응용통계연구 제 **16**권 1호, *2003*년, *pp. 151-167*

혼합자료에서 독립성 검정에 의한 연관성 측정

이승천[1] 허문열[2]

요 약

두 확률변수의 연관성을 측정하는 측도는 많이 있으나, 이러한 측도는 같은 유형인 변수들 간의 관계를 측정하기 위한 것으로 여러 가지 유형의 변수들이 혼재되어 있는 혼합자료에서 사용하기는 곤란하다. 본 논문에서는 두 확률변수의 독립성 검정을 통해 구한 p-값으로 혼합자료에서 사용될 수 있는 새로운 연관성 측도를 구하였으며, 이렇게 구하여 진 연관성 측도가 혼합자료에서 변수들 간의 연관성을 비교하는데 유용하게 사용될 수 있음을 보였다.

주요용어: 연관성 측도, 독립성 검정, 혼합자료, p-값, 순위상관계수, 크루스칼-왈리스 검정, 피어슨의 X^2 검정, 피어슨의 ϕ^2

1. 연구의 필요성

두 확률변수의 연관성 측도는 한 변수가 다른 변수에 대해 어느 만큼의 정보를 갖고 있는지를 측정하는 측도인데, 많은 통계적 또는 데이타 마이닝 모형은 변수들 간의 관계를 규명하는 것을 목적으로 하고 있으므로 분석의 초기 단계에서 연관성 측도는 매우 중요한 역할을 할 수 있다. 예를 들어 통계적 모형에서 너무 많은 수의 변수들은 모형의 해석을 어렵게 함으로 미리 상관계수등과 같은 연관성 측도를 이용하여 필요한 변수만을 선택하여 모형을 설정하는 것이 바람직하다. 또한 데이터 마이닝의 한 기법인 결정나무(decision tree)는 분석 결과가 사람이 이해할 수 있는 규칙으로 표현될 수 있어 선호되기도 한다. 그러나 너무 많은 수의 입력변수가 사용될 경우, 매우 큰 나무가 형성되어 규칙을 이해하기 어려울 수 있다. 비록 가지치기(pruning)에 의해 끝마디의 수를 줄인다고는 하나, 이는 매우 제한적인 것으로 본 논문의 예제에서 보듯이 목표변수와 연관성이 큰 변수만을 선택하여 결정나무를 구성할 경우 보다 좋은 결과를 가져올 수 있다.

이와 같이 통계적 자료분석에서 유용성을 갖는 연관성 측도에 대한 연구는 매우 오랜 역사를 갖고 있으며, 그중 특기할만한 것으로 피어슨의 상관계수, 스피어만 순위상관계수, 켄달의 tau (Kendall 1938) 등을 들 수 있다. 이와는 다른 유형의 연관성 측도로 분할표(contingency table)에서 정의되는 크래머의 V (Cramér 1946), 굿맨-크루스칼의 tau (Goodman and Kruskal 1954), 코헨의 kappa (Cohen 1960), 율의 Q (Yule 1912) 등 여러 가지 연관성 측도가 있다. 그러나 전자와 후자의 연관성 측도는 각각 두 연속형변수 또는 두

1) (447-791) 경기도 오산 양산시 411, 한신대학교 정보통계학과, 교수
E-mail: seung@hanshin.ac.kr
2) (110-745)서울시 종로구 명륜동 3-53, 성균관대학교 통계학과, 교수
E-mail: myhuh@skku.ac.kr

이산형변수간의 연관성만을 측정하기 위한 것으로, 서로 다른 유형의 변수간의 연관성을 비교할 수 있는 측도에 대한 연구결과는 찾아 볼 수 없다. 또한 기존의 연관성 측도를 사용한다고 할지라도 이를 이용하여 여러 유형의 변수들이 섞여 있는 혼합자료에서 어느 변수들이 더 연관성을 갖는지 비교할 수 없다는 문제점을 갖는다. 즉, 통계적 분석을 요구하는 많은 자료들, 특히 데이터 마이닝을 위한 자료들은 흔히 연속형과 이산형 변수들이 혼재된 혼합자료들이 대부분이다. 따라서 혼합자료에서 연관성을 측정할 수 있는 측도에 대한 연구를 필요로 한다.

2. 이론적 배경

확률변수 X와 Y가 연관성을 갖는다면 $\Pr(Y \in A) < \Pr(Y \in A|X)$를 만족하는 사상 A가 존재하며, 연관성이 클수록 이 사상은 큰 확률값을 갖게 된다. Shannon (1948)은 이를 기초로하여 엔트로피를 고안하였고, Goodman과 Kruskal (1954) 그리고 Silvey(1964)도 같은 사실을 기초로 하여 연관성 측도를 고안하였다. 특히 Silvey는 변수의 유형과는 무관한 일반성을 갖는 연관성 측도로서 Δ를 다음과 같이 정의하였다.

$$\Delta = \iint_{\{(x,y):\phi(x,y)>1\}} [\, p(x,y) - p(x)p(y) \,] dxdy.$$

여기서 $\phi(x,y)$는 X, Y 주변분포(marginal distribution)들의 곱(product)에 대한 결합분포의 라돈-니코딤 도함수로서 $p(x,y), p(x), p(y)$를 각각 X와 Y의 결합밀도함수과 주변밀도함수라고 할 때, $\phi(x,y) = p(x,y)/p(x)p(y)$와 같다.

Silvery의 Δ는 변수의 유형과는 무관한 일반성을 갖는 연관성 측도이기는 하지만 표본에 의해 Δ를 추정하려면 결합확률밀도함수를 추정하여야 하는 등 여러 가지 계산상의 문제를 내포하고 있다. 즉, 핵밀도함수추정등과 같은 방법을 사용하여 밀도함수의 추정이 가능할 지라고 이차원의 밀도함수에 대한 추정은 효율이 떨어지는 것이 일반적이며 또한 적분도 간단한 문제는 아니다. 이제 새로운 연관성 측도를 다음과 같이 정의하기로 하자.

$$\alpha(X,Y) = \iint \mathrm{I}(x,y) dF(x) dG(y).$$

단 F와 G는 X와 Y의 주변분포함수이고, $\mathrm{I}(x,y)$는

$$\mathrm{I}(x,y) = \begin{cases} 1 & \text{if } p(x,y) > p(x)p(y) \\ 0, & \text{otherwise} \end{cases}$$

과 같이 정의한다.

