Journal of the Korean Data & Information Science Society
2014, **25**(6), 1253–1262

http://dx.doi.org/10.7465/jkdi.2014.25.6.1253
한국데이터정보과학회지

회귀나무 모형을 이용한 패널데이터 분석[†]

장영재[1]

[1]한국방송통신대학교 정보통계학과

접수 2014년 8월 10일, 수정 2014년 9월 12일, 게재확정 2014년 9월 18일

요 약

회귀나무 (regression tree)는 독립변수로 이루어진 공간을 재귀적으로 분할하고 해당 영역에서 종속변수의 최선의 예측값을 찾고자 하는 비모수적 방법론이다. 회귀나무 모형이 제안된 이래 로지스틱 회귀나무모형이나 분위수 회귀나무모형과 같이 유연하고 다양한 모형적합을 위한 연구가 진행되어 왔다. 최근에 들어서는 Sela와 Simonoff (2012)의 RE-EM 알고리즘, Loh와 Zheng (2013)의 GUIDE 등 패널데이터와 관련하여 진일보한 나무모형 알고리즘도 제안되었다. 본 논문에서는 각 알고리즘을 소개하고 특징을 살펴보는 한편, 실험 데이터를 생성하여 평균제곱오차 (mean squared error)를 바탕으로 예측력을 비교하였다. 분석결과, RE-EM 알고리즘의 예측력이 상대적으로 우수하게 나타났다. 이 알고리즘을 통해 기업경기실사지수 업종별 패널자료를 분석한 결과 최근의 업황에 가장 큰 영향을 미치는 요소는 매출 실적으로 나타났으며 매출 상위 그룹의 경우 비제조업이 제조업에 비해 업황에 대한 판단이 긍정적인 것으로 나타났다.

주요용어: 기업경기실사지수, 패널데이터, 혼합모형, 회귀나무.

1. 서론

1.1. 회귀나무의 정의

회귀나무 (regression tree)는 독립변수로 이루어진 공간을 재귀적으로 분할하고 해당 영역에서 종속변수의 최선의 예측값을 찾고자 하는 비모수적 방법론이다. 나무구조를 형성할 때 변수선택 과정은 독립변수 공간 분할과 직결되는 요소이므로 매우 중요하다고 할 수 있다. 회귀나무를 형성하기 위해서는 적절한 변수를 선택한 뒤 이를 기준으로 데이터를 나누고 나뉘어진 각 독립변수 공간에서 종속변수와 독립변수와의 관계성을 찾아가는 과정을 거치게 된다. 따라서 이러한 변수선택 절차가 회귀나무의 예측력에 큰 영향을 미칠 수 있다.

Figure (1.1)은 반복 이분할 (binary recursive partitioning) 과정을 통해 구현된 회귀나무를 나타내고 있다. 그림에서 원으로 표시된 것이 노드 (node)라고 불리는 것으로서 일정 기준에 따라 분할된 독립변수 공간으로 생각할 수 있다. 즉, 분할의 기준이 되는 변수에 따라 데이터가 분류되는 곳이다. 실선으로 표시된 것이 가지 (branch)라고 불리는 것으로서 각 단계에서 조건에 따라 하위 노드로 데이터를 나누어 가는 과정을 나타낸다. 이러한 분할 과정이 반복적으로 실행되면서 전체 나무의 모습을 이루게 된다. 최상위 노드는 모든 학습샘플 (learning sample)을 포함하고 있으며, 각 분기점이 되는 노드에서 분기변수 (split variable)로 선택된 설명변수 (x_1, x_2, x_3,$\cdots$)의 값에 따라 분기가 반복되는 단계를 거

[†] 이 논문은 2013년도 한국방송통신대학교 학술연구비 지원을 받아 작성된 것임.
[1] (333-749) 서울특별시 종로구 대학로 86, 한국방송통신대학교 정보통계학과, 조교수.
E-mail: yjchang@knou.ac.kr

쳐 최종 노드 (N_4, N_5, N_6, N_8, N_9)에 이르게 된다. 최종 노드 아래에 표시된 값은 회귀나무 모형에 따른 예측값이다. 만약 가장 기본적인 상수항 (constant) 회귀나무 모형을 적합했다면 예측값은 각 최종 노드에 위치한 표본의 평균값 ($\bar{y}_4$, $\bar{y}_5$, $\bar{y}_6$, $\bar{y}_8$, $\bar{y}_9$)이 된다. 즉, 각 노드에 위치한 종속변수들의 평균치를 예측값으로 하는 상수항 모형이다. 상수항 (constant) 회귀나무의 경우 설명변수는 데이터를 적절하게 분할하는 역할을 하면서 최종노드의 예측값 산출에 직접적으로 영향을 미치지는 않는다고 할 수 있다.

