응용통계연구 제 **21**권 2호, *2008*년, *pp. 301–312*

SAMPLE-SPACING 방법에 의한 상호정보의 추정*

허문열[1] 차운옥[2]

요 약

상호정보(mutual information: MI)는 설명변수의 목적변수에 대한 예측정도를 나타내는 척도로서, 목적변수에 대한 설명변수의 중요도 순위를 구하거나 목적변수를 잘 설명해주는 설명변수의 집합을 구하는 변수선택문제에 유용하게 사용된다. 본 논문에서는 연속형 설명변수와 범주형 목적변수로 구성된 데이터로부터 결합확률분포를 추정하지 않고도 MI 추정량을 구할 수 있는 Sample-spacing 방법에 대한 연구를 수행하였다. 몬테칼로 모의실험과 실제 데이터에 대한 실험결과, MI 추정을 위해 Sample-spacing 방법을 사용할 때 $m = 1$을 사용하면 충분히 신뢰할만한 결과를 얻을 수 있다는 것을 알 수 있었다.

주요용어: 엔트로피, 상호정보, Sample-spacing.

1. 서론

상호정보(mutual information: MI)는 변수들 간의 상호의존성(interdependency)을 평가할 수 있는 척도로서 한 변수가 다른 변수에 대해 가지고 있는 정보의 양을 나타낸다. 두 변수가 서로 독립적이면 MI는 0이 되고, 두 변수가 서로 종속적이면 이 값이 커지게 된다. MI는 관련되어 있는 변수들의 형식이 연속형이거나 이산형인 경우에 다 사용할 수 있을 뿐만 아니라 비선형적인 관계에서도 사용할 수 있기 때문에 설명변수의 목적변수에 대한 예측정도를 나타내주는 척도로 많이 연구되고 있다. 목적변수와 각각의 설명변수간의 MI를 구하여, 목적변수에 대한 설명변수의 중요도 순위(variable ranking)를 정하거나 목적변수를 가장 잘 설명해주는 설명변수들의 집합을 구하는 변수선택(variable subset selection)에 사용할 수 있다. 임의의 두 가측공간(measurable space) $\mathcal{X}$와 $\mathcal{Y}$에 정의된 확률밀도 $P(X,Y)$의 MI는 다음과 같이 정의된다.

$$I(X;Y) = \int_{\mathcal{X}\times\mathcal{Y}} dP(x,y) \log \frac{dP(x,y)}{d(P(x) \times P(y))}, \tag{1.1}$$

여기서 $x \in \mathcal{X}, y \in \mathcal{Y}$이다. 확률변수의 불확실성(uncertainty)을 의미하는 엔트로피(entropy)를 h로 나타낼 때, 연속형 확률변수 X의 엔트로피는 다음과 같이 정의된다 (대개 이산형

* 본 연구는 2007년도 한성대학교 교내연구비 지원과제임.

1) (110-745) 서울시 종로구 명륜동 3가 53번지, 성균관대학교 통계학과, 교수.
E-mail: myhuh123@skku.edu
2) (136-792) 교신저자. 서울시 성북구 삼선동 3가 389, 한성대학교 공과대학 멀티미디어공학과, 교수.
E-mail: wcha@hansung.ac.kr

엔트로피는 H, 연속형 엔트로피는 소문자 h를 사용한다).

$$h(X) = -\int_{\mathcal{X}} dP(x) \log\ dP(x), \tag{1.2}$$

그러면 상호정보는 h를 사용하여 다음과 같이 나타낼 수 있다.

$$\begin{aligned} I(X;Y) &= h(X) + h(Y) - h(X,Y) \\ &= h(X) - h(X|Y) = h(Y) - h(Y|X). \end{aligned} \tag{1.3}$$

즉, X와 Y에 대한 MI는 변수들 간의 연관성척도이며, X가 알려져 있을 때 Y의 불확실성이 얼마나 감소되는지를 나타내는 정보의 양의 척도이다. 이는 기존의 상관계수와도 관련이 깊다. 공분산행렬 Σ_{XX}와 Σ_{YY}를 가지는 다변량 정규 확률변수 X와 Y에 대해 다음 식이 성립한다 (Cover와 Thomas, 1991).

$$\begin{aligned} I(X;Y) &= -\frac{1}{2}\log\left(\frac{|\Sigma|}{|\Sigma_{XX}||\Sigma_{YY}|}\right) \\ &= -\frac{1}{2}\sum_i \log\left(1-\rho_i^2\right), \end{aligned} \tag{1.4}$$

여기에서 Σ는 행벡터 $(X;Y)$의 공분산행렬이고 ρ_i는 상관계수이다. 이변량의 경우 MI는 $-1/2\log(1-\rho_i^2)$와 같다.