Rényi (1959)는 연관성 측도가 갖추어야 할 7가지 조건에 대해 열거하였다. Rényi의 조건에 따라 α의 성질을 열거하면 다음과 같다.

보조정리 2.1 연관성 측도로서 α는 다음과 같은 성질을 갖는다.

1. α는 두 변수가 모두 퇴화분포을 갖는 경우를 제외하고 모든 경우에 정의될 수 있다. 특히 하나의 변수가 퇴화분포를 갖을 때에도 정의되며, 이 경우 α는 1의 값을 갖는다.
2. $\alpha(X,Y) = \alpha(Y,X)$. 즉, α는 대칭성을 갖는다.
3. $0 \le \alpha \le 1$
4. $\alpha(X,Y) = 0$인 필요충분조건은 모든 x,y에서 $p(x,y) = p(x)p(y)$이다.
5. 두 변수가 연속형일때 $\alpha = 1$인 필요충분조건은 이변량정규분포에서 상관계수가 -1 또는 1인 경우와 같이 결합분포가 특이분포를 갖는 경우이다. 그러나 이산형분포에서는 이러한 특징을 갖지 않는다. 이는 Δ의 성격과도 같다.
6. h와 k가 보렐 가측 1-1 함수이면 $\alpha(X,Y) = \alpha(h(X),k(Y))$이다.
7. (X,Y)가 이변량정규확률변수이면 α는 $|\rho|$의 단조증감함수이다.

Rényi의 7번째 조건은 원래 이변량 정규분포에서 연관성 측도는 $|\rho|$와 같아야 할 것을 요구하고 있으나, 이는 너무 제약적인 것이므로 Bell(1962)은 Rènyi가 제시한 조건을 위의 7번째 조건으로 대치하였다. 즉, α는 Bell의 조건을 만족한다. 또 α는 정의에서 보듯 Δ와 거의 같은 성격을 갖고 있어, 보조정리 2.1에 관한 증명은 Silvey(1964)를 참조할 수 있다.

α의 추정이 비롯 Δ보다는 쉽게 구할 수 있다고 하지만 역시 이변량분포의 밀도함수를 추정하여야 하는데, 밀도함수추정의 비효율성으로 인하여 시뮬레이션 결과 표본들의 연관성을 그다지 효율적으로 구하여 주지는 못하였다.

한편 α를 다른 측면에서 살펴보면, (X,Y)의 실제 결합확률밀도함수를 $p^0(x,y)$라고 하였을때, α는 가설

$$\begin{aligned} H_0 &: p^0(x,y) = p(x)p(y) \\ H_1 &: p^0(x,y) = p(x,y). \end{aligned} \tag{2.1}$$

에 대한 검정의 유의수준임을 알 수 있다. 즉, 네이만-피어슨 정리에 따라 (2.1) 가설의 최강력 검정은

$$\varphi(x,y) = \begin{cases} 1, & p(x,y) > kp(x)p(y) \text{ 일때,} \\ 0, & p(x,y) < kp(x)p(y) \text{ 일때.} \end{cases} \tag{2.2}$$

와 같이 주어지는 것으로 알려져 있다. α는 $k = 1$로 주어지는 최강력 기각역의 유의수준이다. 그러나 표본에 의해 α를 구할 때에는 $p(x,y), p(x), p(y)$가 주어지지 않으므로 앞에서 설명한 바와 같이 밀도함수에 대한 추정을 필요로 하게 된다.

(2.1)에서 귀무가설은 확률변수 X와 Y가 서로 독립임을 의미한다. 또한 α는 집합 $A = \{(x,y) : p(y|x) > p(y)\}$의 확률이 높을 수록 큰 값을 갖게 된다. 일반적으로 A의 확률이 크면 클 수록 그만큼 독립성과는 거리가 있다는 것을 의미함으로 α는 독립성 검정에서 귀무가설을 부정하는 정도로서 파악할 수도 있다. 한편 표본에 의해 검정을 할 때, 귀무가설을 기각하는 정도는 $1-p$ 값으로 측정할 수 있다. 즉, 연관성 측도의 계산을 검정과 연관시킨

다면, 실제의 검정에 있어서는 (2.2)와 같은 최강력 검정을 구할 수는 없다. 이 경우 검정은 모든 가능한 분포에 대해 검정할 수 있어야 함으로 비모수 검정에 해당된다. 따라서 모든 분포에 대해 같은 기각역을 갖을 수 밖에 없다. 기각력이 같아야 한다면, 표본에 의해 가설을 기각하는 정도를 나타내는 값은 $1-p$ 값이다. 그러므로 표본들의 연관성을 측정하는 측도로서 독립성 검정의 $1-p$ 값을 고려할 수 있다.

통계학에서는 독립성을 검정하기 위한 많은 검정 방법이 제안되어 있는데, 이러한 검정은 변수들의 유형에 의존하는 바가 크다. 예를 들어 두 변수가 연속형 확률변수일 때에는 스피어만의 순위상관계수에 의한 검정을 생각할 수 있으며, 이산형 변수인 경우는 흔히 피어슨의 X^2 검정을 사용할 수 있다. 한편 X는 이산형 변수이고 Y가 연속형 변수일 때 X와 Y가 서로 독립이면 X가 주어질 때 Y의 조건부 분포가 모든 X의 값과 관계없이 같은 분포를 갖게되므로 가설 $H_0 : F(y|x_1) = F(y|x_2) = \cdots = F(y|x_m)$에 대한 검정으로 이해될 수 있다. 그러므로 이산형변수와 연속형변수의 독립성 검정으로 크루스칼-왈리스 검정을 사용할 수 있겠다. 이러한 검정 방법 이외에도 켄달 검정 등 독립성을 검정하기 위한 많은 방법들이 제시되어 있다.

