S-QUEST와 태아발육제한증 (IUGR) 조기진단시스템 개발

차경준† , 박문일‡ , 최항석* , 신영재*

요 약

방대한 양의 데이터에서 의사결정에 필요한 정보를 발견하는 일련의 과정을 데이터 마이닝 (data mining)이라고 하는데, 본 연구에서는 생물정보학 (bioinofmatics)의 한 분야로서 의학분야의 통계적 의사결정 시스템을 제공하는 의사결정나무 (decision tree) 알고리즘 중 QUEST를 S-PLUS로 구현하고(이하 S-QUEST) 발육제한 (Intrauterine Growth Restriction; IUGR) 데이터를 분석하였다.

주요용어 : 의사결정나무, QUEST, 발육제한 (IUGR)

1. 서론

21C는 biotechnology의 세기라는 말을 자주듣고 있다. 이와 더불어 생물정보학(bioinformatics)이라는 말이 함께 나오는데, 생물정보학은 genome에 대한 총체적인 연구를 하는 학문으로, 최근 생명과학분야의 연구 개발에 핵심적인 분야이다.

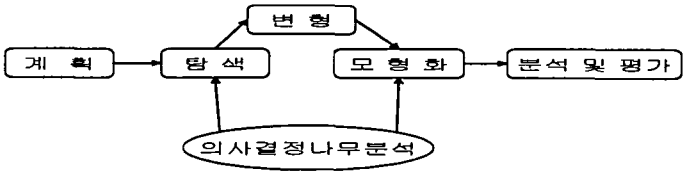
생물정보학의 특징은 데이터의 양이 방대하다는 것인데 데이터의 양만큼 데이터가 복잡한 성질을 담고 있다. 결국 생물정보학 연구를 위해서는 다양한 데이터의 처리가 필요한데 이를 위해서는 방대한 양의 정보를 데이터화 할 수 있는 컴퓨팅 환경이 필수적이다. 이렇게 다양한 형태의 데이터로부터 의사결정에 유용한 정보 및 지식을 발견하려는 일련의 데이터분석 및 모형선정과정을 데이터마이닝 이라고 한다.

데이터마이닝에서 의사결정을 내리기 위한 통계적 도구로서 의사결정나무 (decision tree)는 매우 유용하고 널리 사용되는 기법이다. 의사결정나무는 의사결정규칙 (decision rule)을 나무구조로 도표화하여 분류 (classfication)와 예측 (prediction)을 수행하는 분석방법이다. 또한 의사결정나무는 탐색과 모형화라는 두 가지 특성을 지니는데, 신경망 (neural networks) 분석 이전에 분석에 필요한 유의한 변수를 찾음으로 학습 시간의 단축을 초래할 수 있고, 그 자체가 분류 또는 예측모형으로 사용되기도 한다. 이렇게 데이터 마이닝의 한 도구로 의사결정나무가 사용된 일련의 구조는 <그림 1>과 같다.

의사결정나무에서 분리기준은 나무모형 형성에 가장 중요한 역할을 하게되는데, 1980년대 이후 지금까지 많은 알고리즘이 제안되었다. 그 가운데 잘 알려진 것으로 CHAID (Kass, 1980), CART (Breiman, Friedman, Olshen and Stone, 1984), C4.5 (Quinlan, 1993), QUEST (Loh and Shih, 1997) 등이 있다.

본 연구에서는 산과학분야 중에서 고위험임신의 한 예로서 발육제한 (Intrauterine Growth

† : 한양대학교 수학과 교수
‡ : 한양대학교 의과대학 교수
* : 한양대학교 대학원 수학과



<그림 1> 데이터마이닝 과정에서 의사결정나무의 분석구조

Restriction; IUGR)과 관련한 데이터를 바탕으로 가장 최근에 소개된 의사결정나무 알고리즘인 QUEST를 S-PLUS로 구현하고 데이터 분석을 하고자 한다.

본 논문의 구성은 다음과 같다. 제 2절에서는 산과학에서 발육제한태아에 대한 소개를 한다. 제 3절에서는 의사결정나무의 특징과 장, 단점 소개하고, 제 4절에서는 QUEST 알고리즘을 구현하고 시뮬레이션을 시행한다. 마지막으로 제 5절에서 결론 및 고찰을 기술한다.