확률변수 X가 연속형일 때 엔트로피 $h(X)$를 미분 엔트로피(differential entropy)라 한다. 이 경우 MI를 추정하기 위해서는 결합확률분포를 추정해야 하는 등 많은 문제점이 있다. 미분 엔트로피를 추정하기 위해 모수적 (Lazo와 Rathie, 1978; Ahmed와 Gokhale, 1989; Hulle, 2002), 비모수적 (Dmitriev 등, 1973; Ahmad와 Lin, 1976; Tsybakov와 Meulen, 1994; Mokkadem, 1989; Joe, 1989)인 다양한 방법들이 연구되었는데 이러한 연구방법에서는 비모수적 커널추정치(kernel estimator)나 Parzon window 추정치 (Silverman, 1986)를 구해 식 (1.2)에 적용한다. Beirlant 등 (1997)과 Brillinger (2004)에는 엔트로피 추정방법과 그 통계적 성질에 대해 잘 정리되어 있다. 이러한 다양한 연구에도 불구하고 연속형 데이터에서 MI를 추정하는 데는 여러 가지 문제점이 있다. 첫째로 포본 크기가 작거나 표본크기는 크다고 하더라도 확률변수가 가지는 값의 종류가 적을 때 생길 수 있고, 두 번째로는 확률변수의 분포가 꼬리 부분이 아주 길거나 데이터가 특이값(outlier)을 가지고 있을 때 수치적분 과정에서 생길 수 있다. 미분 엔트로피를 추정할 때 생길 수 있는 이러한 문제점들을 해소시키기 위해 이산화(discretization) 방법과 Sample-spacing 방법을 사용할 수 있다. 본 논문에서는 설명변수가 연속형이고 목적변수가 범주형일 때, 결합밀도함수를 추정하지 않고도 MI 추정량을 구할 수 있는 Sample-spacing 방법에 대한 연구를 수행한다. 본 논문의 구성은 다음과 같다. 제 2장에서는 Sample-spacing 방법으로 MI를 추정하는 방법에 대해 정리하고, 제 3장에서는 몬테칼로 모의실험 과정과 실험 결과를 분석하였다. 제 4장에서는 실제 데이터에 대한 실험을 통해 Sample-spacing 방법에 대한 문제점과 해결방안을 제시하고 제 5장에 결론을 기술하였다.

2. Sample-spacing 방법

$x_1, x_2, \ldots, x_n$을 연속형 확률변수 X의 표본이라 가정했을 때, $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$을 대응하는 순서통계량이라 하자. 이때, $x_{(i+m)} - x_{(i)}$ $(1 \le i < i+m \le n)$를 m차 spacing 또는 m-spacing이라 한다. 다양한 Sample-spacing 방법으로 MI를 추정할 수 있으며 (Brillinger, 2004). Miller와 Fisher (2003)의 ICA(Independent Component Analysis) 문제에서 추정된 엔트로피의 m-spacing 추정량은 다음과 같다,

$$\hat{h}(X) = \frac{1}{n-m} \sum_{i=1}^{n-m} \log\left(\frac{n+1}{m} \left(x_{(i+m)} - x_{(i)} \right) \right). \tag{2.1}$$

범주형 변수 Y가 이진값 1과 2를 가질 때, $I(X;Y)$는 다음과 같이 추정한다.

$$\begin{aligned} \hat{I}(X;Y) &= \hat{h}(X) - \hat{h}(X|Y) \\ &= \hat{h}(X) - \hat{p}\hat{h}(X|Y=1) - (1-\hat{p})\hat{h}(X|Y=2), \end{aligned} \tag{2.2}$$

여기서 $\hat{p} = \hat{P}[Y=1]$이다. Miller와 Fisher (2003)는 m값이 커질수록 Sample-spacing 추정량의 분산이 기대값에 비해 작아지기 때문에 $m = \sqrt{n}$을 사용하는 것이 좋다고 하였다. 다음 장에서는 몬테칼로 방법을 사용하여 어떤 m이 가장 적절한가를 밝히고자 한다. 참고로 Sample-spacing 방법은 순서통계량에 기반을 두기 때문에 1차원에서만 적용가능하다.