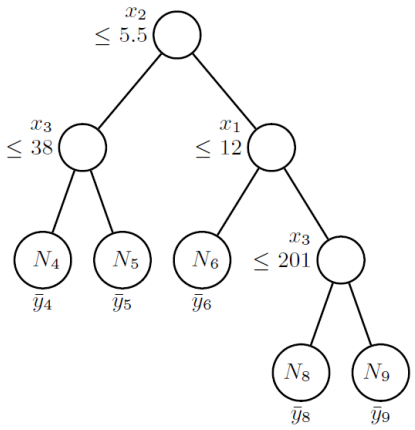

Figure 1.1 Regression tree: At each intermediate node, an observation goes to the left branch if and only if the condition is satisfied.

1.2. 회귀나무의 확장

앞서 살펴 본 기본적인 회귀나무 모형이 제안된 이래 더욱 유연하고 다양한 모형적합을 위한 연구가 진행되어 왔다. Loh (2002)는 선택편의 (selection bias) 문제와 교호작용을 고려하는 회귀나무 모형을 제안하였으며 이를 로지스틱 회귀모형이나 분위수 회귀모형 등 다양한 형태의 조각별 선형 회귀나무 모형 (piecewise linear regression tree)으로 확장하였다. 즉, 상수형 (constant) 회귀나무에서는 최종 노드의 예측값을 상수 모형, 즉 종속변수의 평균값으로 산출하였는데 조각별 선형 회귀나무에서는 각 최종 노드별로 선형모형을 적합함으로써 예측값을 산출하는 것이다. 대체로 나무모형의 최종 노드에서 적합하는 모형이 복잡한 형태를 지닐수록 나무의 크기는 작아지게 마련이며 모형이 단순할수록 나무의 크기는 커지는 경향이 있다. Chang과 Kim (2011)은 이러한 조각별 선형 회귀나무 모형의 특징을 곡률검정 (curvature test) 및 이에 기반한 비선형 모형의 적합 등으로 꼽으면서 확장가능성을 보였다.

이렇듯 다양한 모형을 접목하는 연구가 시도되는 한편 고차원의 데이터에 나무모형을 적합하려는 연구도 진행되었다. 다중반응 변수 (multiresponse variables)로 이루어진 데이터나 시간을 흐름을 고려한 시계열 자료와 관련된 여러 가지 나무모형 방법론들도 개발되었다. Dzeroski와 Zenko (2004)는 분류 (classification)의 문제와 관련하여 다중반응변수 분석 나무모형의 우수성을 보였다. 기존의 방법에 비해 교차타당화 (cross validation)을 적용한 앙상블 방법이 다중반응 변수 분류예측력에 있어서 더욱 우수한 것으로 나타났다. Meek 등 (2002)은 시계열 자료 분석을 위해 자기회귀 나무모형 (autoregressive tree models)을 제안하였으며, Cappelli와 Iorio (2010)와 Rea 등 (2010)은 시계열자료에 있어서 구조변화나 국면전환을 탐지하거나 변화 시점을 찾아내는 회귀나무 방법론을 제안하였다.

최근에 들어서는 패널데이터와 관련하여 진일보한 나무모형 알고리즘도 제안되었다. 패널 (panel)연구는 종단연구 (longitudinal study)의 특별한 한 가지 형태라고 할 수 있다. 종단연구는 연구의 대상 집단을 어느 시점에서 표집하여 이들을 대상으로 오랜 기간에 걸쳐 반복적으로 관찰함으로써 시간의 흐

름에 따라 각종 변인들의 변화 상태를 파악하는 연구방법이다. 관측치의 특성이 시간의 흐름에 따라 변화하는 것을 추적하고 관찰하는 연구 방법으로서 관측 대상의 범위에 따라 특정 집단의 특성을 시간의 흐름에 따라 분석하는 코호트 연구와 동일한 개체의 특성 변화를 시간의 경과에 따라 분석하는 패널연구로 구분 할 수 있다. 본 논문에서는 나무모형의 확장이라는 방법론적 측면과 응용방법에 초점을 맞추고 있으므로 패널 데이터와 종단자료라는 두 가지 개념을 유사한 의미로 사용하기로 한다.

Segal (1992)과 Zhang (1998)은 Breiman 등 (1984)이 제안한 CART (classification and regression tree) 알고리즘을 확장하여 종단자료와 다중 반응변수 자료 등에 적합 가능한 분류나무 (classification trees) 알고리즘을 제안하였다. 이상의 방법론은 다중 반응변수와 종단자료에 나무모형을 적용을 선도한 알고리즘이라고 할 수 있으나 공분산 행렬 등의 계산의 어려움이라는 한계를 지니고 있었다 (Loh와 Zheng, 2013). De'ath (2002)는 이러한 계산상 제약을 해결하기 위해 CART 알고리즘을 이용하였다. 즉, CART에서 추정과 분기의 주요 판단 지표로 사용되는 표본 평균과 불순도 (impurity)의 개념을 차용하여 다변량으로 확장한 뒤 이를 이용함으로써 공분산 행렬 계산을 대체할 수 있었다. Lee (2005)는 종속변수를 일반화하여 종단자료에 관한 나무모형의 적합 방법을 제안하였다. 이 방법이 과거의 나무모형과 차별성을 보이는 것은 각 노드에서 최대우도 (maximum likelihood)를 이용하여 모수를 추정하고 잔차를 고려한 분기 (split)를 시도하였다는 점이다. 최근 들어서는 패널데이터의 회귀분석에 초점을 맞춘 나무모형이 제안되었다. Sela와 Simonoff (2012)는 관측치 간의 차이점을 임의효과 (random effects)로 간주하고 이를 주 효과인 고정효과 (fixed effects)와 함께 고려하는 혼합모형을 구축한 뒤 나무모형과 접목하여 추정하는 방법을 제안하였다. 추정을 위하여는 EM 알고리즘을 사용하였으며 이러한 점에서 RE-EM 나무모형이라 명명하였다. Loh와 Zheng (2013)은 CART의 선택편의 등 문제를 개선한 GUIDE 회귀나무 모형을 확장하여 다변량 자료와 종단자료에 적합이 가능한 방법론을 제안하였다.

