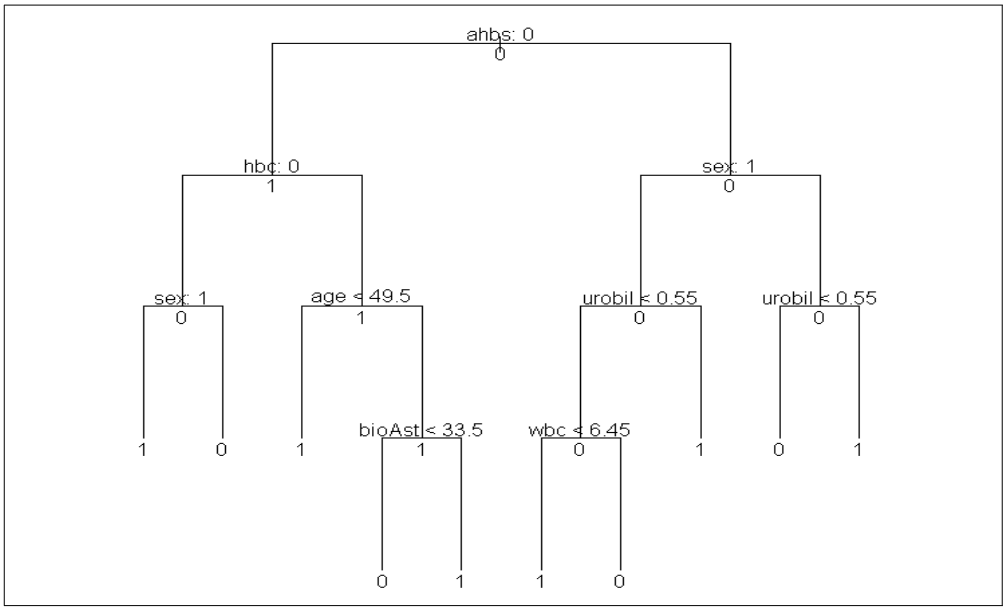

그림 3 CART로 종속변수1의 위험인자 선택

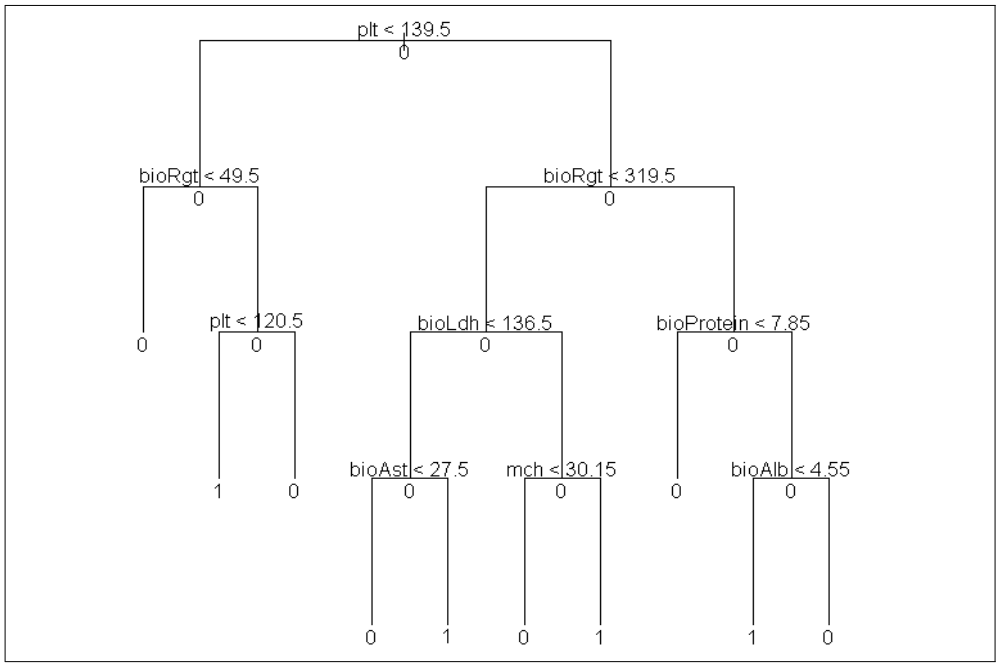

그림 4. CART로 종속변수2의 위험인자 선택

위의 그림은 CART로 종속변수2의 위험인자를 추정한 결과이다. 위험인자로는 plt, bioRgt(감마 글루타밀 트랜스: 간에 이상이 생기면 혈액 중으로 누출되어 이상치를 나타냄, 알코올 섭취시 높게 나타남), bioLdh, bioProtein으로 선택되었으며, plt가 120.5보다 작고 bioRgt 49.5보다 간경변 의심군으로 판명된다.