3. 몬테칼로 모의실험

3.1. 데이터 생성 및 실험방법

X가 연속형, Y가 범주형 변수일 때 MI 추정에 영향을 주는 요소는 X와 Y의 분포, 표본크기 n과 Sample-spacing 모수 m이다. 본 실험에서는 범주형 변수 Y가 이진값 1과 2를 가지는 경우만 생각한다. $p = P[Y=1]$라 하고, $f(x)$를 X의 밀도함수라고 하면 혼합분포는 $pf(x|Y=1) + (1-p)f(x|Y=2)$와 같이 나타낼 수 있다. 다음과 같이 두 가지 X의 분포에 대해 실험한다.

$$\text{[분포 1]} \qquad pN(0,1) + (1-p)N(\mu, \sigma^2), \tag{3.1}$$

$$\text{[분포 2]} \qquad pGamma(\lambda_1, r=1) + (1-p)Gamma(\lambda_2, r=1), \tag{3.2}$$

여기에서 $N(\mu, \sigma^2)$는 정규분포이고, $G(\lambda, r=1)$은 형태(shape) 모수 λ와 크기(scale) 모수 r을 갖는 감마분포이다. 편의상 식 (3.1)의 혼합정규분포는 $Nm(p, \mu, \sigma)$로, 식 (3.2)의 혼합감마분포는 $Gm(p, \lambda_1, \lambda_2)$로 나타내기로 한다. 각 분포의 경우 다음과 같은 모수 값을 실험에 사용한다.

$$\begin{aligned} &\text{[분포 1]} \quad \mu = 5; \qquad \sigma = 1, 3, 5, \\ &\text{[분포 2]} \quad \lambda_1 = 1, 3, 5; \quad \lambda_2 = 10. \end{aligned}$$

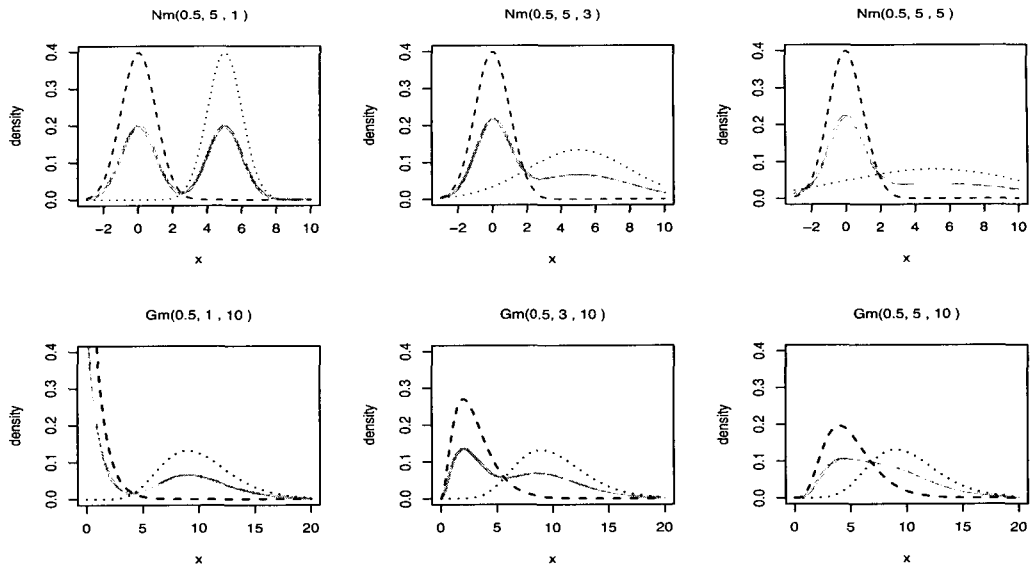

그림 3.1: 모수 값에 따른 혼합정규분포와 혼합감마분포

그림 3.1에는 $p = 0.5$인 경우 각각의 모수 값에 따른 분포를 나타내었다. 그림에서 대시(-)선은 $f(x|Y = 1)$, 점선은 $f(x|Y = 2)$, 굵은 회색선은 혼합밀도함수 $1/2f(x|Y = 1) + 1/2f(x|Y = 2)$를 나타낸다. 위의 세 분포는 Y의 값이 주어졌을 때의 조건부확률분포를 나타낸 혼합정규분포이고, 아래의 세 분포는 혼합감마분포이다. 실험을 위해서 각각의 분포를 따르는 데이터를 필요에 따라 생성한다. 본 실험에서는 표본크기 n과 m에 대해서 다음과 같은 경우를 고려하였다.

$$n : 50, 100, 150, 200, 300, 400, 500, 1000,$$
$$m : 1, 2, 3, 4, 5, 10, 20.$$

모의실험은 각 경우에 대해 500번씩 반복하여 수행하였으며, 실험 결과들을 이론적인 MI 값과 비교분석하였다. 이론적인 값은 다음과 같이 구하였다.