2. 발육제한태아 (Intrauterine Growth Restriction; IUGR)

본 절에서는 산과학에서 고위험임신의 한 예로서 발육제한태아에 대한 소개를 한다.

발육제한태아란 태아가 임신 주수에 비해 작은 경우를 말하며, 그 주수의 10 %이하인 경우이다. 즉 임신 마지막달에 체중이 2.5kg이하인 경우를 발육제한태아라 한다.

발육제한에는 두가지 형태가 있는데, 임신 초기부터 어떠한 것의 영향을 받아 머리와 모든 장기가 작은 경우와 임신 말기에 영향을 받아 머리 크기는 정상이지만 다른 장기들은 작은 경우로 나눈다.

태아 성장에 미치는 요소로는 자궁 및 태반을 통한 혈류, 태아의 인슐린 (insulin), 그리고 포도당등이 알려져 있으며 (Vorherr H, 1982), 자궁 및 태반의 순환장애에 기인한 태아의 저산소증이 발육제한태아의 주원인으로 알려져 있다. 발육제한태아는 주산기 이환율 및 사망률이 높아 (Harvey, D., 1982) 조기식별 및 치료방침의 설정이 산과학 영역의 중요한 과제가 되어왔다 (Crawford CS, 1982).

지금까지 발육제한태아를 식별하는데에는 산모의 자궁고를 측정하거나, 초음파단층촬영에 의하여 태아의 체중을 측정하는 방법 등이 이용되었다 (Neilson, J.P., 1984; Geirsson, R.T., 1984).

이러한 태아의 안녕상태를 예견하는 중요한 지표로 태아의 심박동 변이도 (Fetal Heart Rate; FHR)의 분석은 최근 많이 이용이 되는 것으로 특히 태아 가사상태나 저산소증, 태아산혈증으로 인한 중추신경계의 기능저하가 있는 경우에는 심박동 변이도가 감소됨은 잘 알려져 있다.

3. 의사결정나무의 소개

본 절에서는 의사결정나무 및 알고리즘에 대하여 QUEST를 중심으로 소개하도록 한다.

의사결정나무는 나무그림을 의사결정문제에 적용시킨 것으로 의사결정을 이해하기 쉽게 설명할 수 있다. 의사결정나무는 Sonquist와 Morgan (1963)이 그 효시라고 할 수 있는데 AID (Automatic Interaction Detection)라는 컴퓨터 프로그램과 함께 의사결정나무의 역사가 시작된 것이다.

<표 1> 의사결정나무 분석 알고리즘의 비교

	CHAID	CART	C4.5	QUEST
목표변수	명목형, 순서형, 연속형	명목형, 순서형, 연속형	명목형, 연속형	명목형
예측변수	명목형, 순서형, 연속형	명목형, 순서형, 연속형	명목형, 연속형	명목형, 순서형, 연속형
분리기준	카이제곱-검정, F-검정	지니 계수, 분산의 감소	이득비율	카이제곱-검정, F-검정
분리개수	다지분리	이지분리	이지분리, 다지분리	이지분리

의사결정나무의 대중적인 이용은 Morgan과 Messenger (1973)의 THAID라는 알고리즘이 소개된 이후부터라고 할 수 있다. 그 후 Kass (1980)는 CHAID라는 알고리즘을 소개하게 되는데 카이제곱 적합성검정에 근거한 의사결정나무로써 현재까지 사용되고 있다.

의사결정나무는 누구나 이해할 수 있고 쉽게 설명되어질 수 있는 결과의 간결함으로 의학분야뿐 아니라 고객의 의사결정 패턴을 분석해야하는 상품개발, 마케팅부서, 그리고 문자, 지문인식 등을 연구하는 기계학습 이론분야에서 사용되고 연구되고 있다 (Frawley, 1991). 그밖에 누락된 관측 값에 대한 처리가 다른 모델보다 우수하고 변수들간의 교호작용 (interaction)의 설명과 처리가 용이하다는 장점이 있다. 반면에 의사결정나무에서는 선형 (linear) 또는 주효과 (main effect) 모형에서와 같은 결과를 얻을 수 없다는 한계점이 있다.