Randle과 Wolfe(1979)에 따르면 상기된 검정들은 해당되는 검정에 있어 유용성이 입증되어 있어, 각각의 경우에 연관성 측도로 사용될 수 있겠으나, 본 논문은 모든 유형의 변수들에게 공통적으로 적용된 연관성 측도를 구하는 것이 목적이므로 이러한 검정들의 $1-p$ 값이 연관성을 일관되게 측정할 수 있는가에 대한 의문을 갖게된다. 예를 들어 순위상관계수에 의한 검정에서 얻은 연관성 측도의 값이 X^2 검정에서 구한 값 보다 크다고 할 때, 전자가 더 큰 연관성을 갖는다고 할 수 있겠는가? 즉, 연관성 측도로 사용하려면 변수의 유형과 관계없이 사용할 수 있는 검정이 있는가 라는 문제에 직면한다.

Lancaster (1969)는 피어슨의 ϕ^2 측도를 귀무가설하의 밀도함수와 실제 밀도함수와 차이에 대한 L_2-노름으로 정의하였는데, a_i들 일반화 푸리에 계수라고 할 때, ϕ^2는 Parseval 관계에 의해 다음과 같이 푸리에 계수들의 무한급수로 나타낼 수 있다.

$$\phi^2 = \sum_{i=1}^{\infty} a_i$$

즉, 가설과 실제의 밀도함수가 같을 필요충분 조건은 모든 i에 대해 $a_i = 0$이다.

Eubank, Lariccia, Rosenstein (1987)은 많은 통계적 가설 검정 방법들이 실제로 $H_0 : a_i = 0, H_1 : a_i \neq 0$에 대한 검정인 것을 보였다. 예를 들어, X와 Y가 연속형 확률변수이면, 두 변수가 서로 독립이라는 가설하에서 L_2-노름의 기저가 르장드르 다항식으로 설정하면, 첫번째 푸리에 계수 a_1은

$$\tilde{a}_1 = 12n \sum_{i=1}^{n} \left(R_{X_i} - \frac{n}{2}\right)\left(R_{Y_i} - \frac{n}{2}\right),$$

와 같이 추정될 수 있음을 보였다. 여기서 R_{X_i}와 R_{Y_i}는 X와 Y의 각 그룹에서 X_i와 Y_i의 순위를 나타낸다. 따라서 선형요소인 a_1에 대한 검정은 스피어만의 순위상관검정과 같다. 같은 방법으로 크루스칼-왈리스 검정과 피어슨의 X^2 검정은 각 경우의 선형요소 a_1에 대

한 가설

$$H_0 : a_1 = 0, \quad H_1 : a_1 \neq 0$$

을 검정한다. 즉, 상기된 세 검정은 변수의 유형은 다를지라도 같은 기원을 같는 검정으로 각 검정에서 구한 $1-p$ 값은 연관성 측도로서 일관성을 갖을 것으로 기대되며, 다음 절에서 시뮬레이션에 의해 이를 확인하기로 한다.

3. 시뮬레이션

스피어만의 순위상관 검정, 크루스칼-왈리스 검정, 피어슨의 X^2 검정에서 구한 $1-p$ 값을 각각 $\delta_S, \delta_K, \delta_C$로 정의하기로 하자.

정의 3.1 두변수 X와 Y의 연관성 측도 δ를 다음과 같이 정의한다.

$$\delta = \begin{cases} \delta_S, & X\text{와 } Y\text{가 모두 연속형일 때} \\ \delta_K, & X\text{가 연속형이고, } Y\text{는 이산형일 때} \\ \delta_C, & X\text{와 } Y\text{가 모두 이산형일 때} \end{cases}$$

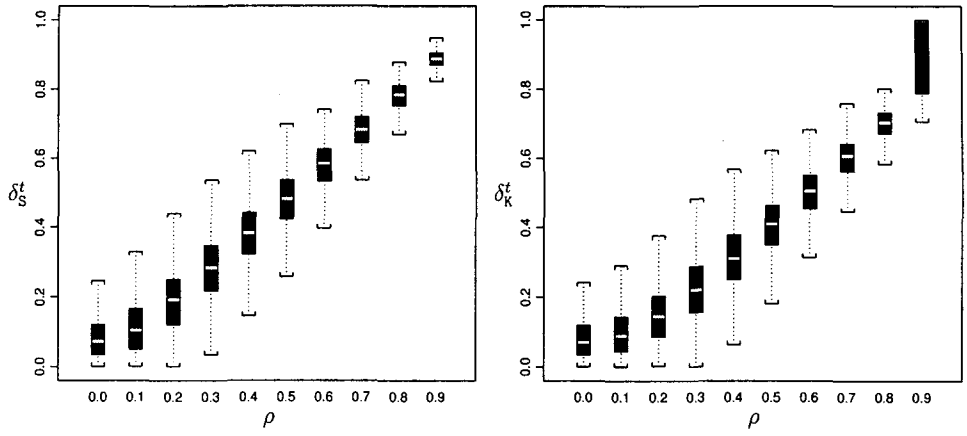
E(δ)은 두 변수들의 연관성이 커짐에 따라 처음에는 매우 급격히 1에 가까운 값까지 증가하다가 1에 가까워지면 이후에는 매우 완만한 경사를 이루는 S형 곡선을 갖는다. 그러므로 어느 변수들이 더 연관성을 갖고 있는지는 δ의 크기 비교를 통해 판단할 수 있지만, δ의 이러한 특성으로 인해 값만으로는 어느 정도의 연관성을 갖고 있는지 판단하기 어렵다. 물론 이러한 해석상의 문제가 연관성 측도로서 δ만이 갖는 단점이라서 할 수는 없다. Bishop, Fienberg, Holland(1975)에 따르면 분할표에서 연관성을 측정하는 많은 연관성 측도가 값 자체의 해석보다는 단지 비교를 위한 것임을 언급하고 있다(386쪽 참조). 그러나 본 논문에서는 연관성 측도로서의 해석을 위해 적절한 변환을 제시하기로 한다.