1. $Nm(p, \mu, \sigma)$ 분포의 경우:

   $h(N(\mu, \sigma^2)) = 1/2 \log(2\pi e\sigma^2)$이 성립하므로 (Lazo와 Rathie, 1978), X가 $Nm(p, \mu, \sigma)$를 따르는 경우 MI는 다음과 같다.

$$\mathrm{MI} = h(Nm(p, \mu, \sigma)) - h_1 - h_2, \tag{3.3}$$

   여기에서,

$$h_1 = \frac{1}{2}\log(2\pi e), \quad h_2 = \frac{1}{2}\log(2\pi e\sigma^2), \tag{3.4}$$

$$h(Nm(p,\mu,\sigma)) = \int f(x)\log f(x)dx, \tag{3.5}$$

$$f(x) = pN(0,1) + (1-p)N(\mu,\sigma^2) \tag{3.6}$$

이다. 식 (3.6)의 엔트로피를 구하기 위해서는 수치적분을 해야 하는데, 수치적분은 R의 함수 integrate를 사용한다.

2. $Gm(p,\lambda_1,\lambda_2)$ 분포의 경우:

역시 Lazo와 Rathie (1978)에 의해, $h(G(\lambda,r=1)) = \log\Gamma(\lambda) + (1-\lambda)\psi(\lambda) + \lambda$, $\psi(\lambda)$는 digamma 함수로서 $\psi(z) = d/dz\psi(z)$이다. 따라서, MI는 다음과 같이 계산할 수 있다.

$$\text{MI} = h(Gm(p,\lambda_1,\lambda_2)) - h_1 - h_2, \tag{3.7}$$

여기에서,

$$h_1 = \log\Gamma(\lambda_1) + (1-\lambda_1)\psi(\lambda_1) + \lambda_1, \tag{3.8}$$

$$h_2 = \log\Gamma(\lambda_2) + (1-\lambda_2)\psi(\lambda_2) + \lambda_2, \tag{3.9}$$

$$h(Gm(p,\lambda_1,\lambda_2)) = \int f(x)\log f(x)dx, \tag{3.10}$$

$$f(x) = pG(\lambda_1,r=1) + (1-p)G(\lambda_2,r=1) \tag{3.11}$$

이다.

3.2. 모의실험 결과

여러 가지 모수 조합에 대해 실험하여본 결과를 그림 3.2에 나타내었다. 표본크기를 $(50,100,150,\ldots,1000)$으로 변화시켜가며 각 그림의 x-축에 나타내었고, y-축에는 MI 추정값을 나타내었다. y-축의 범위는 (이론적인 MI 값 -0.1, 이론적인 MI 값 $+0.1$)이다. 여러 개의 선은 $m\,(1,2,3,4,5,10,20)$에 해당하는 MI 추정량이며, 가운데 두꺼운 회색 직선이 이론적인 값이고 두꺼운 실선은 $m=1$인 경우이다. 이 그림에서 알 수 있는 바와 같이 값이 커질수록 이론적인 값에서 멀어지고, 이러한 추이는 모든 경우에서 동일하다는 것을 알 수 있다. 또한 Sample-spacing 추정량은 모든 경우에 이론적인 값보다 큰 값으로 추정되고 있고(overestimation), 특히 $Nm(0.5,5,5)$에서 매우 과다하게 추정되는 것을 알 수 있다. 그러나 이 경우는 $N(0,1)$과 $N(5,5)$의 두 집단이 혼합되어있는 것으로서 (그림 3.1 참고) 두 집단의 구분이 매우 어려운 경우이다. Sample-spacing의 과다 추정 문제는 식 (2.3)–(2.5)의 엔트로피 추정식에서 상수부분을 적절히 조정하면 가능할 것이다. 본 논문에서는 이 문제를 핵심으로 다루지 않기 때문에 이 문제에 대해서는 더 이상 고려하지 않기로 한다. 중요한 것은 모든 경우에 공통적으로 $m=1$이 이론적인 값에 가장 가까운 것을 알 수 있다.