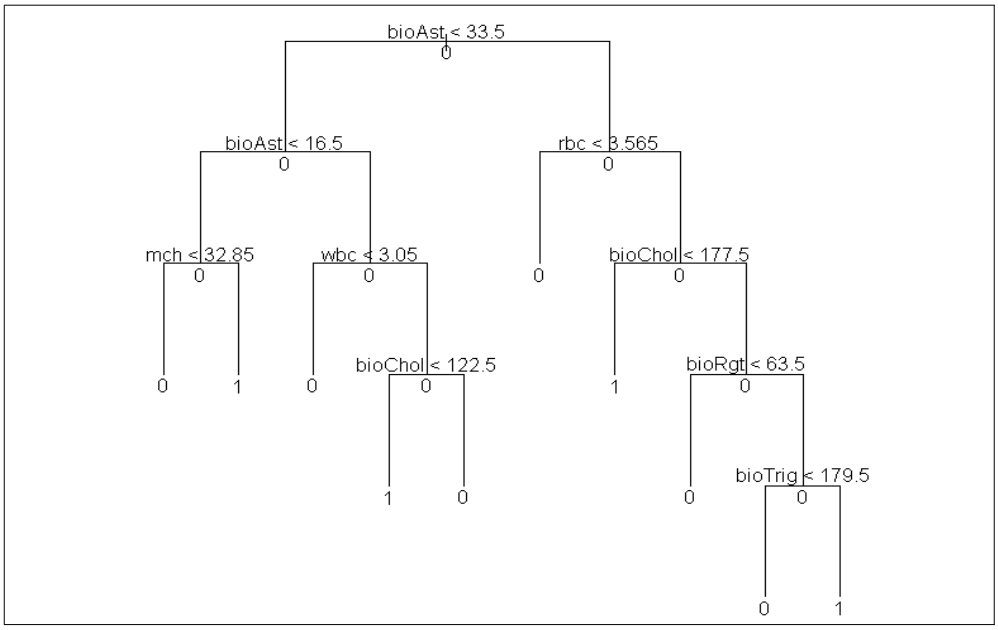

그림 5. CART로 종속변수3의 위험인자 선택

위의 그림은 CART로 종속변수3의 위험인자를 추정한 결과이다. bioAst가 16.5보다 작고 mch가 32.85보다 크면 간경변 의심군으로 판명되고, bioAst가 33.5보다 크고 rbcrk 3.565보다 크고 bioChol이 177.5보다 작으면 간경변 의심군으로 판명된다.

4.5 간경변 의심군 예측 모형

간경변 의심군의 위험인자를 선택하는데 사용했던 LDA, QDA, 로지스틱 회귀모형, CART에 대한 예측 모형의 우수성을 민감도와 특이도에 대해 알아본다. 위의 방법으로 얻어진 예측 모형을 통해 판별된 '간경변 의심군'과 '간경변 비의심군'이 실제 판별된 '간경변 의심군'과 '간경변 비의심군'과 얼마나 정확한지에 따라 예측 모형의 우수성이 결정되는데, 민감도와 특이도는 아래 표와 같이 Confusion Matrix에서 나타낼 수 있다.

표 12. Confusion Matrix

		Predicted model	
		간경변 비의심군	간경변 의심군
Actual data	간경변 비의심군	Specificity(특이도)	False Positive
	간경변 의심군	False Negative	Sensitivity(민감도)

위의 표에서 나타나듯이, 실제 간경변 의심군을 예측 모형을 통해서도 간경변 의심군으로 판명하는 정도를 민감도라 하고, 실제 간경변 비의심군을 예측 모형을 통해서도 간경변 비의심군으로 판명하는 정도를 특이도라 한다. 또한 실제 간경변 비의심군을 의심군으로(False Positive), 실제 의심군을 비의심군으로(False Negative) 잘못 판명한 정도를 오분류율이라고 한다. 민감도와 특이도는 높을수록 오분류율은 낮을수록 예측 모형의 우수성이 높게 나타나므로 여러 방법에 의해 예측된 모형의 민감도와 특이도를 비교함으로써 예측 모형의 우수성을 비교할 수 있다.

