5장. 의사결정나무

최호식

경기대학교 응용통계학과

Mar, 2018(Kyonggi University)

학습내용

1. 개요
2. 의사결정나무의 형성
3. 여러가지 불순도의 측도
4. 여러가지 의사결정나무 알고리즘
5. 의사결정나무의 특징

의사결정나무 I

- 의사결정나무는 지도학습 기법으로 각 변수의 영역을 반복적으로 분할함으로써 전체 영역에서의 규칙을 생성한다.
- 의사결정나무의 예측력은 다른 지도학습 기법들에 비해 대체로 떨어지나 해석력이 좋다. 즉, 의사결정나무에 의하여 생성된 규칙은 if-then 형식으로 표현되어 이해가 쉽고 SQL(structured query language)과 같은 데이터베이스 언어로 쉽게 구현되는 장점이 있다.

- 의사결정나무의 구성요소
  - 뿌리마디(root node): 시작되는 마디로 전체 자료를 포함
  - 자식마디(child node): 하나의 마디로부터 분리되어 나간 2개 이상의 마디들
  - 부모마디(parent node): 주어진 마디의 상위마디
  - 끝마디(terminal node): 자식마디가 없는 마디
  - 중간마디(internal node): 부모마디와 자식마디가 모두 있는 마디
  - 가지(branch): 뿌리마디로부터 끝마디까지 연결된 마디들
  - 깊이(depth): 뿌리마디부터 끝마디까지의 중간마디들의 수

