Communications of the Korean Statistical Society Vol. 15, No. 6, 2008, pp. 977–991

상호정보 추정을 위한 k-최근접이웃 기반방법†

차운옥[1], 허문열[2]

요 약

본 논문에서는 연속형 변수에 대한 결합확률분포를 추정하지 않고도 상호정보(MI) 추정량을 구할 수 있는 k-최근접이웃 기반방법에 대하여 연구하였다. 변수가 동일한 값들을 가지는 경우 k-최근접이웃을 구할 때 생기는 문제점을 해결하기 위하여 지터링(jittering)과 붓스트랩(bootstrap) 방법을 제안하였다. 몬테칼로 모의실험과 실제 데이터에 대한 실험을 수행한 결과, $k = 1$과 같이 작은 값을 사용한 k-최근접이웃 기반방법에 의해 효율적인 MI 추정량을 구할 수 있었다. k-최근접이웃 기반방법은 연속형 설명변수, 범주형 또는 연속형인 목적변수 형태의 데이터에 적용할 수 있으며, 목적변수에 영향을 주는 중요한 설명변수의 순서를 구할 수 있을 뿐만 아니라 다차원에도 적용할 수 있기 때문에 중요변수의 집합을 구하는 변수 선택(feature subset selection) 문제에도 적용할 수 있다.

주요용어: 상호정보; k-최근접이웃; 지터링; 붓스트랩.

1. 서론

임의의 연속형 변수 사이의 상호정보 MI 는 다음과 같이 정의된다.

$$I(X;Y) = \int_Y \int_X p(x,y) \log \left(\frac{p(x,y)}{p_X(x) p_Y(y)} \right) dxdy, \tag{1.1}$$

여기서 $p(x,y)$는 X와 Y의 결합확률밀도함수이고, $p_X(x)$와 $p_Y(y)$는 각각 X와 Y의 주변확률밀도함수이다. 또한 엔트로피(entropy)는 확률변수의 불확실성(uncertainty)을 의미하며, 연속형 확률변수 X의 미분 엔트로피(differential entropy) h는 다음과 같이 정의된다.

$$h(X) = -\int_X p_X(x) \log p_X(x) dx. \tag{1.2}$$

MI는 $h(X)$를 사용하여 다음과 같이 나타낼 수 있다.

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

† 본 연구는 2008년도 한성대학교 교내연구비 지원과제임.

1) (136-792) 서울시 성북구 삼선동 3가 389, 한성대학교 공과대학 멀티미디어공학과, 교수.
교신저자: wcha@hansung.ac.kr

2) (110-745) 서울시 종로구 명륜동 3가 53번지, 성균관대학교 통계학과, 교수.

MI는 X가 알려져 있을 때 Y의 불확실성이 얼마나 감소되는지를 나타내는 정보의 양의 척도로서 변수들 간의 연관성을 측정해주는 척도로 이용될 수 있다. 두 변수가 서로 독립적일 때 MI는 0이 되고, 두 변수가 서로 종속적이면 이 값은 커지게 된다. MI는 관련되어 있는 변수들이 가지는 값의 형식이 연속형이거나 범주형인 경우에 다 사용할 수 있고 비선형적인 관계에서도 사용할 수 있기 때문에 설명변수의 목적변수에 대한 예측정도를 나타내주는 척도로 많이 연구되고 있다. 이는 기존의 상관계수와도 관련이 깊다. Beirlant 등 (1997)과 Brillinger (2004)에는 다양한 엔트로피 추정방법과 그 통계적 성질에 대해 잘 정리되어 있다. 미분 엔트로피를 사용하여 MI를 추정하는 경우에는 결합확률밀도함수를 추정해야 하는 등 많은 문제점이 있다. 연속형 변수의 결합확률밀도함수를 추정하지 않고도 MI 추정량을 구할 수 있는 방법으로는 이산화(discretization)방법 (Cha와 Huh, 2005), 표본간격(Sample-spacing) 방법 (Miller와 Fisher, 2003)과 k-최근접이웃 기반방법 (Kraskov, 2004)이 있다. 이산화방법에서는 연속형 변수를 이산형 변수로 이산화 시켜서 MI를 추정한다. 표본간격 방법은 설명변수가 연속형이고 목적변수가 범주형일 때 사용할 수 있으며, 표본간격 방법에서의 모수 m에 관한 연구 (허문열과 차운옥, 2008)에서는 m=1일 때 MI 추정에서 좋은 결과를 얻을 수 있음을 보였다. 표본간격 방법은 설명변수가 연속형일 때 목적변수가 범주형인 경우에만 사용할 수 있으므로, 본 논문에서는 목적변수가 범주형, 연속형에 상관없이 MI 추정에 사용할 수 있는 k-최근접이웃 기반방법에 대한 연구를 수행하였다. 다양한 모의실험을 통해 k값에 대한 분석을 하였으며, 목적변수가 범주형인 경우에는 표본간격 방법과 비교하였다. 본 논문의 구성은 다음과 같다. 제 2장에서는 지터링(jittering)과 붓스트랩(bootstrap), 제 3장에서 k-최근접이웃 기반방법으로 MI를 추정하는 방법에 대해 정리하였고, 제 4장에서는 몬테칼로 모의실험 과정과 실험 결과를 분석하였다. 제 5장에서는 실제 데이터에 대한 실험과 실험결과 분석, 제 6장에 결론을 기술하였다.