최근에 의사결정나무의 논리적인 구조가 인간의 사고구조와 유사하여 인공지능 (artificial intelligence) 분야에서도 이 모형을 연구하고 있다. Hunt, Marin과 Stone (1966)의 연구가 컴퓨터 과학분야에서의 효시라고 할 수 있고 Quinlan (1986, 1993)의 ID3와 C4.5는 대표적인 의사결정나무를 구현한 알고리즘이다. 현재 컴퓨터 과학에서는 인공지능 분야뿐 아니라 패턴인식 (pattern recognition), 기계학습 분야에서 활발히 연구되고 있다.

데이터마이닝에서 통계적 방법을 적용할 때 모수적 방법과 최근 컴퓨터의 발전과 더불어 활발히 연구되고 있는 비모수적 방법을 사용하고 있다. 의사결정나무 알고리즘에서 CART와 C4.5는 비모수적인 분석방법임에 반해 QUEST와 FACT (Loh and Vanichetakul, 1988)는 모수적인 접근 방법을 토대로 의사결정나무를 연구된 것으로 다양하게 변화하고 있는 데이터 형태에 따라 다양한 알고리즘을 적용할 수 있다.

QUEST (Quick, Unbiased, Efficient, Statistical Tree)는 이름에서 알 수 있듯이 빠르며 (Quick), 불편의 (Unbiased)인 유용한 통계적 나무 모형이다. 즉, 빠르게 변수선택 편의를 줄이기 위해 변수선택의 단계와 선택된 변수에 기초한 분리기준을 찾는 단계를 나누는 것이 특징이다.

변수선택은 ANOVA F-검정이나 분할표의 카이제곱검정에서 p-값에 대응되는 변수를 변수로 선택하고, 분리를 위하여 2-평균 군집분석 (two-means clustering)을 수행하여 두 개의 그룹을 만든 후 최적분리를 찾기 위하여 2차 판별분석 (quadratic discriminant analysis)을 수행하여, 목표변수를 가장 잘 분류하는 변수의 최적분리를 이용하여 자식마디를 형성한다.

QUEST는 이지분리 (binary split)를 수행하는 알고리즘으로, 변수선택 편의 (bias)나 계산시간을 줄이는 방법으로 관측치의 수가 많거나 복잡한 데이터에 대해서는 효율적이다.

다른 알고리즘과 변수선택 확률을 추정하여 편의와 선택력을 비교할 때, C4.5와 CART는 모두 변수선택에서 편의가 있는 것으로 나타났으며, CHAID는 편의가 덜하며, QUEST가 가장 안정된 결과를 보여주고 있었다 (송문섭, 윤영주, 2001). <표1>은 위에 언급된 내용을 요약한 것이다.

4. 실증분석

본 절에서는 발육제한태아 (IUGR) 데이터를 이용하여 의사결정을 내리는 시뮬레이션 과정을 소개할 것이다.

1995년 3월부터 2000년 12월까지 한양대학교 부속병원 산부인과를 내왕한 산모 중, 분만전 비수축 검사 (Nonstress Test; NST)를 20분 시행 받은 산모 6,589예 중에서 임신 30-35주에 해당되는 쌍태임신과 임신성고혈압 (PIH), 전치태반 (placenta previa), 태반조기박리 (abruptio placenta), 태아 곤란증 (fetal distress), 둔위 (breech), 기형 (fetal anomaly), 당뇨병 (GDM) 그리고 태아빈맥 (getal tachycardia) 등 비정상 요소를 제외한 정상태아 남녀 각각 100예와 목표변수 발육제한 (IUGR)이 있는 남녀 태아 각각 50예를 최종 연구대상으로 하였다. 설명변수로는 NST 측정 주수 (NST weeks(wks)), 평균 태아 심박동 (Mean FHR(bmp)), FHR 변이도 (Amplitude(bmp)), 분만 주수(Deliver weeks), 1분 Apgar 점수 (Apgar 1 min), Ponderal Index (g/㎝x100)와 특별히 근사 엔트로피 (Approximate Entropy)를 변수로 추가 하였다.