본 논문에서 사용한 예측 모형인 LDA, QDA, 로지스틱 회귀모형, CART의 특이도, 민감도, 오분류율을 비교해보면 아래의 표와 같다.

표 13. 간경변 의심군의 종속변수1에 대한 예측 모형 분류표

		예측모형	Sensitivity	Specificity	False Positive	False Negative	Accuracy
종속변수1	Training	LDA	66.51	70.15	29.85	33.49	68.35
		QDA	24.82	90.33	9.67	75.18	58.01
		Logistic Regression	66.25	71.42	28.58	33.75	68.87
		CART	68.77	64.95	35.05	31.23	66.84
	Test	LDA	64.73	69.86	30.14	35.27	67.27
		QDA	23.85	97.62	2.38	76.15	55.99
		Logistic Regression	64.20	69.82	30.18	35.80	66.98
		CART	67.95	64.50	35.50	32.05	66.24

검증용 자료에서 종속변수1의 예측 모형 정확도는 로지스틱 회귀모형(66.98%), LDA(67.27%), CART(66.24%), QDA(55.99%)로 로지스틱 회귀모형이 제일 높게 나왔고, 민감도에서도 마찬가지로 CART가(67.95%) 높게 나왔다. 대체로 종속변수1에 대해서는 만족스러운 결과가 나온 것으로 보인다.

표 14. 간경변 의심군의 종속변수2에 대한 예측 모형 분류표

		예측모형	Sensitivity	Specificity	False Positive	False Negative	Accuracy
종속변수2	Training	LDA	13.42	99.30	0.70	86.58	98.08
		QDA	22.82	98.07	1.93	77.18	96.79
		Logistic Regression	2.68	99.98	0.02	97.32	98.61
		CART	14.77	99.97	0.03	85.23	98.76
	Test	LDA	13.33	99.19	0.81	86.67	98.33
		QDA	13.33	97.64	2.36	86.67	96.79
		Logistic Regression	2.22	99.98	0.02	97.78	98.99
		CART	2.22	99.89	0.11	97.78	98.91

표 15. 간경변 의심군의 종속변수3에 대한 예측 모형 분류표

		예측모형	Sensitivity	Specificity	False Positive	False Negative	Accuracy
종속변수3	Training	LDA	20.33	99.13	0.87	79.67	98.00
		QDA	33.33	98.49	1.51	66.67	97.73
		Logistic Regression	5.69	99.99	0.01	94.31	98.88
		CART	5.69	99.99	0.01	94.31	98.83
	Test	LDA	17.14	99.03	0.97	82.86	98.40
		QDA	14.29	98.00	2.00	85.71	97.35
		Logistic Regression	2.86	99.97	0.03	97.14	99.22
		CART	2.86	99.97	0.03	97.14	99.22

위의 표 4.14와 표 4.15은 종속변수2와 종속변수3에 대한 예측 모형 분류표이다. 종속변수1에서 보다는 민감도가 많이 떨어지는 결과를 보이는데 이는 간경변 의심군과 비의심군의 사전확률이 차이가 많이 나서 나타나는 결과이다. 종속변수2에서는 로지스틱 회귀모형의 정확도가 98.99%로 제일 높았고, 종속변수3에서는 CART와 로지스틱 회귀모형이 99.22%로 높았다.

4.6 FINCO에 의해 선택된 변수

본 논문에서 소개한 Feature Selection Method인 FINCO는 공집합에서 변수를 하나씩 더하면서 불일치율을 계산해 미리 정해진 불일치율 값보다 작은 변수를 선택하게 된다. 그러므로 각각 변수가 더해질 때마다 불일치율도 함께 계산되어 나온다. 이제 FINCO에 의해 선택된 변수를 가지고 예측 모형을 통해 위험인자를 선택해 본다. 먼저 FINCO에 의해 계산된 불일치율과 선택된 변수들은 다음과 같다.