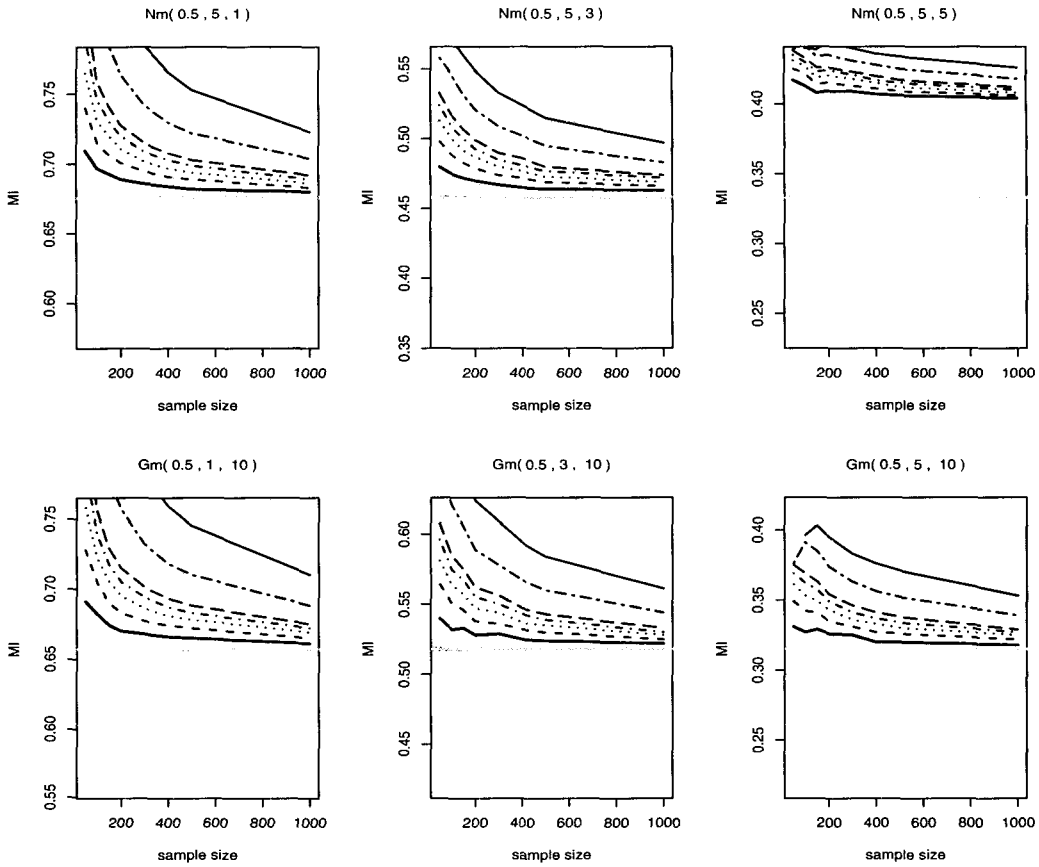

그림 3.2: 정규혼합과 감마혼합 두 가지 분포에서 여러 모수조합에 대한 MI 추정량

표 3.1에는 $Nm(0.5, 5, 3)$분포에 대해서, $m = 1, 2, 3, 5, 10, 20$을 사용하여 Sample-spacing 방법으로 구한 MI 추정값의 표준편차를 정리하였다. 표본의 크기 $n = 100, 200, 300, 500, 1000$을 사용하였으며 반복회수는 500번이다. 다른 분포의 경우도 이와 유사한 결과를 보였다.

Sample-spacing 방법에서는 표 3.1에서 알 수 있는 바와 같이 표본의 크기가 300이하 정도로 작을 때는 m값을 점점 크게 하면 표준편차는 조금 줄어들었다가 오히려 증가하고, 표본의 크기가 클 때는 큰 m값을 사용하면 표준편차는 조금씩 줄어들지만 정도는 미미하다. $n = 1000$인 경우, $m = 1$에서 MI 추정량의 표준편차는 0.0067이고 $m = 20$에서 0.0051이다. $Nm(0.5, 5, 3)$ 분포의 이론적인 MI 값은 그림 3.2에 의하면 0.46 정도이므로, $m = 1$일 때의 표준편차 0.0067에 대한 표본변동계수는 $0.0067/0.46 \simeq 0.014$ (1.4%)이다. 따라서 이 정도의 표준편차이면 충분히 신뢰할만한 MI 추정량을 제공하고 있다고 할 수 있다. 이상의 모의실험 결과를 가지고 볼 때, 연속형 설명변수와 범주형 목적변수로 구성된 데이터에 대해 Sample-spacing 방법으로 MI를 추정하면 대부분의 경우 표본크기가 500 이상이면 이론적

표 3.1: MI 추정값에 대한 표준편차

		n				
		100	200	300	500	1000
SS	$m=1$	0.022	0.017	0.0120	0.0089	0.0067
	$m=2$	0.020	0.015	0.0100	0.0080	0.0062
	$m=3$	0.020	0.014	0.0096	0.0076	0.0059
	$m=5$	0.021	0.013	0.0093	0.0071	0.0057
	$m=10$	0.025	0.014	0,0096	0.0066	0.0053
	$m=20$	0.039	0.018	0.0110	0.0069	0.0051

인 값과의 차이가 0.01 (표본변동계수 0.015) 이내에 있으며, 표본크기가 100 이하로 작더라도 0.03 (표본변동계수 0.05) 이내에 있는 것을 알 수 있었다. 따라서 $m=1$을 사용하는 것이 모든 경우에서 가장 적합하다고 할 수 있다. 이 결과는 과거의 기존 연구와 매우 다르다는 것을 알 수 있다 (예: Miller와 Fisher, 2003). 다음 장에서는 실제 자료를 사용하여 m 선택에 대한 문제를 다시 고려해 보기로 한다.