본 논문에서는 최근 제안된 알고리즘 중 패널데이터 적합을 위한 회귀나무 모형 알고리즘을 소개하고 각각의 특징을 살펴보았다. 비교의 기준으로는 다변량 회귀나무 (multivariate regression tree; MRT)를 설정하였으며 시뮬레이션 모형을 통해 생성된 데이터를 바탕으로 혼합모형의 나무구조 적합 방법론인 RE-EM 나무모형, 곡률검정을 통한 변수선택과 불편성을 고려한 GUIDE 등의 예측력을 비교하였다. 본 논문의 구성은 다음과 같다. 2절에서는 비교 대상 분석방법인 회귀나무 알고리즘에 관하여 살펴보았다. 3절에서는 시뮬레이션을 통해 생성된 데이터를 이용하여 각 회귀나무의 예측력을 비교하고 평가해 보았으며 RE-EM 알고리즘을 실제 데이터인 한국은행의 기업경기실사지수 자료에 적용해 보았다. 마지막으로 4절에서는 본 논문의 내용에 대하여 간략히 정리하고 향후 연구방향에 대해 살펴보았다.

2. 종단자료 적합을 위한 회귀나무 알고리즘

2절에서는 비교 대상 분석방법인 회귀나무 알고리즘에 관하여 살펴보기로 한다. 혼합모형을 도입한 RE-EM 알고리즘과 패널데이터 분석을 위한 GUIDE 모형의 확장 알고리즘을 간략히 정리하면서 각 방법론의 특징을 살펴보도록 한다.

2.1. RE-EM 알고리즘

Sela와 Simonoff (2012)는 관측치 간의 차이점을 임의효과 (random effects)로 간주하고 이를 주 효과인 고정효과 (fixed effects)와 함께 고려하는 혼합모형을 구축한 뒤 나무모형과 접목하여 추정하는 방법을 제안하였다. 우선, $t = 1, \cdots, T_i$ 각 시점에서의 $i = 1, \cdots, I$와 같은 개별 개체의 관측값으로 이루

어진 패널데이터를 얻었다고 가정하고 다음과 같은 모형을 고려하자.

$$y_{it} = Z_{it}b_i + f(x_{it1}, \cdots, x_{itK}) + \epsilon_{it} \tag{2.1}$$

$$(\epsilon_{i1}, \cdots, \epsilon_{iT_i})' \sim N(0, R_i) \tag{2.2}$$

$$b_i \sim N(0, D) \tag{2.3}$$

여기서 R_i는 대각행렬이 아닌 임의의 양정치 행렬을 의미하며 f는 모수에 관해 선형인 함수를 의미한다. 만약 b_i가 고정된 형태로 속성값들과 상관되어 있다고 가정하면 선형고정효과모형 (linear fixed effects model)이 된다. 반면, Jo와 Chang (2013)에서처럼 b_i가 랜덤인 형태로 속성값들과 상관되어 있지 않다고 가정하면 혼합효과모형 (mixed effects model)의 형태가 된다. 이러한 모형을 바탕으로 Sela와 Simonoff (2012)에서 제시된 RE-EM 나무의 추정방법을 간략하게 소개하면 다음과 같다.

1. 임의효과와 관련된 $\hat{b_i}$ 값을 0으로 놓는다.
2. 추정된 임의효과 $\hat{b_i}$가 수렴할 때까지 다음과 같은 과정을 반복한다. 여기서 수렴의 기준은 우도함수 (likelihood function)나 제한된 우도함수 (restricted likelihood function) 값의 변화가 사전에 지정한 일정한 기준치보다 작을 경우로 정한다.
   (a) $t = 1, \cdots, T_i$ 각 시점에서의 $i = 1, \cdots, I$와 같은 개별 개체에 대해, 목표변수 $y_{it} - Z_{it}\hat{b_i}$와 속성 $x_{it} = (x_{it1}, \cdots, x_{itK})$을 바탕으로 f를 근사 (approximating)하는 회귀나무를 추정한다. 추정된 회귀나무를 이용하여 지시변수 (indicator variable), $I(x_{it} \in g_p)$를 생성한다. 여기서 g_p는 회귀나무의 모든 최종노드 (terminal nodes)를 나타내는 영역이다.
   (b) 선형혼합모형 $y_{it} = Z_{it}b_i + I(x_{it} \in g_p)\mu_p + \epsilon_{it}$를 적합시킨다. 그리고 추정된 모형으로부터 $\hat{b_i}$을 추출한다.
3. 각 회귀나무의 최종노드에서의 예측 반응값을 선형 혼합모형 적합과정을 통해 추정된 모수 수준의 예측 반응값 μ_p로 대체한다.