의사결정나무 II

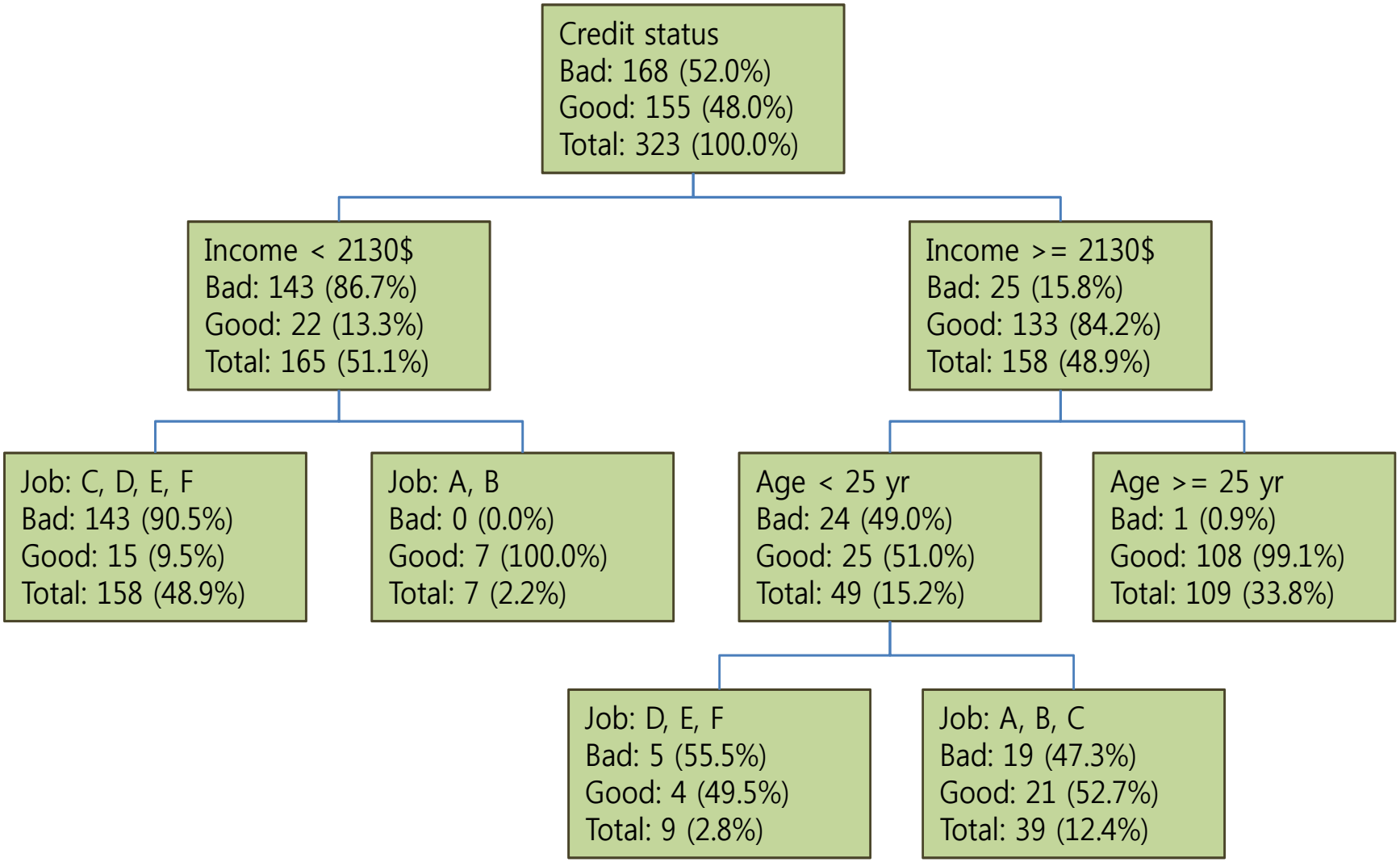

Figure: 신용자료에 대한 의사결정나무

의사결정나무의 형성 I

- 의사결정나무의 종류
  - 회귀나무(regression tree)
  - 분류나무(classification tree)
- 형성과정
  - 성장(growing): 각 마디에서 적절한 최적의 분리규칙을 찾아서 나무를 성장. 정지규칙을 만족하면 중단
  - 가지치기: 오차를 크게 할 위험이 높거나 부적절한 추론규칙을 가지고 있는 가지 또는 불필요한 가지를 제거
  - 타당성 평가: 이익도표(gain chart), 위험도표(risk chart), 혹은 시험자료를 이용하여 의사결정나무를 평가
  - 해석 및 예측: 구축된 나무모형을 해석하고 예측모형을 설정한 후 예측에 적용

회귀나무의 성장 I

- 훈련자료: (x_i, y_i), $i = 1, \ldots, n$, $x_i = (x_{i1}, \ldots, x_{ip})^T$
- 전체 영역을 M개의 영역 $R_1, \ldots, R_M$으로 나누고 각 영역에서 상수값 c_m으로 예측하는 나무모형

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m).$$

  c_m과 R_m은 불순도(impurity)의 측도를 이용하여 값을 정함. 흔히 오차제곱합 $Q_m(T) = \sum_{i=1}^{n}(y_i - f(x_i))^2$을 사용
- 분리변수(split variable) x_j와 분리기준에 따라 영역을 분리
  - 연속형 분리변수: 분리점 s에 대하여 $R_1(j, s) = \{x : x_j \leq s\}$와 $R_2(j, s) = \{x : x_j > s\}$로 분리
  - 범주형 분리변수: 전체 범주를 두 개의 부분집합으로 나눔

회귀나무의 성장 II

- 분리기준

$$\min_{j,s}\left(\min_{c_1}\sum_{x_i\in R_1(j,s)}(y_i-c_1)^2+\min_{c_2}\sum_{x_i\in R_2(j,s)}(y_i-c_2)^2\right).$$

- $\hat{c}_1$과 $\hat{c}_2$: $R_1(j,s)$와 $R_2(j,s)$에 속하는 자료의 y_i값들의 평균
- 분리변수가 주어지면 분리점 s는 쉽게 찾을 수 있음
- 최적 분리를 찾은 후에는 두 영역에 대하여 동일한 과정을 반복
- 정지규칙
  - 나무의 크기를 모형의 복잡도로 볼 수 있고 최적의 나무 크기는 자료로부터 추정
  - 마디에 속하는 자료가 일정 수(가령 5) 이하일 때 분할

회귀나무의 가지치기 I

- T_0: 성장시킨 나무모형, $T \subset T_0$: 가지치기하여 얻을 수 있는 나무모형
- 불순도

$$Q_m(T) = \frac{1}{N_m} \sum_{x_i \in R_m} (y_i - \hat{c}_m)^2$$

- $|T|$는 T에서의 끝마디 개수, N_m은 T의 영역 R_m에 속하는 자료수, $\hat{c}_m$은 영역 R_m에 속하는 자료에 대한 y값들의 평균
- 비용함수