4. 실제 데이터에 대한 실험

모든 설명변수들이 연속형 값을 가지고 클래스의 개수가 2인 데이터 WBCD를 UCI 창고(machine learning repository (Blake와 Merz))에서 구하였다. WBCD는 569명의 유방암 환자에 대한 데이터로서 30개의 연속형 설명변수와 1개의 목적변수 (양성(benign) 환자: 357, 악성(malignment) 환자: 212)로 구성되어 있다. $m=1$을 사용하여 각 변수와 목적변수 사이의 MI를 구했을 때 MI 값에 따른 변수의 순서와 해당 MI 값은 다음과 같다.

변수의 순서:

V2 V12 V22 V15 V17 V20 V16 V30 V26 V13 V11 V1 V14 V29 V4 V27 V6 V3 V7 V23 V21 V10 V24 V19 V8 V5 V18 V28 V9 V25

해당 MI x 1000:

80 91 91 98 180 190 190 260 270 280 300 300 300 310 310 340 340 350 360 360 370 400 400 420 430 430 430 460 470 490

이렇게 구한 MI의 타당성을 살펴보기 위해, DAVIS (Huh, 2005)를 사용하여 MI의 크기로 변수를 재배열하여 평행좌표계를 그리고 $Y=1$과 2에 대해 붉은 색과 푸른색으로 표현한 것이 그림 4.1에 나타나 있다. 이 그림을 보면 $m=1$에 의한 MI 추정량은 그 유효성이 매우 의심된다. 실제로 MI가 가장 큰 변수 (V25, MI = 0.490)에 대해 각 집단의 기술통계량을 구한 것이 표 4.1과 같다. 이 결과에 의하면 V25의 분포가 $Y=1$과 $Y=2$의 두 집단에서 별로 차이가 없는 것을 알 수 있다. 왜 이런 문제가 발생하는가? V25의 값을 살펴보면 같은 값을 가지는 것이 전체 569개 데이터 중에서 159개나 된다. m-spacing 추정량은 m차 차분을 택하므로 $m=1$일 때 같은 값이 나타나면 이 값들은 추정에서 제외된다. 실제로는

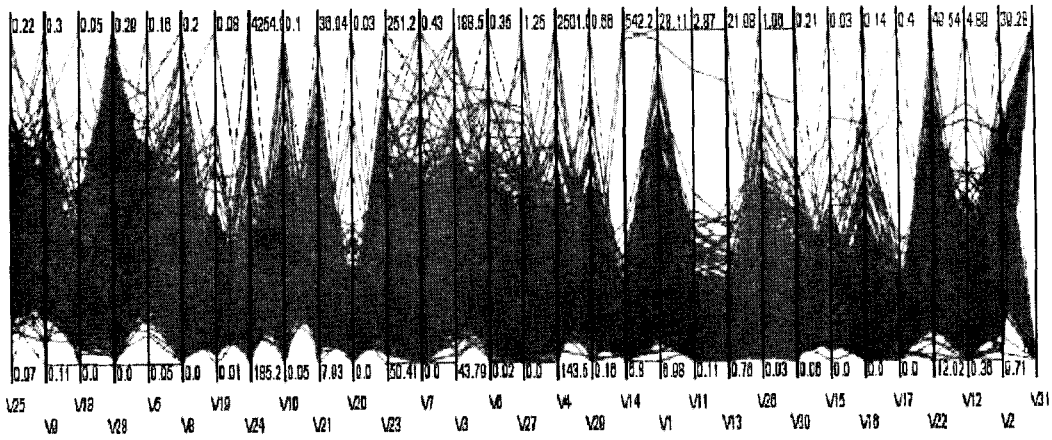

그림 4.1: 연속형 설명변수 30개를 이진 목적변수 V31 (가장 오른편)에 대해 $m = 1$을 사용하여 MI를 구하고 크기 순서대로 (가장 왼편이 가장 큰 MI) 재배열한 결과

표 4.1: V25 변수의 Y=1과 Y=2 집단에서의 기술통계량

	최소값	하4분위수	중앙값	평균	상4분위수	최대값
Y=1	0.071	0.110	0.125	0.125	0.137	0.200
Y=2	0.088	0.130	0.143	0.144	0.156	0.222

같은 값이 나타나는 경우에는 확률밀도추정량을 구할 때 이들은 매우 높은 가중값을 가져야 한다. V25의 2차 차분을 택하는 $m = 2$인 경우를 살펴보면 2차 차분에서도 36개의 0이 나타났다. 3차 차분에서는 9개의 0이 그리고 4차 차분에서는 0이 나타나지 않았다. 따라서 실제 데이터의 경우, 측정 등의 문제로 인해 데이터가 같은 값을 갖는 경우가 많이 나타나므로 $m = 1$이나 $m = 2$를 사용하면 매우 위험할 수 있다. 이를 보완하는 방법으로는 다음 두 가지를 생각할 수 있다.