표 16. FINCO에 선택된 변수와 불일치율

Selected Feature	hbs	bioAlk	mch	age	mono	weight
불일치율	0.24281211	0.24080619	0.22227529	0.10726908	0.00907441	0.00009552

Feature Selection 결과 불일치율이 큰 순서대로 변수가 선택되었고, 선택된 변수로는 hbs, bioAlk, mch, age, mono, weight가 있다. 이제 이 변수들로 예측 모형을 통해 위험인자를 선택 보도록 한다. 아래 표는 판별 모형을 통해 간경변 의심군의 위험인자를 선택한 결과이다.

표 17. 판별 분석을 통한 간경변 의심군의 위험인자

	종속변수1		종속변수2		종속변수3	
	F value	p-value	F value	p-value	F value	p-value
bioAlk	16.60	<.0001	118.83	<.0001	152.59	<.0001
hbs	4989.93	<.0001	72.64	<.0001	126.56	<.0001
mono	3.75	0.0527	5.63	0.0177	10.37	0.0002
mch	109.41	<.0001	4.31	0.0379	13.87	0.0013
age	57.31	<.0001	2.18	0.1398		
weight	150.58	<.0001				

이제 FINCO로 Filtering된 변수로 간경변 의심군 예측 모형들의 민감도, 특이도, 정확도를 비교해 본다. 이 결과와 Filtering을 하지 않은 결과와 비교해 보도록 한다.

표 18. Filtering 통한 간경변 의심군 예측 모형의 분석용 자료의 분류표

		Sensitivity	Specificity	False Positive	False Negative	Accuracy
종속변수1	LDA	94.29	97.66	2.43	5.71	95.12
	QDA	64.64	90.98	9.02	35.36	76.31
	Logistic Regression	75.61	80.99	19.01	24.39	92.78
	CART	64.86	98.55	1.45	35.14	81.94
종속변수2	LDA	96.64	99.76	0.24	3.35	98.66
	QDA	50.51	98.74	1.26	49.49	74.69
	Logistic Regression	2.68	99.98	0.02	97.32	98.99
	CART	14.77	99.97	0.03	85.23	98.76
종속변수3	LDA	56.23	99.13	0.87	43.77	73.61
	QDA	30.11	97.89	2.11	69.89	90.46
	Logistic Regression	4.56	98.13	1.87	95.44	99.03
	CART	10.46	97.89	2.11	89.54	98.83

표 19 Filtering 통한 간경변 의심군 예측 모형의 검증용 자료의 분류표

		Sensitivity	Specificity	False Positive	False Negative	Accuracy
종속변수1	LDA	90.18	97.68	2.32	9.82	92.11
	QDA	65.11	98.98	1.02	34.89	82.28
	Logistic Regression	67.66	94.40	5.60	32.34	90.31
	CART	63.56	97.43	36.44	2.57	80.19
종속변수2	LDA	96.64	99.76	0.24	3.36	98.39
	QDA	52.75	98.74	1.26	47.25	75.88
	Logistic Regression	3.03	98.23	1.77	96.97	98.76
	CART	13.88	99.98	0.02	86.12	97.64
종속변수3	LDA	5.69	99.72	0.28	94.31	98.61
	QDA	15.45	98.84	1.16	84.55	97.68
	Logistic Regression	5.01	97.99	2.01	94.99	92.47
	CART	11.05	96.67	3.33	88.95	94.29

Filter Method를 통해 간경변 의심군 예측 모형에서는 종속변수1을 제외하고 종속변수2와 종속변수3에서는 민감도가 증가한 것으로 보이며, 전체적으로 QDA의 정확도가 떨어지는 것으로 보인다. 종속변수1에 대해 LDA에서 민감도와 정확도가 각각 90.18%, 92.11%로 제일 높게 나타났으며, 종속변수2에 대해서도 LDA에서 민감도와 정확도가 각각 96.64%, 98.39%, 종속변수 3에 대해서도 LDA에서 정확도가 98.61%로 제일 높게 나타났다. 간경변 의심군 예측 모형에서 있어서 정확도도 높고 해석이 용이한 CART를 그림으로 표현해 보았다.