$\tilde{\rho}$를 순위상관계수라 하면, $\sqrt{n-1}\,\tilde{\rho}_n$는 표본크기 n이 커짐에 따라 표준정규분포에 수렴한다. 따라서 표본크기 n이 크다고 할 때, δ_S는 $2\Phi(\sqrt{n-1}\,|\tilde{\rho}_n|)) - 1$에 의해 구하여 진다. 여기서 Φ는 표준정규분포의 분포함수이다. 즉, δ_S와 $|\tilde{\rho}|$ 는 단조증가함수에 의해 1-1 대응관계를 갖는다. 그러므로 연관성 측도로서의 δ_S는 $|\tilde{\rho}|$와 같은 특성을 갖는다고 할 수 있다. 일반적으로 순위상관계수는 두 연속형 확률변수의 연관성을 측정하는 측도로서 많은 장점을 갖고 있으므로 이러한 장점이 그대로 δ_S에 상속됨을 알 수 있다. 한편 분할표에서 연관성을 측정하는 측도로 잘알려져 있는 크래머의 V는 δ_C와 마찬가지로 피어슨의 X^2 통계량을 기초로 만들어져 δ_C와 1-1 대응관계를 갖는다.

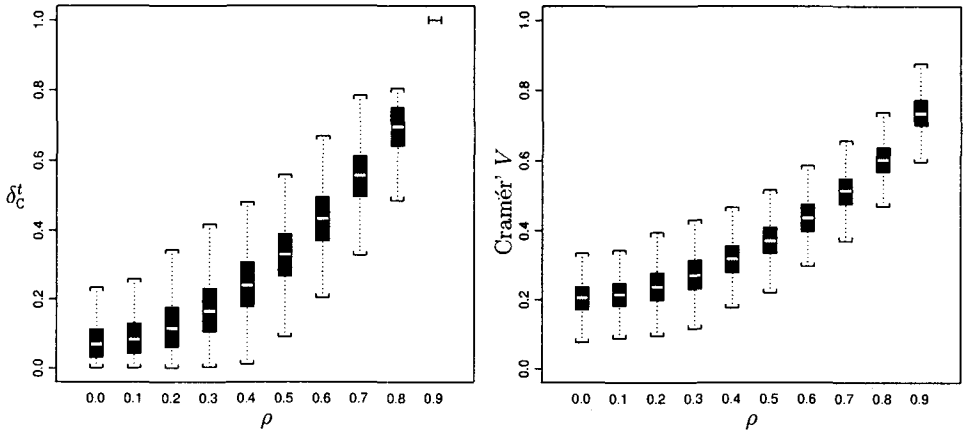
이제 모든 δ에 대해 다음과 같은 변환을 고려하기로 하자.

$$\delta^t = \frac{1}{\sqrt{n-1}}\Phi^{-1}\left(\frac{\delta+1}{2}\right) \tag{3.1}$$

δ^t는 $\delta = \delta_S$일때 순위상관계수가 된다. 연관성 측도로서 δ^t의 특성을 살펴보기 위해 상관계수가 각각 $\rho = 0.0, 0.1, \ldots, 0.9$인 이변량 정규분포에서 100개의 표본을 추출하여 δ^t를 구



(a) 스피어만의 순위상관 검정 (b) 크루스칼-왈리스 검정



(c) 피어슨의 X^2 검정 (d) 크래머의 V

그림 3.1: 이변량정규분포에서 상관계수값에 따른 δ^t들의 분포

하였다. 이러한 과정을 1000회 반복하여 구한 δ^t의 분포를 상자그림으로 표현하였다. δ_K^t와 δ_C^t는 추출된 표본을 같은 확률을 갖는 네개의 구간에 의해 X 또는 X와 Y를 모두를 이산화(discretize)하여 δ_K 또는 δ_C를 구한 후 (3.1)에 의해 δ_K^t, δ_C^t를 구하였다. 그 결과는 그림 3.1 (a), (b), (c)와 같다. 그림 3.1 (d)는 분할표에서 연관성 측도로서 흔히 사용되는 크래머의 V에 대한 분포를 나타낸다. 이 그림은 δ_C^t와의 비교를 위해 수록하였다.

그림에 나타난 결과를 요약하면 다음과 같다.

1. δ_S^t는 분포의 중심이 상관계수 ρ의 증가에 비례하여 같은 거의 같은 비율로 증가한다. δ_S^t는 순위상관계수이므로 이는 예상된 결과와 같다.

2. δ_K^t도 역시 ρ의 증가에 따라 중심이 증가하고 있으나, 완만한 곡선에 따라 증가하고 있다. ρ가 같은 값일 때, δ_S^t와 비교하면, δ_K^t가 약간씩 작은 값을 갖는 것을 볼 수 있으나, 이는 연속형변수를 이산화하는 과정에서 손실되는 정보에 의해 연관성이 약간씩 작아진 것으로 판단된다.

3. δ_C^t는 δ_K^t보다 더욱 곡선에 가까운 형태로 증가하고 있고, 같은 ρ 값에서 δ_S^t나 δ_K^t보다 작은 값을 갖는 경향이 있다. 이 역시 이산화에 따른 연관성의 손실로 인한 것으로 보인다. δ_K^t와 비교한다면, δ_C^t의 경우는 두 변수 모두를 이산화하는 과정에서 더 많은 연관성의 상실이 일어 났다고 판단된다.

4. 크래머의 V는 ρ에 따라 분포의 중심에 대략 0.2에서 0.7 사이의 값을 갖는다. 따라서 δ_S^t 또는 δ_K^t와 비교하여 연관성의 크고 작음을 판단하기는 어렵다.

δ_C^t와 δ_K^t는 $\rho = 0.9$에서 대부분 1의 값을 갖고 있는 것을 볼 수 있다. 그러나 이러한 현상은 계산상의 문제로 인하여 발생한 것이다. 즉, 두 변수간의 연관성이 큼에 따라 δ의 값이 1에 매우 가까운 값을 갖게 되어, 수치해석에 의해 구한 $\Phi^{-1}((\delta+1)/2)$값이 무한값을 갖게 되는데, 이러한 경우를 모두 1로 처리한 때문이다.