1. $m = 3$ 이상을 사용한다.
2. $m = 1$을 사용하되, 값을 jittering 시킨다.

첫 번째 방법을 사용하더라도 같은 값이 나타나는 위험성을 배제할 수 없다. 그러므로 2번 방법을 사용하는 것이 더 타당하다고 할 수 있다. jittering 방법은 다음과 같다.

$$x^* = x + \frac{1}{100} R\bar{x}, \tag{4.1}$$

여기서 R은 (0,1) 사이에서 택한 난수이고, $\bar{x}$는 변수가 가지는 값의 평균이다. $m = 1$인 경우 jittering 방법을 사용하고, $m = 2, 3, 4$를 가지고 데이터를 다시 분석하여 MI가 큰 순서대로 중요변수를 정리하면 표 4.2와 같다. $m^* = 1$은 $m = 1$을 jittering 방법으로 조정한 것을 나타내고, =기호는 동일한 MI 추정량이 얻어진 것을 나타낸다.

표 4.2: 중요변수 순위

	변수 순위
$m* = 1$	24, 23 = 21, 28, 8 = 3, 7, 1, 4, 27
$m = 2$	23 = 24, 21, 28, 8, 3, 1, 7, 4, 27
$m = 3$	23, 21 = 24, 28, 8, 3 = 4, 7, 1, 27
$m = 4$	23, 24, 21 = 28, 8, 3, 7, 1, 4, 27

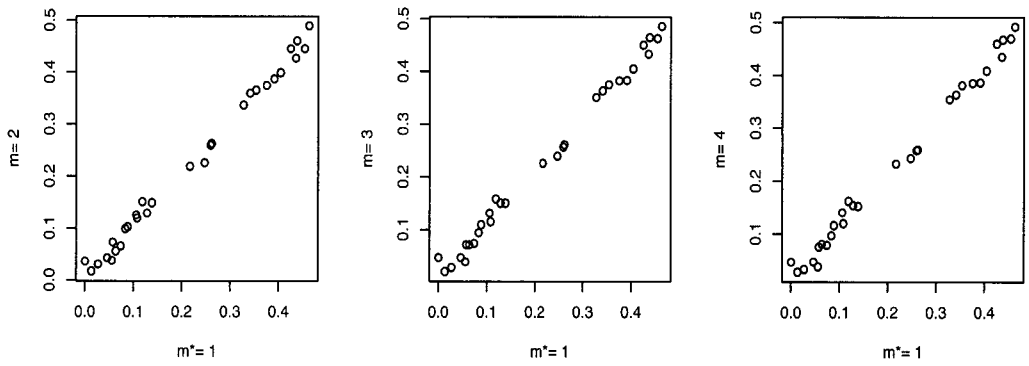

그림 4.2: Jittering에 의해 자료를 변화시키고 x-축에는 $m^* = 1$, y-축에는 $m = 2, 3, 4$에 의한 추정량을 그려본 결과

표 4.2를 보면 Sample-spacing 방법에서 $m^* = 1$로 했을 때, $m = 2, 3, 4$로 했을 때와 순위는 약간 다르지만 중요변수 10개를 다 잘 구해주는 것을 알 수 있다. 또, $m^* = 1$에 의해 구한 MI와 $m = 2, 3, 4$에 의한 MI 값을 산점도로 나타낸 것이 그림 4.2에 주어져 있다. 이 그림들을 보면 $m^* = 1$이 적절한 방법임을 다시 확인할 수 있다. 참고로 V25 변수의 MI는 0.068로 나타났는데 $m = 1$일 때 0.490인 것과 비교하면 큰 차이이다. 이상의 내용을 모의실험의 결과와 종합하면 다음과 같다. 즉, 연속형 설명변수와 범주형 목적변수로 구성된 데이터에서 MI 추정량을 구해 중요변수를 선택하기 위해서 Sample-spacing 방법을 사용하는 경우, Miller와 Fisher (2003)는 표본의 크기가 클 때 큰 값을 사용하도록 추천하였으나 $m = 1$을 사용하여도 충분히 신뢰할만한 결과를 얻을 수 있다.