RE-EM알고리즘은 R에서 REEMtree라는 패키지를 통하여 구현할 수 있다.

2.2. GUIDE의 확장

Loh와 Zheng (2013)은 다중변수 데이터와 관련하여 개별적인 다중회귀 모형을 각각의 종속변수에 적합할 경우, 최종적인 나무모형 결과의 해석 등에 있어서 어려움을 초래하는 문제점이 있음을 지적하였다. 예를 들어 서로 다른 세 가지 종속변수와 관련된 개별적인 세 개의 나무모형이 적합되었다고 가정해 보자. 이 경우, 각 나무를 이루는 설명변수들인 분기변수가 개별 나무모형을 넘어서 다른 나무모형의 분기변수와 함께 작용하는 교호작용 (interaction)이나 변수 간 결합으로 인한 영향 등을 고려할 수 없는 한계점이 존재한다. 만약 이러한 개별적인 나무모형 대신, 세 개의 종속변수를 동시에 추정하는 하나의 나무모형을 적합한다면 이러한 문제점을 보완할 수 있다는 것이다. 더불어 GUIDE의 분기변수 선택 방법의 특징으로 꼽을 수 있는 불편성 (unbiasedness)에 기인하여 더욱 정확한 예측이 가능하다는 것이었다. 다중반응 회귀나무모형에서 변수를 선택하는 과정은 잔차벡터의 부호를 고려하여 분할표 (contingency table)를 만들고 카이제곱 (chi-squared) 통계량을 기준으로 선택하는 과정을 거치게 되는데, 이는 단변량 GUIDE 회귀나무의 곡률검정에서 사용하였던 방법과 유사한 것이다. Loh와 Zheng (2013)에 제시된 GUIDE 다중반응 회귀모형의 변수선택 알고리즘을 간략하게 살펴보면 다음과 같다.

1. 임의의 노드 t에서 반응변수 또는 종속변수의 평균벡터 $\bar{y} = (\bar{y}_1, \cdots, \bar{y}_d)$를 계산한다.

2. 부호벡터 $Z = (Z_1, \cdots, Z_d)$를 계산한다. 여기서 부호벡터는 반응벡터의 k번째 원소가 $Y_k > \bar{y}_k$를 만족할 때 $Z_k = 1$로, $Y_k \leq \bar{y}_k$를 만족할 때 $Z_k = -1$로 설정한다.
3. 각 설명변수 X에 대하여 다음과 같이 주효과 검정 (main effects test)을 실시한다.
   (a) X가 수치형 변수일 경우에는 정해진 기준에 따라 m개의 구간으로 나눈다. 데이터 개수가 일정 기준보다 적을 경우에는 세 구간, 그렇지 않을 경우에는 네 구간으로 나누도록 한다.
   (b) X가 범주형 변수일 경우에는 각 범주가 그룹을 형성하게 된다.
   (c) 결측치가 있을 경우에는 별도의 범주를 생성하여 이 범주로 분류한다.
   (d) Z의 2^d (Z는 -1이나 $+1$ 둘 중 한 값을 가지며 d는 관측시점 전 범위를 적당히 나눈 구간의 개수이므로 Z의 부호의 곱집합의 개수는 2^d로 산출)개의 패턴을 열로 하고 X의 그룹들을 행으로 하는 분할표 (contingency table)를 작성한 뒤, 카이제곱 (chi-squared) 독립성 검정을 위한 유의확률 (p-value)을 산출한다.
4. 만약 산출된 가장 작은 유의확률 값이 $0.05/d$보다 작다면, 이와 관련된 설명변수 X를 선택한다.
5. 그렇지 않다면 X_i, X_j의 쌍으로 이루어진 변수들에 관해 교호작용 검정 (interaction test)을 실시한다.
6. 만약 교호작용 검정을 실시하여 산출된 가장 작은 유의확률 값이 $0.05/d(d-1)$보다 작다면 해당하는 변수쌍을 선택한다.
7. 교호작용 검정 결과, 위의 기준을 만족시키지 못할 경우에는 이전 단계인 4단계에서 고려한 가장 작은 유의확률 값을 산출하게 되는 설명변수 X를 선택한다.