δ를 구하기 위해 사용된 세 개의 검정은 모두 일치검정(consistent test)으로 두 변수간에 연관성이 있다면 표본크기 n이 커짐에 따라 δ는 1로 수렴하게 된다. 따라서 n이 크면 δ^t는 상기된 계산상의 문제점을 안고 있다. δ의 계산에 있어서도 이러한 문제가 일어날 수 있으나, 일반적으로 분위수(quantile)를 구하는 것보다는 p-값을 구하는 것이 보다 효율적인 것으로 알려져 있다(박성현, 허문열, 1983 참조). 그러므로 표본크기가 큰 경우에는 δ^t보다는 δ에 의해 연관성 크기를 비교하는 것이 바람직하다.

4. 사용 예제

이 절에서는 실제 자료를 사용하여 본 논문에서 제안한 연관성 측도의 효율을 살펴보기로 한다. 첫째 예제는 결정나무에 연관성 측도를 적용시키는 것으로서 이를 위해 "http://www.ics.uci.edu/ mlearn/MLSummary.html" 사이트에서 인용한 "독일신용자료"를 사용하였다.

두번째 예제는 로지스틱 회귀모형에 δ를 적용한 것으로 "http://worldcup.espnsoccernet.com/teamstats" 사이트에 나타나있는 FIFA 자료를 사용하였다.

4.1. 사례 1

결정나무는 예측 또는 분류를 위한 강력한 데이터 마이닝 도구로서 모형의 추정 결과가 사람이 이해하기 쉬운 규칙에 의해 표현되기 때문에 모형을 이해하기 쉽다는 장점을 갖는다. 이러한 장점때문에 결정나무는 때때로 모형에 의한 예측보다, 변수간의 관계를 파악하는데 더욱 주안점을 두게 되는 경우도 있다.

결정나무의 규칙이 쉽게 이해되려면 가짓수가 가능한 적을 수록 좋다. 또한 너무 많은 가짓수는 과적합 문제를 야기시켜 모형의 훈련에 사용되지 않은 새로운 자료를 예측에 사

표 4.1: 독일 신용자료

변수명	속성	변수명	속성	변수명	속성
checking	ordinal	duration	continuous	history	ordinal
purpose	nominal	amount	continuous	savings	ordinal
employed	ordinal	installp	ordinal	martial	ordinal
coapp	ordinal	resident	ordinal	property	ordinal
age	continuous	other	ordinal	housing	ordinal
existcr	ordinal	job	ordinal	depends	binary
telephon	binary	foreign	binary	good_bad	binary

용하였을 때 예측오차가 커질 수도 있다. 그러므로 적절한 정지규칙이나 가지치기 등에 의해 적정규모의 결정나무를 형성하도록 하는 것이 바람직하다. 그러나 Breiman, Friedman, Olshen, Stone(1984) 그리고 Quinlan(1990)은 정지규칙에 의존하기보다는 가지치기를 사용할 것을 권장하고 있다. 즉, 그들은 과적합에 의한 큰 나무로부터 가지치기를 통하여 적정규모의 나무를 구하도록 권장하고 있다. 특히 Breiman 등은 자식마디에 속한 객체가 모두 하나의 클래스에 속할 때까지 또는 더 이상의 분류가 불가능할 때 까지 분류를 계속한 후에 가지치기를 하도록 하고 있다.

일반적으로 많은 수의 입력변수로부터 형성된 결정나무는 가짓수가 많아지는 경향이 있다. 비록 가지치기에 의해 불필요한 가지를 제거할 수 있다고는 하지만, 불필요한 입력변수가 들어 있을 때, 때때로 과적합문제가 발생되는 경우가 있다. 그러므로 사전에 목표변수와 연관성이 없는 변수들을 필터링하였을 경우, 보다 바람직한 결정나무를 구할 수도 있다.

독일 신용자료는 고객의 신용등급(good_bad)과 이를 설명하기 위한 20개의 변수들의 1000개의 관찰값으로 구성된 자료로서 자료에 포함된 변수명과 속성은 표 4.1과 같다. 이 자료의 분석목적은 고객의 신용평가(good_bad)를 예측하기 위한 것으로 먼저 20개의 입력변수를 설정하고 SPSS AnswerTree Release 2.1.1를 이용하여 CART알고리즘에 따라 모형을 추정하였다.