5. 결론

본 논문에서는 연속형 설명변수와 범주형 목적변수로 구성된 데이터에서 결합확률분포의 추정없이 MI 추정량을 구하기 위한 Sample-spacing 방법에 대해 연구하였다. 몬테칼로 모의실험과 실제 데이터에 대한 실험 결과, Sample-spacing 방법으로 MI 추정량을 구해 중요변수를 선택할 때 $m = 1$을 사용하면 매우 효율적이라는 것을 알 수 있었다. 특히 모든 모

의실험에서 MI 추정량과 이론적인 MI 값은, 표본크기가 작지 않을 때 차이가 0.01이내이고 상대오차 (오차/실제값)가 1.5% 이내, 표본크기가 크지 않을 때에도 상대오차가 5% 이내인 것을 알 수 있었다. 그러나 실제 데이터에 대한 실험 결과를 분석해 보면 연속형 데이터에서 같은 값이 나타나는 경우가 많을 때에는 엉뚱한 결과가 나타날 수 있었다. 이를 보완하는 방법으로 jittering 방법을 사용하면 매우 신뢰할 수 있는 결과를 얻을 수 있다. 본 논문에서는 Sample-spacing 방법으로 목적변수와 각각의 설명변수간의 MI를 구하여 목적변수에 대한 설명변수의 중요도 순위를 구하는데 중점을 두었다. 설명변수들 간의 상관관계를 고려하여 목적변수에 대한 중요변수 집합을 구하려면 다른 MI 추정방법을 사용하는 것이 바람직하다.

참고문헌

Ahmad, I. A. and Lin, P. E. (1976). A nonparametric estimation of the entropy for absolutely continuous distribution, *IEEE Transactions on Information Theory*, **22**, 372–375.

Ahmed, N. A. and Gokhale, D. V. (1989). Entropy expressions and their estimators for multivariate distribution, *IEEE Transactions on Information Theory*, **35**, 688–692.

Beirlant, J., Dudewicz, E. J., Györfi, L. and Meulen, E. (1997). Nonparametric entropy estimation: An overview, *International Journal of Mathematical and Statistical Sciences*, **6**, 17–39.

Blake, C. and Merz, C. J. UCI machine learning repository, http://www.ics.uci.edu/ mlearn /MLRepository.

Brillinger, D. R. (2004). Some data analysis using mutual information, *Brazilian Journal of Probability and Statistics*, **18**, 163–183.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*, John Wiley & Sons, New York.

Dmitriev, G., Yu, G. and Tarasenko, F. P. (1973). On the estimation of functionals of the probability density and its derivatives, *Theory of Probability and Its Applications*, **18**, 628–633.

Huh, M. Y. (2005). *DAVIS(http://stat.skku.ac.kr/myhuh/DAVIS.html).*

Joe, H. (1989). Estimation of entropy and other functionals of a multivariate density, *Annals of Institute of Statistical Mathematics*, **41**, 683–697.

Lazo, A. V. and Rathie, P. (1978). On the entropy of continuous probability distributions, *IEEE Transactions on Information Theory*, **24**, 120–122.

Miller, E. G. L. and Fisher III, J. W. (2003). ICA using spacings estimates of entropy, *The Journal of Machine Learning Research*, **4**, 1271–1295.

Mokkadem, A. (1989). Estimation of the entropy and information of absolutely continuous random variables, *IEEE Transactions on Information Theory*, **35**, 193–196.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman & Hall/CRC, London.

Tsybakov, A. B. and van der Meulen, E. C. (1994). Root-n consistent estimators of entropy for densities with unbounded support, *Scandinavian Journal of Statistics*, **23**, 75–83.

van Hulle, M. M. (2002). Multivariate edgeworth-based entropy estimation, *Neural Computation*, **14**, 1887–1906.

[2007년 10월 접수, 2008년 1월 채택]

Sample-spacing Approach for the Estimation of Mutual Information*

Moon Yul Huh[1] Woon Ock Cha[2]

ABSTRACT

Mutual information is a measure of association of explanatory variable for predicting target variable. It is used for variable ranking and variable subset selection. This study is about the Sample-spacing approach which can be used for the estimation of mutual information from data consisting of continuous explanation variables and categorical target variable without estimating a joint probability density function. The results of Monte-Carlo simulation and experiments with real-world data show that $m = 1$ is preferable in using Sample-spacing.

Keywords: Entropy, mutual information, sample-spacing,

* This research was financially supported by Hansung University in the year of 2007.

1) Professor, Dept. of Statistics, Sungkyunkwan University, Seoul, 110-745, Korea.
E-mail: myhuh123@skku.edu

2) Corresponding author. Professor, Dept. of Mutimedia Engineering, Hansung University, Seoul 136-792, Korea.
E-mail: wcha@hansung.ac.kr