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha|T|$$

로 정의

- 가지치기는 α에 대하여 $C_\alpha(T)$를 최소화하는 $T_\alpha \subset T_0$를 찾는 문제가 됨

회귀나무의 가지치기 II

- $\alpha \geq 0$는 나무모형의 크기와 자료에 대한 적합도를 조절하는 조율모수로 α값이 크면(작으면) T_α의 크기는 작아(커)짐
- $\alpha = 0$이면 가지치기는 일어나지 않고 T_0를 최종모형으로 준다.
- $T_{\hat{\alpha}}$: 가지치기된 최종 모형
- 시험자료가 $x \in R_m$에 대하여 $\hat{y} = \hat{c}_m$으로 예측

분류나무

- 불순도의 측도: 카이제곱 통계량, 지니지수(Gini index), 엔트로피지수(entropy index) 등
- 성장: 회귀나무와 동일
- 가지치기: 흔히 오분류율을 불순도의 측도로 사용하여 회귀나무와 동일한 방식으로 실시하여 최종 분류나무모형 $T_{\hat{\alpha}}$을 얻음
- $\hat{p}_{mk}$를 최종 모형의 영역 R_m에 속하는 자료중 출력변수의 범주가 k인 자료의 비율
- $x \in R_m$이면 예측값은 $\hat{y} = \arg\max_k \hat{p}_{mk}$로 예측
- 분류나무는 예측값을 각 마디에서 다수결(majority vote) 원칙으로 정함

불순도의 여러 가지 측도 I

- 성장 단계에서 최적 분리기준을 정하는데 사용
- 분류나무의 측도
  - 카이제곱(χ^2) 통계량
  - 지니지수
  - 엔트로피지수

불순도의 여러 가지 측도 II

- 예제 자료

	Good	Bad	Total
Left	32 (56)	48 (24)	80
Right	178 (154)	42 (66)	220
Total	210	90	300

실제도수(O)와 괄호안의 값인 기대도수(E)

1. 카이제곱 통계량
   - 정의: 각 셀에 대한 ((기대도수−실제도수)의 제곱/기대도수)의 합
   - 최대가 되는 분리를 사용

$$\frac{(56-32)^2}{56}+\frac{(24-48)^2}{24}+\frac{(154-178)^2}{154}+\frac{(66-42)^2}{66}=46.75$$

불순도의 여러 가지 측도 III

2. 지니지수
   - 정의:

$$2\,(\Pr(\text{Left에서 Good})\Pr(\text{Left에서 Bad})\Pr(\text{Left}) + \Pr(\text{Right에서 Good})\Pr(\text{Right에서 Bad})\Pr(\text{Right}))$$

   - 최소가 되는 분리를 선택

$$2\left(\frac{32}{80}\times\frac{48}{80}\times\frac{80}{300}+\frac{178}{220}\times\frac{42}{220}\times\frac{220}{300}\right)=0.355$$

불순도의 여러 가지 측도 IV

3. 엔트로피지수=엔트로피(Left)Pr(Left) + 엔트로피(Right)Pr(Right)

$$\begin{aligned} \text{엔트로피(Left)} &= -\Pr(\text{Left에서 Good}) \log_2 \Pr(\text{Left에서 Good}) \\ &\quad -\Pr(\text{Left에서 Bad}) \log_2 \Pr(\text{Left에서 Bad}) \end{aligned}$$

- 최소가 되는 분리를 선택

$$-\left(\frac{32}{80}\log_2\left(\frac{32}{80}\right)+\frac{48}{80}\log_2\left(\frac{48}{80}\right)\right)\frac{80}{300}$$
$$-\left(\frac{178}{220}\log_2\left(\frac{178}{220}\right)+\frac{42}{220}\log_2\left(\frac{42}{220}\right)\right)\frac{220}{300}=.7747$$

불순도의 측도 예제 I

Temperature	Humidity	Windy	Class
Hot	High	False	N
Hot	High	True	N
Hot	High	False	P
Mild	High	False	P
Cold	Normal	False	P
Cold	Normal	True	N
Cold	Normal	True	P
Mild	High	False	N
Cold	Normal	False	N
Mild	Normal	False	P
Mild	Normal	True	P
Mild	High	True	P
Hot	Normal	False	N
Mild	High	True	P