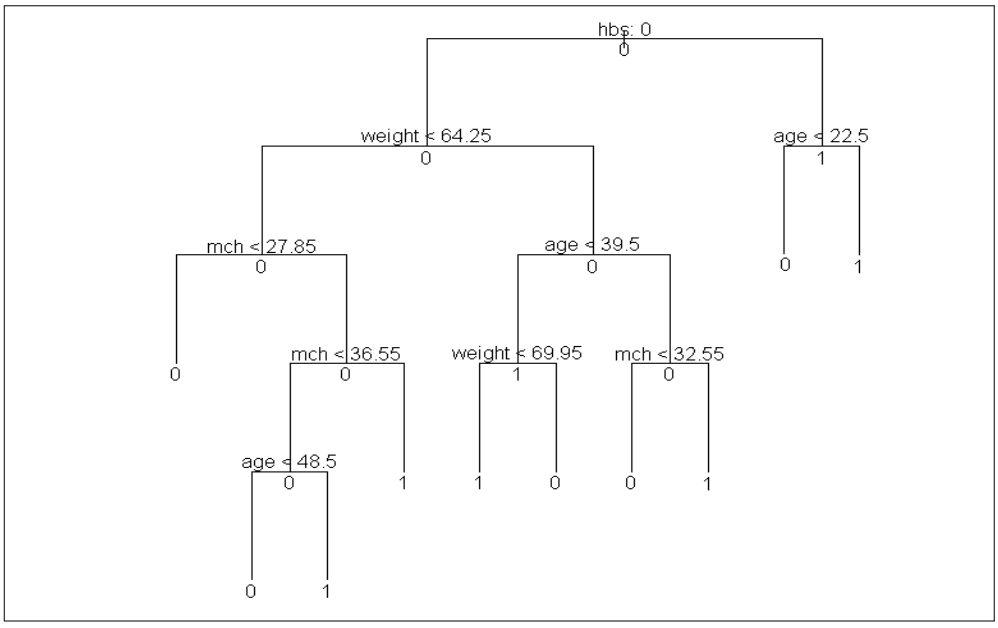

그림 6. Filtering을 통한 간경변 의심군 종속변수1의 CART 모형

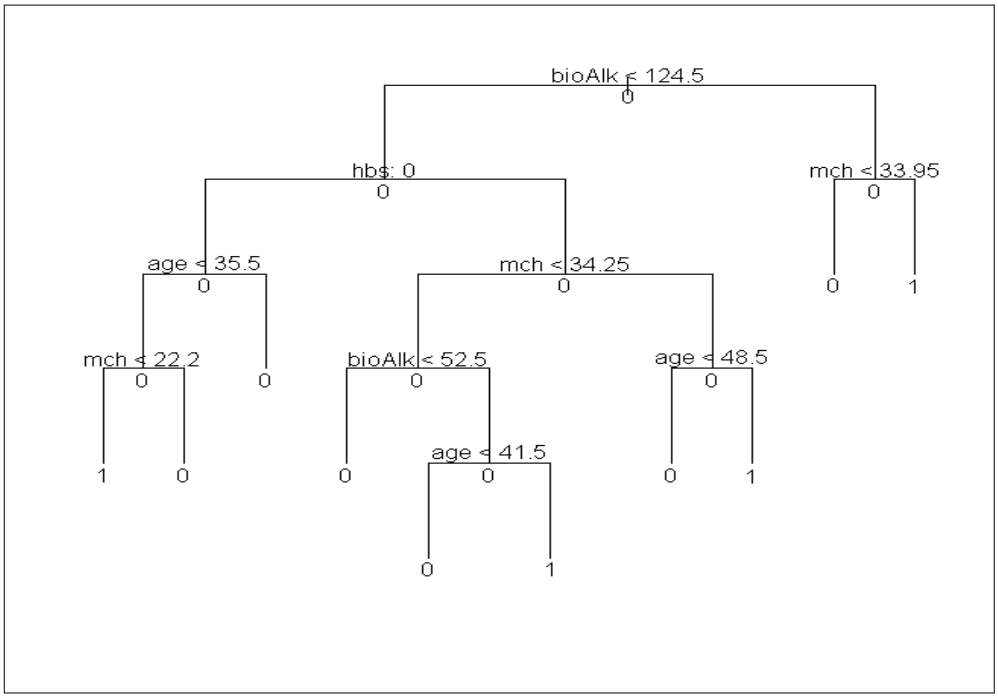

그림 7. Filtering을 통한 간경변 의심군 종속변수2의 CART 모형

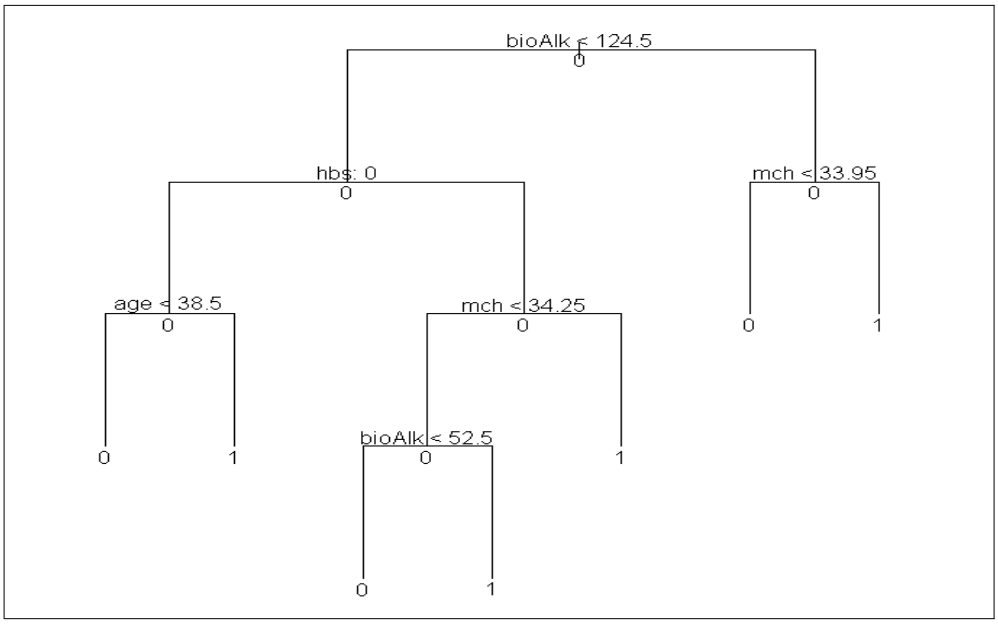

그림 8. Filtering을 통한 간경변 의심군 종속변수3의 CART 모형

제 5장 결론 및 고찰

본 논문은 간경변 의심군의 위험인자를 선택하기 위해서, 예측 모형을 적용하기 전에 Filter Method인 FINCO 알고리즘을 사용하여 58개의 독립변수를 차원 축소를 통해 건강 검진 자료에서 Feature Selection을 거쳐 예측 모형을 통해 위험인자를 선택해 보았다. Feature Selection 과정에는 feature를 생성하는 방법과 평가 함수에 따랐고, FINCO는 Sequential Forward Generation으로 feature를 생성하여, 일치성 측도로 평가하였다.

간경변의 유무에 대한 판단을 할 변수가 존재하지 않아 초음파 검사, 내시경 검사, 혈액 검사 결과의 조합을 통해 '간경변 의심군' 3개를 설정했다. Filtering을 하지 않고 예측 모형에 적용했을 때에는 알부민, platelet(혈소판), B형간염S항원, 총단백, 총콜레스테롤, LDH(젖산탈수소효소), γ-GT(감마 글루타밀 트랜스)가 간경변 의심군의 위험인자로 선택되었으며, Filtering을 통해 예측 모형에 적용했을 때에는 Alk.Phos(알칼리 포스파타제), B형간염S항원, MCH(적혈구 1개당 혈색소 양), 몸무게, 나이가 위험인자로 선택되었다. 이러한 결과에 간경변이라고 해서 간 기능이나 간염 검사 결과만으로 간경변의 유무를 판단하는 것이 아님을 알 수 있었다. 갑상선 기능, 종양혈청, 소변검사 등 신체 모든 기관과 관련된 검사도 위험인자가 될 수 있다는 것이다.