GUIDE의 다중반응 회귀나무 모형은 각 관측시점에서의 측정되는 반응변수를 서로 다른 차원의 반응변수로 간주함으로써 패널데이터에 적합 가능한 형태로 응용할 수 있다. 앞서 소개한 알고리즘의 (1)단계의 다중반응변수 대신 i번째 개체의 j번째 관측시점 u_{ij}에서의 종속변수 Y_{ij}를 고려한다. 그리고 u_{ij}의 전 범위를 적당한 d개의 구간으로 나누고 (u_{ij},Y_{ij})로 이루어진 데이터에 LOWESS (Cleveland, 1979) 방법을 적용하여 평활화된 곡선으로 Y_{ij}의 평균수준을 추정한다. (2)단계에서는 부호벡터 Z를 정의할 때 평균벡터 대신 평활화된 곡선을 기준으로 비교를 한다. 나뉘어진 d 개의 각 구간에서 곡선보다 위쪽에 위치한 관측치가 많으면 $Z_k = +1$로, 그렇지 않으면 $Z_k = -1$로 설정한 뒤, (3)단계 이후는 기존 알고리즘과 동일한 과정을 거치게 된다. 실제 데이터의 예를 들어 보자. 만약 1주차부터 9주차까지 환자들에 대한 처치 (treatment)와 이에 관한 관측변수 값이 지속적으로 측정된다면, 9개의 원소로 이루어진 9차원 반응변수 벡터로 간주하고 각각의 측정시점 변수와 짝을 이루도록 데이터를 가공한다. 개별 개체들의 데이터는 성별이나 나이, 처치 등 설명변수와 함께 측정 시점 및 반응변수 벡터로 구성된다.

3. 실증분석

지도학습 (supervised learning)의 목적은 예측오차를 최소화하는 모형의 구축에 있다. 본 논문에서는 식 (3.1)과 같이 각 알고리즘의 예측치와 실제치의 차이인 오차의 제곱을 평균하여 평균제곱오차 (mean squared error)를 구한 뒤 이를 비교하였다. 즉, 평균제곱오차는 각 알고리즘의 오차를 측정한 것으로서 예측력을 비교하기 위한 도구라고 볼 수 있다. 여기서 m은 각각의 개별 데이터 포인트 총 개수이며 d는 총 관측시점 개수를 의미한다. 또한 p개의 독립변수가 있다고 가정하였으며 $\hat{f}$는 회귀나무

모형에 의한 추정값이다. 따라서 높은 MSE 값을 나타내는 알고리즘은 상대적으로 예측력이 낮은 것이라 할 수 있다.

$$MSE = \frac{\sum_{i=1}^{m}\sum_{j=1}^{d}(\hat{f}_j(x_{i1}\cdots,x_{ip}) - E(y_j|x_{i1}\cdots,x_{ip}))^2}{md} \tag{3.1}$$

3.1. 시뮬레이션 모형 I

Loh와 Zheng (2013)을 참고하여 세 가지 종류의 패널데이터 D1, D2, D3를 생성한 뒤 각 데이터를 이용하여 알고리즘의 예측력을 평가하였다. 아래의 모형을 사용하여 데이터를 생성하되, 모형에 사용된 균일분포 $U(-1,1)$를 따르는 X_1, X_2 외에 3~4개의 무상관 변수를 추가하였다.

$$Y_u = 1 + X_{1u} + X_{2u} + 2X_{1u}X_{2u} + 0.5u + b_0 + b_1u + \epsilon_u \tag{3.2}$$

여기서, $u = 1, 2, \ldots, d$인 관측시점을 의미하며 $b_0 \sim N(0, 0.5^2)$이고 $b_1 \sim N(0, 0.25^2)$인 임의효과(random effects)이다. ϵ_u는 오차항으로서 $N(0, 0.25^2)$을 따른다. 한편 각 데이터 별로 추가된 무상관 변수는 다음과 같다.

(D1) $X_3, \cdots, X_5$는 $U(-1,1)$을 따르는 변수

(D2) $X_3, \cdots, X_5$는 $U(-1,1)$을 따르는 수치형 변수이며 X_6는 5개의 범주를 갖는 범주형 변수

(D3) $X_3, \cdots, X_5$는 $U(-1,1)$을 따르는 수치형 변수이며 X_6는 10개의 범주를 갖는 범주형 변수

예측력 평가를 위해서는 $(-1,1)$구간을 균등하게 6개의 구간으로 나누고 각 구간 내에서 균일분포를 따르는 랜덤변수를 생성한 뒤 이를 조합하여 생성되는 총 7776개 (6^5) 데이터 포인트로 이루어진 데이터를 생성하였다. 각각의 데이터 포인트에 대해 관측시점 u를 감안하여 식 3.2에 따라 실제 기댓값을 산출하였으며 동일한 데이터를 사용하여 각 알고리즘으로 추정된 모형의 예측값도 산출하였다. 각 알고리즘의 예측력을 비교하기 위해서 식 3.1을 사용하였으며 이러한 일련의 과정은 100회 반복되었다. 식 3.2과 데이터 (D1), (D2), (D3)를 이용하여 각 알고리즘의 예측력을 비교한 결과는 Table 3.1과 같다. 해당 알고리즘과 데이터 별로 100회의 시뮬레이션을 통해 산출된 각각의 MSE값의 평균치 및 표준오차를 제시하였다.