모형 추정에 사용된 정지규칙은 부모마디와 자식마다의 수가 각각 10과 5미만이면 분

표 4.2: 13개의 종료마디를 결정나무의 오분류표

		실제 범주		
		good	bad	합
예측 범주	good	223	49	272
	bad	40	53	93
		263	102	365

위험 추정값: 0.243836
위험 추정값의 SE: 0.0224755

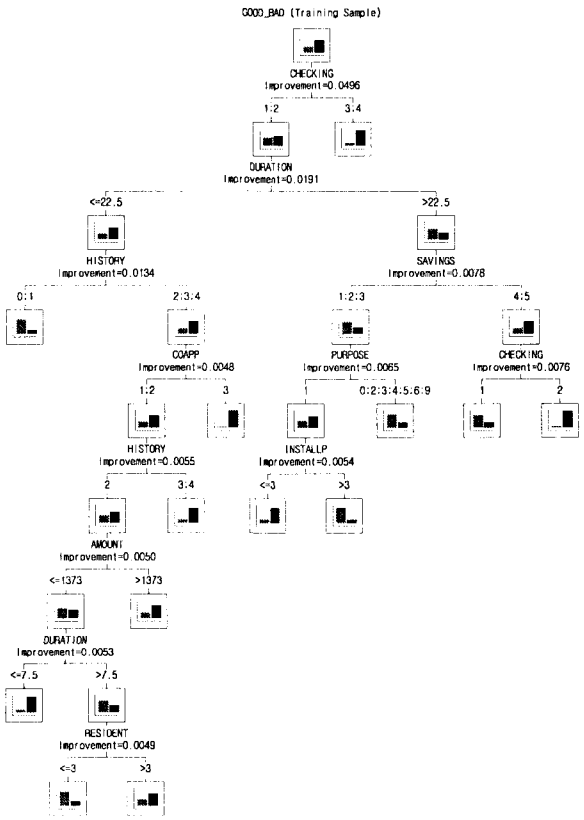


그림 4.1: 20개 입력변수로 구하여진 결정나무

리를 중지하도록 설정하였고, 나무의 깊이에 대한 제약은 두지 않았다. 10과 5라는 숫자는 Breiman이 제안한 수와 비교하여 다소 큰 값이지만, 이것보다 작은 값으로 설정하였을 경우 가지치기를 한 이후에도 너무 큰 나무가 형성됨으로, 나무의 크기를 줄이기 위해 이 값으로 설정하였다. 자료는 각각 635와 365개의 훈련자료와 검증자료로 나누었고, 635개의 훈련자료로 추정된 모형을 비용합성(cost-complexity) 가지치기 알고리즘과 1 SE 규칙을 적용하여 얻어진 결정나무는 13개의 종료마디를 갖게 되었다. 이렇게 구하여진 결정나무는 그림 4.1와 같다. 또 검증자료를 예측한 결과에 대한 정보는 표 4.2에 주어져 있다.

20개의 입력변수 가운데 목표변수와 연관성이 큰 일부의 변수만을 선택하기 위해 구한 δ 값을 크기순으로 나열하면 그 결과는 표 4.3와 같다. δ-값은 1000개의 자료로부터 계산되었는데, 표본크기의 값이 크면 δ^t의 계산에 어려움이 있어 δ^t 대신 δ값을 구하였다.

몇 개의 변수를 입력변수로 선택할 것인가는 주관적인 문제이기는 하지만, δ의 값을 살펴보면 checking에서 coapp까지 δ 값의 변화가 약 0.036에 불과하지만 coapp과 installp의 사이의 δ 값의 변화는 약 0.104로 값의 변화가 상대적으로 큰 것으로 판단된다. 그러므로 coapp의 δ 값보다 큰 값을 갖는 변수들을 입력변수로 선택한다. 이렇게 선택된 변수들의 δ

표 4.3: 크기 순으로 나열된 δ의 값

변수명	δ	변수명	δ	변수명	δ
checking	1.0000000	age	0.9996089	installp	0.8599667
duration	1.0000000	employed	0.9989545	telephon	0.7211238
history	1.0000000	other	0.9983707	existcr	0.5548559
savings	0.9999997	amount	0.9940845	job	0.4034184
property	0.9999714	foreign	0.9841692	resident	0.1384479
housing	0.9998883	marital	0.9777620	depends	0.0000000
purpose	0.9998843	coapp	0.9639440		

표 4.4: 7개의 종료마디를 결정나무의 오분류표

		실제 범주		
		good	bad	합
예측 범주	good	229	56	285
	bad	34	46	80
		263	102	365

위험 추정값: 0.246575
위험 추정값의 SE: 0.0225605

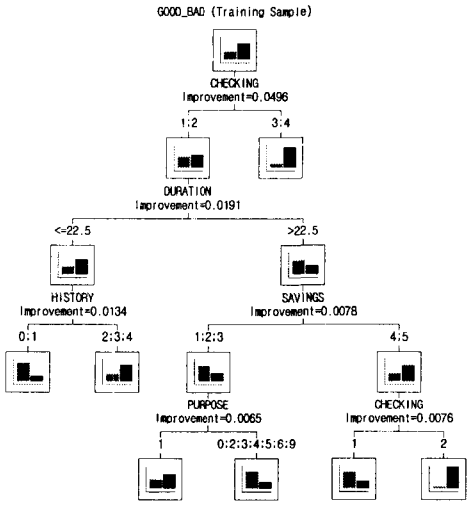


그림 4.2: 선택된 14개의 입력변수에 의해 구하여진 결정나무

값은 0.95 이상으로 5% 유의수준에서 독립성 검정의 가설을 기각한 변수들이기도 하다.

선택된 14개의 입력변수를 이용하여, 앞에서 설정한 정지규칙 및 가지치기 알고리즘을 적용하여 구하여진 결정나무는 그림 4.2와 같이 7개의 종료마디를 갖는 것으로 나타났으며, 검증자료에 대한 오분류표는 표 4.4와 같다.

표 4.2와 표 4.4를 비교하면 13개의 종료마디를 갖는 결정나무는 365개의 검증자료에서 89개를 오분류한 것에 비하여, 7개의 종료마디를 갖는 결정나무는 90개를 오분류하고 있어 단지 1개의 차이를 보이고 있다. 그러나 두번째 결정나무는 첫번째와 비교하여 현저하게 적은 수의 종료마디의 갖고 있어 규칙에 대한 이해가 쉬울뿐 아니라, 새로운 자료의 예측에서도 큰 차이를 보이지 않으므로 두번째 결정나무가 선호된다고 하겠다.

4.2. 사례 2

2002 한일월드컵이 끝난 후, 출전팀의 성적과 시합중에측정된 여러 가지 변수들의 측정값이 `http://worldcup.espnsoccernet.com/teamstats`에 발표되었다. 원 자료에는 출전국을 포함한 31개 변수로 구성되어 있으나 여기에서는 팀의 성적 및 전력과 관련된 26개의 변수를 추려 16강 진출 여부에 대해 어떠한 변수들의 영향력을 미치는지를 로지스틱 회귀분석(logistic regression analysis)에 의해 파악하려고 한다. 분석에 사용된 변수들을 요약하면 표 4.5과 같다. 각 변수들중 게임수와 관계된 변수는 게임수로 나누어 평균 1 게임당 횟수로 환산하였다. 반응변수인 `sixteen`은 16강 진출여부에 따라 0과 1의 값을 갖도록 설정하였다.