Filtering을 거치지 않고 간경변 의심군의 위험인자를 선택하는 예측 모형에서는 로지스틱 회귀모형에서 민감도와 정확도가 우수하게 나타남을 알았으며, CART에서도 민감도가 뛰어남을 알았다. Filtering을 거쳐서 간경변 의심군의 위험인자를 선택하는 예측 모형에서는 LDA를 제외하고 QDA, 로지스틱 회귀모형, CART에서 민감도가 조금씩 향상되었고, 정확도에서는 변화가 없음을 알았다.

Feature Selection 방법은 Filter Method와 Wrapper Method로 나뉜다. Filter Method에는 본 논문이 소개한 FINCO를 포함해, Relief, Focus, Las Vegas Filter가 있으며, Wrapper Method에는 Sequential Forward Selection(SFS), Sequential Backward Selection(SBS), Sequential Floating Forward

Selection(SFFS)이 있다. 본 논문이 소개하지 못한 Filter Method와 Wrapper Method들은 어떤 방법을 통해 변수를 선택하는지, 또 선택된 변수들은 어떤 것들인지 비교해 보는 연구도 이루어지길 바란다.

참 고 문 헌

김 부 성, 간경변증, 대한소화기학회 총서 6, 대한소화기학회, 군자출판사

김 영 선, 건강검진 자료에서 성장곡선을 이용한 간 질환 예측모형, 2002, 연세대학교 대학원 석사 학위연구

이 은 주, 만성 간질환 환자의 영양 상태, 1999, 연세대학교 생활환경 대학원 석사 학위연구

한 은 정, 건강검진 자료에서 Random Forests를 이용한 백내장 발생 위험군 예측 모형, 2004, 연세대학교 대학원 석사 학위연구

Avrim L. Blum, Pat Langley, Selection of Relevant Features and Examples in Machine Learning, *Artificial Intelligence*, 1997, 245-271

Edgar Acuna, A comparison of filters and wrappers methods for feature selection methods in supervised classification, *Proceedings of the Interface 2003 Computing Science and Statistics* vol 34, 2003

Hussein Almuallim, Thomas G. Dietterich, Efficient Algorithms for Identifying Relevant Features, *Proc. 9th Canadian Conf. on AI*, 1992, 38-45

Isabelle Guyon, Andre Elisseeff, An Introduction to Variable and Feature Selection, *Journal of Machine Learning Research 3*, 2003, 1157-1182

Jennifer G. Dy, Carla E. Brodley, Feature Selection for Unsupervised Learning, *Journal of Machine Learning Research 5*, 2004, 845-889

M. Dash, H. Liu, Feature Selection for Classification, *Intelligent Data Analysis 1*, 1997, 131-156

M. Dash, H. Liu, Consistency-based sezrch in feature selection, 2003, *Artificial Intelligence* vol 151, 155-176

Ron Kohavi, George H. John, Wrappers for Feature Subset Selection, *Artificial Intelligence* vol 97, 1996, 273-324

W.N. Venables, B.D. Ripley, Modern Applied Statistics With S-PLUS, Third Edition, Springer

ABSTRACT

Risk Factor Selection to Predict the Susceptible Group in the Development of Liver Cirrhosis Using Filter Methods

Kim, Jeong Sook
Dept. of Biostatistics and Computing
The Graduate School
Yonsei University

The National Statistical Office demonstrated that the mortality of 40's men is three times as much as women, and mentioned because it was mainly liver diseases. Actually the probability that more men die than women on liver diseases is 4.4 times. In this paper, to protect liver diseases for health care and screening test we predicted risk factors of screening positives of liver cirrhosis. We analyzed screening data which were collected from the people who had screening tests form 1994 to 2005.

We used prediction models for Liver Cirrhosis including Linear Discriminant Analysis, Quadratic Discriminant Analysis, Logistic Regression, and Classification and Regression Tree, to select risk factors of screening positives of liver cirrhosis. Also we used Forward and Inconsistency, one of the filter methods, whose main purpose was to reduce the dimension of data in the pre-processing step. Without filter methods, we selected the risk factors

as albumin, platelet, HBS, total protein, total cholesterol, LDH, and γ-GT, with filter methods we selected Alk. phos, HBS, MCH, weight, and age.

Key Words : liver disease, screening test, Liver Cirrhosis, Linear Discriminant, Quadratic Discriminant, Logistic Regression, Classification and Regression Tree, Forward and Inconsistency, filter methods