2. 지터링과 붓스트랩

실제 데이터의 경우 변수가 연속형 값을 가지더라도 여러 개의 관측 값이 동일한 경우가 많이 있다. 예를 들어 Fisher의 붓꽃(Iris) 데이터에서 변수 꽃잎너비(petal width)를 살펴보기로 한다. 그림 2.1에 꽃잎너비와 꽃받침너비(sepal width)의 산점도가 주어져 있다. 이 그림에서 볼 수 있는 바와 같이 꽃잎너비의 경우에 여러 개의 관측 값이 같은 값으로 나타나고 있다. 동일한 값이 여러 개이면 k-최근접이웃 기반방법으로 MI를 추정할 때, 여러 k에 대해 k-최근접이웃의 거리가 0이 되므로 문제가 된다. 본 논문에서는 이러한 문제점을 해결하기 위하여 지터링 방법을 제안한다.

X가 연속형 변수일 때 지터링은 다음과 같이 한다.

$$X \leftarrow X + \epsilon \cdot \sigma_X \cdot z, \tag{2.1}$$

여기서, ϵ은 작은 값(예를 들어 10^{-5})이며, σ_X는 X의 표준편차, z는 $N(0,1)$에서 택한 하나의 난수이다.

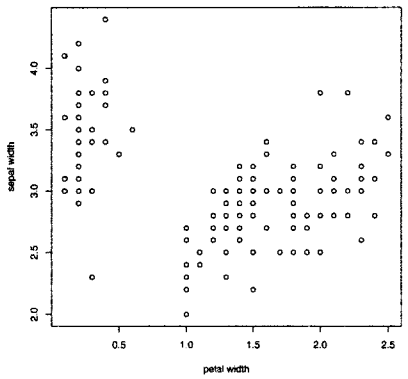

그림 2.1: 꽃잎너비와 꽃받침너비의 산점도

지터링은 난수 발생에 따라 결과가 많이 좌우되기 때문에 이러한 문제점을 해결하기 위하여 붓스트랩 방법을 사용한다. 붓스트랩 방법에서는 여러 번(예를 들어 1000번) 지터링을 수행하고 수행결과의 평균값을 활용하기 때문에 지터링에서 발생하는 결과의 랜덤성을 제거할 수 있다. 지터링과 붓스트랩 방법을 실제 데이터에 적용하였을 때 k-최근접이웃 기반방법은 MI 추정에 매우 효율적인 것으로 나타났다. 자세한 내용은 5장에서 다루기로 한다.