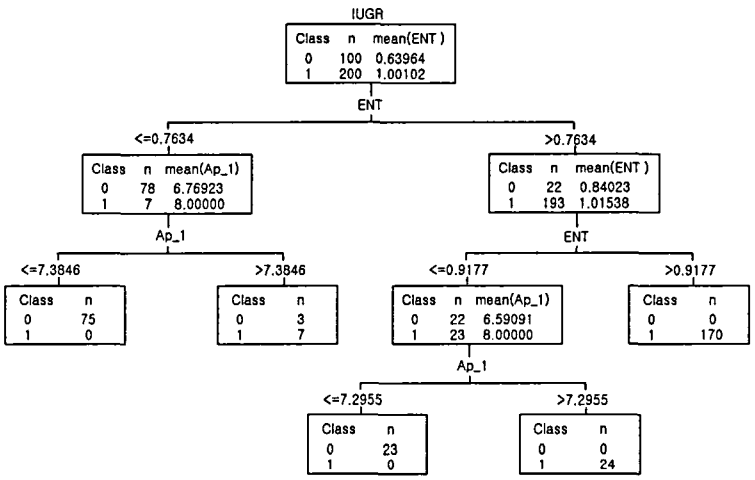
태아심박동과 제 변수 자료는 한양대학교 산부인과에서 개발된 FHR분석 시스템인 HYFM-I & II (1998)를 이용하여 수집 처리되었고, 의사결정 알고리즘인 QUEST는 S-PLUS를 사용하여 작성하였으며, 데이터의 통계분석은 SAS, version 8.2를 이용하였고 그룹간 유의성 검정은 Student's t-test를 적용하였다.

<표 2> 정상·발육제한 그룹간의 심박동에 관한 통계적 비선형적 지표

	Normal(N=200)	IUGR(N=100)	p-value
NST weeks(wks)	33.04±1.60	33.24±1.68	NS
Mean FHR(bpm)	144.15±5.24	143.26±6.24	NS
Amplitude(bpm)	17.23±3.40	16.52±5.02	NS
Delivery weeks(wks)	40.02±2.12	39.24±1.98	NS
Apgar 1 min	7.47±0.26	6.16±0.66	<.0001
Approximate Entropy	1.02±0.11	0.61±0.21	<.0001
Ponderal Index	2.70±0.39	2.57±0.31	NS

<표 3> S-QUEST 실행한 결과에 대한 오분류표

	S-plus Coding Program IUGR		
관 측	0 (IUGR)	1 (정상)	분류정확%
0 (IUGR)	97	3	97%
1 (정상)	0	200	100%
Total			99%



<그림 2> S-QUEST 결과의 나무모형

<표 2>는 각각의 설명변수에 대하여 정상태아군 ($N=200$)과 발육제한태아군 ($N=100$) 두 그룹간에 차이가 있는지 검정하였다. 그 결과, 1분 Apgar 점수 (Apgar 1 min)는 각각 7.47 (±0.26), 6.16 (±0.66)로서 발육제한태아보다 정상태아가 높게 나타났으며 ($p<0.0001$), 비선형적인 분석인 근사엔트로피 (Approximate Entropy)의 경우는 각 1.02 (±0.11), 0.61 (±0.21)로서 발육제한 태아보다 정상태아가 높게 나타났다 ($p<0.0001$).

<표 3>은 S-QUEST를 이용하여 발육제한 데이터를 분석한 결과에 대한 오분류표이다. 정상태아군 ($N=200$)에 대하여는 100%의 정확한 예측력을 보였으며, 발육제한태아군 ($N=100$)에 대하여도 97%라는 높은 예측 결과를 보였다.

<그림 2>은 S-QUEST를 이용한 결과를 나무모형으로 나타낸 것이다. <그림 2>에서 근사엔트로피는 태아의 발육제한을 설명하는 첫번째 설명변수의 역할을 하고, 다음으로 1분 Apgar 점수이다. 이것은 앞에서 살펴본 통계적 유의성 테스트 결과와 일치한다.

<그림 2>의 첫 번째 분리결과에 의하면 발육제한을 진단하기 위한 첫 번째 설명변수는 근사엔트로피이다. 근사 엔트로피가 0.7634이상이고 1분 Apgar 점수가 7.2955이상이면 태아는 정상인 태아로 판단되며, 근사 엔트로피가 0.9177이상이면 1분 Apgar 점수에 관계없이 항상 정상으로 판단 할 수 있다. 반면에 근사 엔트로피가 0.7634이하이고 1분 Apgar 점수가 7.3846이하이면 발육제한 태아로 판단된다.