각 변수들은 월드컵에 출전한 32개국에 대해 관찰한 값으로 표본크기 $n = 32$이지만 모형에 사용될 변수가 26개이므로 반응변수와의 연관성을 δ'에 측정하고, 이를 토대로 일부 변수를 선택한 후, 선택된 변수에 의해 적절한 로지스틱 회귀모형을 작성하기로 한다. 측정

표 4.5: Fifa 자료

변수명	변수 설명	변수명	변수 설명
GL	Goals	OG	Opponent goals
AS	Assists	OA	Opponent assists
SH	Shots faced	OS	Opponent shots
SG	Shots on goal	OSG	Opponent shots on goal
SV	Saves	OSV	Opponent saves
PS	Penalty shots	OPS	Opponent penalty shots
PG	Penalty goals	OPG	Opponent penalty goals
CS	Clean sheets	OCS	Opponent clean sheets
FC	Fouls committed	OF	Opponent fouls
OFF	Offsides	OOFF	Opponent offsides
CK	Corner kicks	OCK	Opponent corner kicks
YC	Yellow cards	OYC	Opponent yellow cards
RC	Red cards	ORC	Opponent red cards

표 4.6: 측정된 반응변수와의 연관성 δ^t

변수명	δ^t	검정방법	변수명	δ^t	검정방법
SG	0.7499846	Kruskal-Wallis	OA	0.3291379	Kruskal-Wallis
CS	0.6230639	Pearson의 X^2	PS	0.2934405	Pearson의 X^2
OSV	0.6105098	Kruskal-Wallis	RC	0.2932178	Pearson의 X^2
OCS	0.5979098	Pearson의 X^2	YC	0.1973748	Kruskal-Wallis
SV	0.5935007	Kruskal-Wallis	OCK	0.1865103	Kruskal-Wallis
OSG	0.5229511	Kruskal-Wallis	OYC	0.1697424	Kruskal-Wallis
GL	0.486091	Kruskal-Wallis	OF	0.1489618	Kruskal-Wallis
AS	0.474295	Kruskal-Wallis	SH	0.09482866	Kruskal-Wallis
OG	0.4693986	Kruskal-Wallis	OS	0.07110845	Kruskal-Wallis
OPS	0.4342716	Pearson의 X^2	OOFF	0.04746226	Kruskal-Wallis
OPG	0.3625759	Pearson의 X^2	OFF	0.03393596	Kruskal-Wallis
ORC	0.3433114	Pearson의 X^2	CK	0.02372241	Kruskal-Wallis
PG	0.3425208	Pearson의 X^2	FC	0.02035033	Kruskal-Wallis

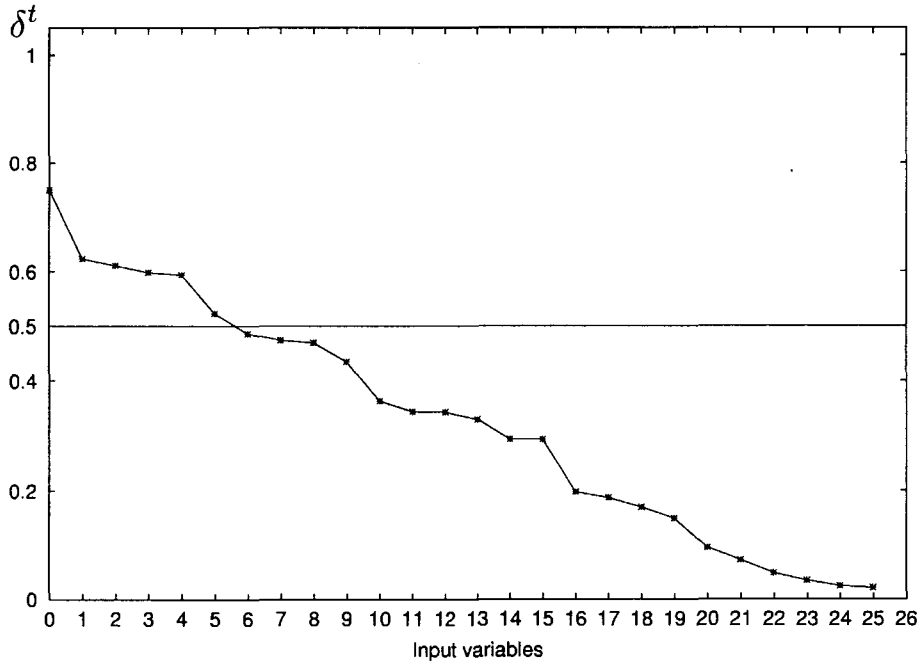


그림 4.3: 크기순으로 나열된 δ^t

표 4.7: 가능한 모든 변수를 대상으로 단계별 변수선택법에 의해 구하여진 최종모형

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-9.0441	3.3521	7.2796	0.0070
SG	1	0.4854	0.1837	6.9827	0.0082

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	46.361	23.848
SC	47.827	26.780
-2 Log L	44.361	19.848

표 4.8: 선택된 6개 변수에 의한 모형의 SAS 추정 결과

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-143.3	94.9223	2.2806	0.1310
SG	1	15.8177	10.5853	2.2330	0.1351
CS	1	-62.8425	52.1576	1.4517	0.2283
OSV	1	-13.5508	9.4041	2.0763	0.1496
OCS	1	32.9164	32.5138	1.0249	0.3114
SV	1	3.9258	3.0862	1.6181	0.2034
OSG	1	-1.4707	2.0792	0.5003	0.4794

된 δ^t는 크기순으로 나열하면 표 4.6와 같다. 표에 나타난 값으로 몇 개의 변수를 선택하여야 할 것인가는 역시 주관적인 판단에 의존하겠다. 그러나 이 경우 그림 4.3에서 보듯, 가장 큰 δ^t 값과 다음 크기의 δ^t의 값과의 차이가 큰 것으로 판단하여, 가장 큰 δ^t값을 갖는 변수 SG만을 선택하거나, $n = 32$임을 고려하여 너무 많은 수의 변수를 선택해서는 곤란함으로, 0.5 이상의 값을 갖는 여섯 개의 변수 SG, CS, OSV, OCS, SV, OSG를 포함한 로지스틱 회귀모형을 고려할 수 있다.