Table 3.1 MSE's of MRT, RE-EM and GUIDE (standard errors in parentheses)

Data	MRT	RE-EM	GUIDE
D1	0.5194	0.2198	0.4227
	(0.002)	(0.001)	(0.02)
D2	0.5185	0.2382	0.4241
	(0.003)	(0.001)	(0.02)
D3	0.5193	0.2629	0.4242
	(0.002)	(0.004)	(0.01)

시뮬레이션 결과, RE-EM 알고리즘의 예측력이 상대적으로 우수하게 나타났다. 생성된 데이터가 랜덤효과를 감안한 것이므로 RE-EM 알고리즘 적용이 가장 타당하다고 할 수 있다. 다만, 데이터에 범주형 변수가 추가되고 범주가 증가할수록 MSE가 커지는 것으로 나타났다. 한편 GUIDE의 예측력은 RE-EM 방법보다 좋지 않지만, MRT보다는 다소 좋은 것으로 나타났다. 또한, GUIDE의 경우 MSE 평균치의 표준오차가 상대적으로 큰 것으로 나타났는데, 이것도 선형 랜덤효과를 감안한 모형의 특성에 기인한다. MRT의 경우에는 종단자료를 시간의 흐름을 고려하여 적합할 수 없다는 근본적인 한계점 때문에 모든 알고리즘 중 가장 예측력이 좋지 않은 것으로 나타났다.

3.2. 시뮬레이션 모형 II

시뮬레이션 식 3.2을 통해 생성한 데이터는 관측시점에 따라 변하는 시변 (time-varying) 독립변수들을 포함하고 있었다. 관측시점에 상관없이 한 개체에 관하여는 모든 독립변수들이 고정 (fixed)되어 있는 데이터 즉, 시간과 독립인 예측변수 (time independent predictors)들로 구성된 데이터를 생성하여 각 알고리즘의 예측력을 비교해 보았다. R의 mmm package 내의 데이터 생성방법을 그대로 인용하여 다변량 정규분포로부터 4개의 관측시점에 따른 반응변수와 예측변수를 생성하였다.

$$Y \sim N(0, \Sigma) \tag{3.3}$$

여기서, 반응변수와 예측변수 분산은 각각 1,1, 4.0으로 가정하였다. 모형을 통해 생성된 데이터 (D4)는 각 개체번호 (ID), 반응변수, 독립변수, 관측시점, 독립변수와 관측시점의 교호작용 등 5개의 변수를 포함하고 있다. 1000개의 학습샘플 (learning sample)을 생성하였으며 RE-EM과 GUIDE를 통해 모형을 적합한 뒤 1000개의 시험샘플 (testing sample)로서 각각의 예측력을 평가하였다. 앞선 예와 마찬가지로 100회의 시뮬레이션을 통한 평균제곱오차의 평균치를 산출하여 예측력을 비교하였다.

Table 3.2 MSE's of RE-EM and GUIDE based on model (3.3) (standard errors in parentheses)

Data	RE-EM	GUIDE
D4	1.04	1.05
	(0.005)	(0.005)

시뮬레이션 결과, 시변 (time-varying) 예측변수가 포함되어 있었던 경우와는 달리 RE-EM 알고리즘과 GUIDE의 예측력이 큰 차이점을 보이지 않았다. 이는 관측시점에 상관없이 각 개체별로 고정인 예측변수를 바탕으로 하고 종속변수의 평활화를 통해 예측치를 산출하는 GUIDE의 알고리즘 특징에 기인한다. 또한 데이터 (D4)에는 식 (3.2)와 같이 명시적인 시변 랜덤효과가 반영되지 않았기 때문에 랜덤효과 모형 적합에 유리한 RE-EM 알고리즘의 상대적인 장점이 부각되지 않았다.

3.3. 업종패널 기업경기실사지수 자료 분석

본 논문에서는 시변 (time-varying) 독립변수가 포함된 패널자료 적합에 가장 적절하고 예측력도 좋은 것으로 나타난 RE-EM 알고리즘을 통해 기업경기실사지수 업종패널자료를 분석해 보았다. 기업경기실사지수 (business survey index; BSI)란 기업의 총체적인 경제활동에 대한 심리상태를 경제 전반에 걸쳐 계량하여 지수화한 것으로서 동 지수가 100보다 크면 조사항목에 대해 긍정적으로 생각하는 업체가 많은 것으로, 100보다 작으면 반대로 조사항목에 대해 부정적으로 생각하는 업체가 많은 것으로 해석하게 된다. 본 논문에서는 한국은행에서 발표한 BSI 자료 중 2012년 7월부터 2014년 6월까지 총 37개 업종의 업황 BSI를 종속변수로, 매출 BSI (sales), 채산성 BSI (profit), 인력사정 BSI (manpower), 자금사정 BSI (financial), 제조업 여부 (manu), 시간 (t)을 독립변수로 하고 업종별로 내재한 차이를 임의효과로 반영하여 패널 모형을 적합하였다. 2012년 7월부터의 자료를 쓴 이유는 최근 한국은행의 표본개편에 따른 새로운 표본을 바탕으로 한 지수 편제가 2013년 7월부터 이루어졌기 때문이다.