한편, 근사 엔트로피가 0.7634이하이고 1분 Apgar 점수가 7.3846이상인 곳은 결과적으로 정상인 태아로 판단되었으나, 정상태아 7명 발육제한 태아 3명으로 표본의 수가 작기 때문에 쉽게 정상 태아로 판단을 내릴 수가 없다.

5. 결론 및 고찰

본 연구는 생물정보학의 한 분야로서, 정상태아와 발육제한태아 두 그룹간의 분류를 위하여 의사결정나무를 이용하여 다음과 같은 결론을 얻었다.

의사결정나무를 이용하여 출산 전에 태아의 발육제한을 진단하는 과정을 나무구조로 도표화하여 표현함으로서 통계적 지식이 부족한 의료진에게 보다 쉽게 결과를 이해시킬수 있고, 출산

전에 태아의 상태를 쉽게 판단할 수 있다. 출산 전에 태아의 건강상태를 빠르게 판단하므로 건강한 태아를 임신한 산모는 편안하게 출산 할 것이며, 발육제한 태아를 임신한 산모는 출산 전에 태아의 치료가 가능하게 되어 보다 근원적인 치료가 가능할 것이다.

데이터마이닝 기법 중에서 의사결정나무를 적용하여 사례 분석한 결과 정상 태아는 100%, IUGR 태아는 97%의 정확한 분류율을 보여 주었다. 이와 같은 결과는 생물통계학 분야에 의사결정나무를 적용하여 소기의 목적을 달성하였던 바, 이와 관련한 활발한 연구의 필요성을 갖는다.

현재 S-QUEST은 변수 선택과 변수의 분할 기준 값을 텍스트 형식으로 제시하는데, 향후에는 결과를 나무 모형으로 표현 할 수 있게 개발하여 좀더 쉽고 빠르게 결과를 해석할 수 있게 프로그램을 발전시킬 것이고 의사결정나무는 완전한 모형을 제시 할 수 없다는 단점을 보완 할 수 있는 방법론적 연구에 초점을 맞추어야 할 것으로 사료된다.

참고문헌

Breiman, L. et al.(1984), Classification and Regression Trees, Wadsworth, Belmont.

Crawford, C.S.(1982), The growth retarded newborn. In Management of the high risk fetus and neonate, Baltimore, Williams & Wilkins. 501.

Frawley, W.J. et al.(1991), Knowledge discovery in databases, AAAI Press/The MIT Press.

Geirsson, R.T. et al.(1984), Diagnosis of intrauterine growth retardation using ultrasound, Clin Obstet Gynecol, 11, 457-479.

Harvey, D. et al.(1982), Abilities of children who were small for gestational age babies, Pediatrics, 69, 296-300.

Hunt, E.B, et al.(1966), Experiments in induction, Academic Press.

Kass, G.V.(1980). An exploratory technique for investigating large quantities of categorical data, Applied Statistics, 29, 119-127.

Loh, W.Y. and Shih, Y.S.(1997), Split selection methods for classification tree, Statistica Sinica, 7, 815-840

Loh, W.Y. and Vanichsetakul, N.(1988), Tree-structured classification via generalized discriminant analysis, Journal of the American Statistical Association, 83, 715-728.

Morgan & Messenger (1973), THAID-a sequential analysis program for the analysis of nominal scale dependent variables, Survey Research Center, U of Michigan.

Neilson, J.P. et al.(1984), Screening for small for dates fetuses: A controlled trial. Br Med J, 289, 1179-82.

Quinlan, J.R.(1986), Induction of decision trees, in machine learning, 1, 81-106.

Quinlan, J.R.(1993), C4.5: Programs for machine learning. Morgan Kaufmann, Los Altos.

Sonquist, J. A. and Morgan, J. N. (1963). Problems in the analysis of survey data and a proposal, Journal of American Statistical Association, 58, 415-434.

Vorherr, H.(1982), Factors influencing fetal growth. Am J Obstet Gynecol, 142, 577-588.

송문섭, 윤영주(2001), 데이터마이닝 패키지에서 변수선택 편의에 관한 연구, 응용통계연구, 제14권 2호, 475-486