3. k-최근접이웃 기반방법에 의한 상호정보의 추정

연속형 확률변수 X, Y에 대해서, k-최근접이웃 기반방법에 의한 엔트로피와 MI의 Kozachenko-Leonenko 추정량은 다음과 같다 (Staugbauer 등, 2004).

$$\hat{h}(X) = -\psi(k) + \psi(n) + \log c_{d_X} + \frac{d_X}{n}\sum_{i=1}^{n}\log\epsilon(i), \tag{3.1}$$

$$\hat{h}(Y) = -\psi(k) + \psi(n) + \log c_{d_Y} + \frac{d_Y}{n}\sum_{i=1}^{n}\log\epsilon(i), \tag{3.2}$$

$$\hat{h}(X,Y) = -\psi(k) + \psi(n) + \log(c_{d_X}c_{d_Y}) + \frac{d_X+d_Y}{n}\sum_{i=1}^{n}\log\epsilon(i), \tag{3.3}$$

여기서 $\psi(x)$는 digamma 함수

$$\psi(x) = \Gamma(x) - 1 - \frac{d\Gamma(x)}{dx} \tag{3.4}$$

이고, $\psi(1) = -0.5772157$이다. 또한 n은 표본의 크기, $\epsilon(i)$는 확률변수가 가지는 i번째 값으로부터 k번째 이웃(k^{th} neighbor)까지의 거리의 2배, d는 확률변수의 차원이며 c_d는 d차원 단위구(unit ball)의 부피이다. 따라서 k번째 이웃 추정 MI는 다음과 같다.

$$\hat{I}_k(X;Y) = \hat{h}(X) + \hat{h}(Y) - \hat{h}(X,Y). \tag{3.5}$$

또, Kraskov 등 (2004)에서 제시한 k-최근접이웃에 기반을 둔 MI 추정 알고리즘에 대해, 지터링과 붓스트랩 방법으로 조정한 알고리즘은 다음과 같다.

A. 연속형 설명변수와 범주형 목적변수인 경우

연속형 주변 확률변수 X에 대한 엔트로피를 다음 식과 같이 추정한다.

$$\hat{h}(X) = \psi(n) - \frac{1}{n}\sum_{i=1}^{n}\psi[n_x(i)+1] + \log c_{d_x} + \frac{d_x}{n}\sum_{i=1}^{n}\log\epsilon(i). \tag{3.6}$$

[알고리즘 A] ($Z=(X,Y)$, X는 연속형 확률변수, Y는 범주형 확률변수)

1. X를 지터링 시킨다.
2. 순위 $d_{i,j1} \le d_{i,j2} \le d_{i,j3} \le \cdots$를 구한다.
   여기서 $d_{i,j} = \|x_i - x_j\|$는 x_i와 x_j사이의 거리.
3. $\epsilon(i)$를 구한다.
   $\epsilon(i)/2$: x_i부터 k번째 이웃까지의 거리.
4. $\|x_i - x_j\| < \epsilon(i)/2$인 x_j의 개수 $n_x(i)$를 구한다.
5. 식 (3.6)을 이용하여 MI 추정량을 다음 식에 의해 구한다.
$$\hat{I}_k(X;Y) = \hat{h}(X) - \hat{h}(X\,|\,Y). \tag{3.7}$$
6. 1~5 과정을 붓스트랩 횟수만큼 반복해서 추정량의 평균을 구한다.

B. 연속형 설명변수와 연속형 목적변수인 경우

[알고리즘 B] ($Z=(X,Y)$: X, Y는 연속형 확률변수)

1. X, Y를 지터링시킨다.
2. 순위 $d_{i,j1} \le d_{i,j2} \le d_{i,j3} \le \cdots$를 구한다.
   여기서 $d_{i,j} = \|z_i - z_j\| = \max[\|x_i - x_j\|, \|y_i - y_j\|]$는 점 $z_i = (x_i, y_i)$에 대한 점 $z_j = (x_j, y_j)$ 사이의 거리.
3. $\epsilon(i) = \max[\epsilon_x(i), \epsilon_y(i)]$를 구한다.
   (a) 여기서 $\epsilon(i)/2$: z_i부터 k번째 이웃까지의 거리.
   (b) $\epsilon_x(i)/2$: z_i와 z_i의 k번째 이웃 점을 X 주변공간으로 사영시킨 점들 사이의 거리.
   (c) $\epsilon_y(i)/2$: z_i와 z_i의 k번째 이웃 점을 Y 주변공간으로 사영시킨 점들 사이의 거리이다.