적합된 나무모형은 Figure 3.1과 같다. 전반적으로 최종 노드의 수치가 100을 하회하는 것으로 보아 최근 업황에 대한 평가는 다소 부정적임을 알 수 있다. 회귀나무의 첫번째 가지 분기변수가 매출로 나타난 것은 기업의 업황을 판단하는 중요한 실적지수가 매출인 것을 의미한다. 반면 독립변수 중 인력사정 지수만이 나무구조에 나타나지 않음으로써 기업측면에서 바라 보았을 때 인력사정은 업황에 크게 영향을 주지 못하는 것으로 분석되었다. 매출실적을 상대적으로 좋지 않은 그룹에서는 채산성이 높고 자금사정이 좋을수록 상대적으로 업황을 좋게 평가하는 것으로 나타났다. 오른쪽으로 분기한 가지를 따라가면, 매출에 대한 평가가 가장 좋은 그룹을 만나게 되는데, 이 경우 비제조업의 업황에 대한 평가 (93.44)가 제조업 (85.46)보다 더 좋은 것으로 분석되었다.

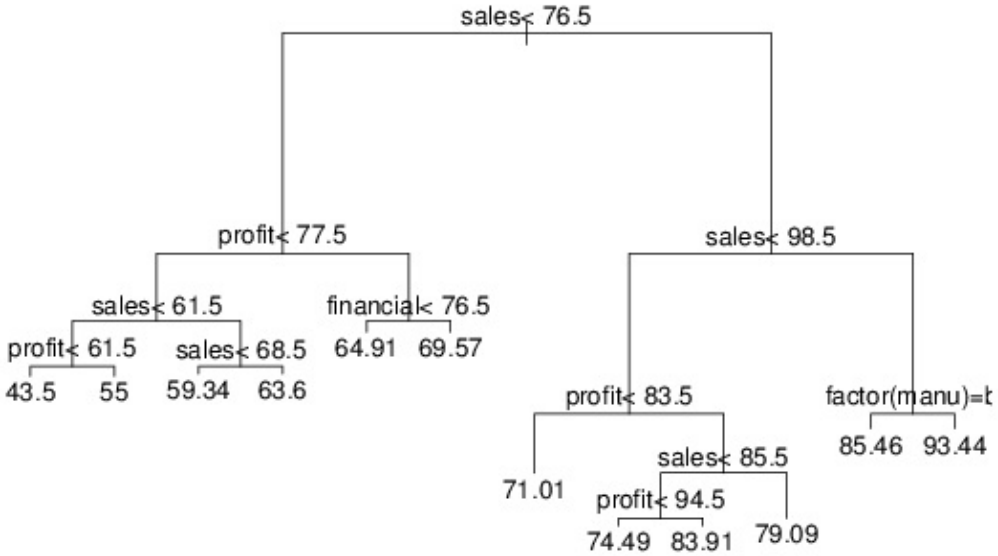

Figure 3.1 RE-EM tree model for the analysis of BSI

4. 결론 및 향후 연구방향

Sela와 Simonoff (2012)는 관측치 간의 차이점을 임의효과 (random effects)로 간주하고 이를 주 효과인 고정효과 (fixed effects)와 함께 고려하는 혼합모형을 구축한 뒤 나무모형과 접목하여 추정하는 방법을 제안하였다. 전반적으로 상당히 우수한 적합 결과를 나타내었으며 이는 랜덤 효과를 고려한 데 기인한다. GUIDE의 경우, 단변량 종속변수를 지닌 회귀나무의 경우 우수한 예측력을 나타내는 것으로 알려져 있으나 패널데이터 분석에 있어서는 상대적으로 예측력이 좋지 않은 것으로 나타났다. 다중반응변수 데이터를 적합하는 알고리즘을 변형하여 종단자료를 분석하는 방법을 사용하므로 시변 독립변수를 적절하게 반영하지 못하는 한계점이 있기 때문이다. 다만, 시간과 독립인 예측변수 (time independent predictors)들로 구성된 데이터를 생성하여 분석한 결과 그 예측력이 RE-EM 알고리즘의 경우와 크게 차이가 나지는 않는 것으로 나타났다.

비교결과를 바탕으로 시변독립변수를 지닌 패널 자료 적합에 적절한 RE-EM 알고리즘을 선택하여 업종패널 BSI 자료를 분석해 보았다. 기업의 업황을 판단하는 중요한 실적지수 매출인 것으로 나타난 반면, 기업측면에서 바라 보았을 때 인력사정은 업황에 크게 영향을 주지 못하는 것으로 분석되었다. 매출이 가장 좋은 그룹을 세부적으로 분해해 보면, 비제조업의 경우가 상대적으로 업황에 대한 평가가 좋은 것으로 나타났다.