모형설정을 위한 첫번째 시도로 가능한 모든 변수들이 모형에 포함될 수 있도록 26개의 변수 모두를 대상으로 단계별 변수선택법에 의해 최종모형을 구하였다. SAS에 의해 구한 결과는 표 4.7와 같다. 전진선택과 후진제거과정의 검정은 SAS에서 기본으로 설정된 유의

표 4.9: δ^t에 의해 선택된 6개의 변수로부터 단계별 변수선택법에 의해 구하여진 최종모형

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-14.0739	5.5078	6.5295	0.0106
SG	1	1.4046	0.5733	6.0029	0.0143
OSV	1	-0.9454	0.4776	3.9189	0.0477

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	46.361	19.243
SC	47.827	23.641
-2 Log L	44.361	13.243

수준 값(SLENTRY=0.05, SLSTAY =0.05)을 그대로 사용하였는데, 모형 탐색과정에서 변수 SH가 모형에 추가되었었으나 후진제거과정에서 이 변수가 제거되어, 최종모형에는 SG 하나만이 들어 가게 되었다. 이렇게 구하여진 최종모형은 δ^t 값 사이의 차이가 큰 것을 고려하여 하나의 변수만을 선택하였을 때의 모형과 같다.

한편 표 4.8는 δ^t가 0.5 이상인 6개의 변수로 설정된 로지스틱 회귀모형의 추정 결과이다. 각 변수들의 회귀계수들에 대한 유의성 검정 결과는 모든 변수들이 그다지 유의하지 않은 것으로 판정되었다. 따라서 새로운 모형을 선택할 필요가 있다. 모형설정을 위한 두번째 시도는 δ^t-값의 비교를 통해 선택된 6개의 변수만을 대상으로 하여 최종모형을 구하는 것으로, 앞서와 같이 단계별변수선택법을 실시하였을 때 구하여진 최종모형은 표 4.9와 같다.

모형에 포함된 두 변수는 모두 유의수준 5%에서 유의한 것으로 판정되었으며, 모형추정에 사용된 자료를 로지스틱 회귀모형에 의해 반응변수를 분류한 결과는 두 모형에서 차이를 보이지 않았지만 표 4.7 모형과 표 4.9 모형의 아카이케 정보는 각각 23.848과 19.243으로, 이 결과는 δ^t에 의해 선택된 6개의 변수로부터 얻어진 최종모형이 전체변수를 대상으로 구한 최종모형보다 선호된다고 하겠다.

5. 결론

본 논문에서는 연속형 및 이산형자료들이 섞여있는 혼합자료에서 변수들간의 연관성을 비교할 수 있는 일반적인 연관성 측도를 제안하였다. 제안된 연관성 측도는 독립성 검정의 p 값을 이용한 것으로 시뮬레이션 결과에 의하면 제안된 연관성 측도는 서로 다른 유형의 변수들 간에도 일관되게 연관성을 비교할 수 있는 것으로 판정되었다.

요즈음 자료들은 대부분 매우 많은 수의 서로 다른 유형의 변수들을 포함한 혼합자료가 대부분으로 본 논문에서 제안된 연관성 측도들은 통계적 자료분석에서 중요한 역할을 담당할 수 있을 것으로 기대되며, 이러한 유용성의 예제로서 연관성 측도에 의해 선택된 변수들만으로 결정나무와 로지스틱 회귀모형을 수립하였을 때, 보다 좋은 결과를 얻을 수 있다는 것을 살펴 보았다.

참고문헌

[1] 박성현, 허문열. (1983). <전산통계>, 박영사.

[2] Bell, C. B. (1962). Mutual information and maximal correlation as measures of dependence, *Annal of Mathematical Statistics*, **33**, 587–595쪽.

[3] Bishop, Y. M. M. Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis-Theory and practive,* The MIT press, Cambridge, Massachusetts.

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

[5] Cramér, H. (1946). *Mathematical methods of statistics.* Princeton, New Jersey: Princeton University press.

[6] Cohen, J. (1960). A coefficient of agreement for nominal scales, *Educ. Psychol. Meas.* , **20**, 37–46.

[7] Eubank, R. L. Lariccia, V. N. and Rosenstein, R. B. (1987) Test statistics derived as components of Pearson's Phi–squared distance measure, *Journal of the American Statistical Association*, **82**, 816–825.

[8] Goodman, L. A. and Kruskal, W. H. (1954). Measure of association for cross classifications. *Journal of the American Statistical Association*, **49**, 732.

[9] Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, **30**, 81–93.

[10] Quinlan, J. R. (1988). *C4.5: Programs for machine learning.* Morgan Kaufmann Publishers, San Mateo, California.

[11] Randle, R., and Wolfe, D. (1979). *Introduction to the theory of Nonparametric statistics*, Wiley, New York.

[12] Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Tech. Journal*, **27**, 379–423 and 623–656.

[13] Silvey, S. D. (1964). On a measure of association. *Annals of Mathematical Statistics*, **35**,

1157–1166.

[14] Yule, G. U. (1912). On the methods of measuring association between two attributes (with discussion). *Journal of Royal Statistical Society*, **75**, 579–642.

[2002년 8월 접수, 2003년 1월 채택]

A Unified Measure of Association for Complex Data Obtained from Independence Tests

Seung-Chun Lee[1] Moon Yul Huh[2]

ABSTRACT

Although there exist numerous measures of association, most of them are lacking in generality in that they do not intend to measure the association between heterogeneous type of random variables. On the other hand, many statistical analyzes dealing with complex data sets require a very sophisticate measure of association. In this note, the p-value of independence tests is utilized to obtain a measure of association. The proposed measure of association have some consistency in measuring association between various types of random variables.

Keywords: Measure of association, Independence test, Complex data, p-value, Rank correlation coefficient, Kruskal-Wallis test, Pearson's X^2 test, Pearson's ϕ^2

1) Department of Statistics, Hanshin University.
E-mail: seung@hanshin.ac.kr
2) Department of statistics, Sung Kyun Kwan University.
E-mail: myhuh@skku.ac.kr