불순도의 측도 예제 II

1. Temperature를 기준으로 분리하는 경우
   - Left={Hot}, Right = {Mild, Cold}일 때

	N	P	계
Left	3	1	4
Right	3	7	10
계	6	8	14

$$\text{지니지수} = 2\left(\frac{1}{4}\times\frac{3}{4}\times\frac{4}{14}+\frac{3}{10}\times\frac{7}{10}\times\frac{10}{14}\right) = 0.4071.$$

   - Left={Mild}, Right = {Hot, Cold}일 때

	N	P	계
Left	1	5	6
Right	5	3	8
계	6	8	14

불순도의 측도 예제 III

$$\text{지니지수} = 2\left(\frac{1}{6}\times\frac{5}{6}\times\frac{6}{14}+\frac{5}{8}\times\frac{3}{8}\times\frac{8}{14}\right) = 0.3869.$$

- Left={Cold}, Right = {Hot,Mild}일 때

	N	P	계
Left	2	2	4
Right	4	6	10
계	6	8	14

$$\text{지니지수} = 2\left(\frac{2}{4}\times\frac{2}{4}\times\frac{4}{14}+\frac{5}{10}\times\frac{6}{10}\times\frac{10}{14}\right) = 0.4860.$$

2. Humidity를 기준으로 분리하는 경우
   Humidity는 High와 Normal의 두 가지 값만 가지므로 Left를 둘중 어느 한 범주로 잡아도 동일한 결과를 준다. 따라서 편의상 Left={High}, Right = {Normal}라 하자.

불순도의 측도 예제 IV

	N	P	계
Left	3	4	7
Right	3	4	7
계	6	8	14

$$\text{지니지수} = 2\left(\frac{3}{7}\times\frac{4}{7}\times\frac{7}{14}+\frac{3}{7}\times\frac{4}{7}\times\frac{7}{14}\right) = 0.4897.$$

3. Windy를 기준으로 분리하는 경우
   Humidity와 마찬가지로 범주가 둘이므로 Left=$\{$False$\}$, Right = $\{$True$\}$라 하자.

	N	P	계
left	4	4	8
right	2	4	6
계	6	8	14

$$\text{지니지수} = 2\left(\frac{4}{6}\times\frac{2}{6}\times\frac{6}{14}+\frac{4}{8}\times\frac{4}{8}\times\frac{8}{14}\right) = 0.4762.$$

불순도의 측도 예제 V

모든 가능한 분리에 대하여 결과를 종합해보면 Temperature에 대하여 Left $= \{\text{Mild}\}$, Right $= \{\text{Hot, Cold}\}$로 분리하는 것이 지니지수 측면에서 최적임

여러 가지 의사결정나무 알고리즘

- CART(classification and regression trees)
  - 가장 널리 사용되는 의사결정나무 알고리즘으로 이진분리(binary split)
  - 분류나무:지니지수, 회귀나무: 분산
  - 개별 입력변수뿐만 아니라 입력변수들의 선형결합들중에서 최적의 분리를 찾을 수도 있음
- C4.5와 C5.0
  - 다지분리(multiple split)가 가능
  - 엔트로피지수를 사용
- CHAID(chi-squared automatic interaction detection)
  - 가지치기를 하지 않고 적당한 크기에서 나무모형의 성장을 중지
  - 입력변수는 범주형
  - 카이제곱 통계량을 사용

의사결정나무의 특징

- 가장 설명력이 있는 변수에 대하여 최초로 분리가 일어남
- 장점
  - if-then 형식의 이해하기 쉬운 규칙을 생성
  - 연속형 변수와 범주형 변수를 모두 취급할 수 있음
  - 모형에 대한 가정(예: 선형회귀의 선형성, 등분산성 등)이 필요 없는 비모수적 방법
- 단점
  - 회귀모형에서는 그 예측력이 떨어짐
  - 일반적으로 복잡한 나무모형은 예측력이 저하되고 해석 또한 어려우며, 상황에 따라 계산량이 많을 수도 있음
  - 분산이 매우 큰 불안정한 방법, 배깅(bagging)과 같은 앙상블(ensemble) 알고리즘을 적용하여 분산 감소