이상과 같이 시뮬레이션과 실제 데이터를 바탕으로 패널데이터 분석을 위한 회귀나무 알고리즘을 살펴보았다. 이들은 좋은 예측력과 분석결과를 나타내긴 하지만 다음과 같은 면을 고려하여 개선하면 예측력도 향상되면서, 보다 확장된 데이터에 응용가능한 알고리즘으로 발전시킬 수 있을 것이다. 한 가지 방향은 다중반응변수를 지닌 패널데이터의 나무구조 모형 적합에 관한 것이다. 종속변수들간 상관관계를 고려하면서 나무모형 알고리즘을 응용할 수 있을 것이다. 예를 들면, Charbonneau (2014)가 제안한 이항 (binary) 반응변수 다중 고정효과 패널데이터 모형의 나무구조화 등이다. 이와 함께 다중변수의 구조를 반영하는 새로운 종속변수를 생성하여 회귀나무간의 결합을 고려하는 앙상블 형태의 회귀나무 방법의 개발도 패널데이터 적합모형의 예측력 향상을 위해 필요한 후속연구라 하겠다.

References

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and regression trees*, Wadsworth, Belmont, CA.

Cappelli, C. and Iorio, D. (2010). Detecting contemporaneous mean co-breaking via ART and PCA. *Quaderni di STATISTICA*, **12**, 169-184.

Chang, Y. and Kim, H. (2011). Tree-structured nonlinear regression. *The Korean Journal of Applied Statistics*, **24**, 759-768.

Charbonneau, K. B. (2014). *Multiple fixed effects in binary response panel data models*, The Bank of Canada Working paper, The Bank of Canada, Canada.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of American Statistical Association*, **74**, 829-836.

De'ath, G. (2002). Multivariate retression trees: A new technique for modeling species-environment relationships. *Ecology*, **83**, 1105-1117.

De'ath, G. (2013). *mvpart: Multivariate partitioning*, R package version 1.6-1. Available from `http://CRAN.R-project.org/package=mvpart`.

Dzeroski, S. and Zenko, B. (2004). Is combining classifiers with stacking better than selecting the best one? *Machine Learning*, **54**, 255-273.

Jo, J. and Chang, U. J. (2013). A statistical analysis of the fat mass repeated measures data using mixed model. *Journal of the Korean Data & Information Science Society*, **24**, 303-310.

Lee, S. K. (2005). On generalized multivariate decision tree by using GEE. *Computational Statistics and Data Analysis*, **49**, 1105-1119.

Loh, W.-Y. (2002). Regression trees with unbiased variable selection and interaction detection. *Statistics Sinica*, **12**, 361-386.

Loh, W.-Y. and Zheng, W. (2013). Regression trees for longitudinal and multiresponse data. *The Annals of Applied Statistics*, **7**, 495-522.

Meek, C., Chickering, D. M. and Heckerman, D. (2002). Autoregressive tree models for time-series analysis. *Proceedings of the Second International SIAM Conference on Data Mining*, 229–244.

Rea, W. S., Relae M., Cappelli, C. and Brown J. A. (2010). Identification of changes in mean with regression trees: An application to market research. *Econometric Reviews*, **29**, 754-777.

Segal, M. R. (1992). Tree structured methods for longitudinal data. *Journal of American Statistical Association*, **87**, 407-418.

Sela, R. J. and Simonoff, J. S. (2012). RE-EM trees: A data mining approach for longitudinal and clustered data. *Machine Learning*, **86**, 169-207.

Zhang, H. (1998). Classification trees for multiple binary responses. *Journal of American Statistical Association*, **93**, 180-193.

Journal of the Korean Data & Information Science Society
2014, **25**(6), 1253–1262

http://dx.doi.org/10.7465/jkdi.2014.25.6.1253
한국데이터정보과학회지

Panel data analysis with regression trees[†]

Youngjae Chang[1]

[1]Department of Information Statistics, Korea National Open University

Received 10 August 2014, revised 12 September 2014, accepted 18 September 2014

Abstract

Regression tree is a tree-structured solution in which a simple regression model is fitted to the data in each node made by recursive partitioning of predictor space. There have been many efforts to apply tree algorithms to various regression problems like logistic regression and quantile regression. Recently, algorithms have been expanded to the panel data analysis such as RE-EM algorithm by Sela and Simonoff (2012), and extension of GUIDE by Loh and Zheng (2013). The algorithms are briefly introduced and prediction accuracy of three methods are compared in this paper. In general, RE-EM shows good prediction accuracy with least MSE's in the simulation study. A RE-EM tree fitted to business survey index (BSI) panel data shows that sales BSI is the main factor which affects business entrepreneurs' economic sentiment. The economic sentiment BSI of non-manufacturing industries is higher than that of manufacturing ones among the relatively high sales group.

Keywords: Business survey index, mixed effects model, panel data, regression tree.

[†] This research was supported by Korea National Open University Research Fund.
[1] Assistant professor, Department of Information Statistics, Korean National Open University, Seoul 110-791, Korea. E-mail: yjchang@knou.ac.kr