[알고리즘 B-1]

4-1. $n_x(i)$, $n_y(i)$를 구한다.

여기서 $n_x(i)$는 $\|x_i - x_j\| < \epsilon(i)/2$인 x_j의 개수, $n_y(i)$는 $\|y_i - y_j\| < \epsilon(i)/2$인 y_j의 개수이다.

5-1. MI 추정량을 다음 식에 의해 구한다.

$$\hat{I}_k^{(1)}(X;Y) = \psi(k) - \frac{1}{n}\sum_{i=1}^{n}[\psi(n_x(i)+1) + \psi(n_y(i)+1)] + \psi(n). \tag{3.8}$$

6-1. 1~5-1 과정을 붓스트랩 횟수만큼 반복해서 추정량의 평균을 구한다.

[알고리즘 B-2]

4-2. $n_x(i)$, $n_y(i)$를 구한다.

여기서 $n_x(i)$는 $\|x_i - x_j\| < \epsilon_x(i)/2$인 x_j의 개수, $n_y(i)$는 $\|y_i - y_j\| < \epsilon_y(i)/2$인 y_j의 개수이다.

5-2. MI 추정량을 다음 식에 의해 구한다.

$$\hat{I}_k^{(2)}(X;Y) = \psi(k) - \frac{1}{k} - \frac{1}{n}\sum_{i=1}^{n}[\psi(n_x(i) + \psi(n_y(i))] + \psi(n). \tag{3.9}$$

6-2. 1, 2, 3, 4-2, 5-2 과정을 부스트래핑 횟수만큼 반복해서 추정량의 평균을 구한다.

본 연구에서는 확률변수 X, Y는 1차원의 경우만 고려하였으며, 연속형 설명변수와 연속형 목적변수의 경우 [알고리즘 B-1]과 [알고리즘 B-2]를 사용하여 실험해 본 결과 큰 차이가 나타나지 않았으므로 [알고리즘 B-1]만 사용하기로 한다.

4. 몬테칼로 모의실험

4.1. 데이터의 생성

A. 연속형 설명변수 X와 범주형 목적변수 Y

X가 연속형, Y가 범주형 변수일 때 MI 추정에 영향을 주는 요소는 X와 Y의 분포, 표본크기 n과 k-최근접이웃 기반방법의 모수 k이다. 본 실험에서는 범주형 변수 Y가 이진값 1과 2를 가지는 경우만 고려한다, $p = P[Y=1]$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.

[분포 1] $pN(0,\,1) + (1-p)N(\mu,\,\sigma^2)$, (4.1)

[분포 2] $p\text{Gamma}(\lambda_1,\,r=1) + (1-p)\text{Gamma}(\lambda_2,\,r=1)$, (4.2)

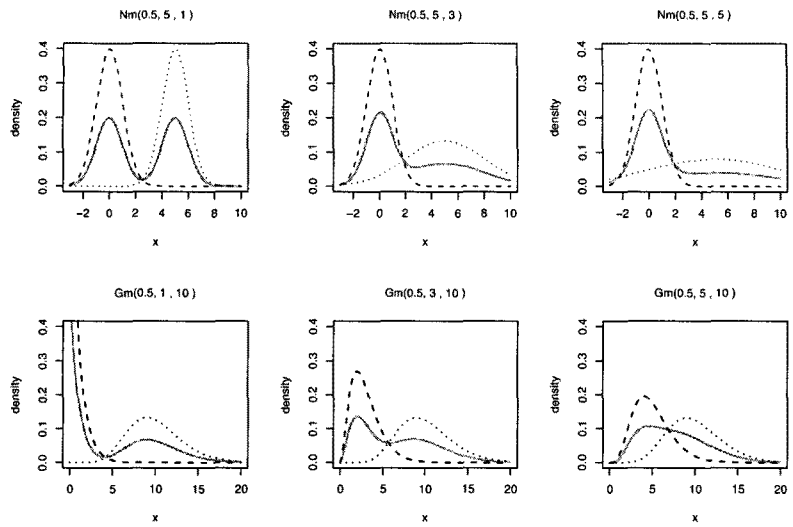

그림 4.1: 모수 값에 따른 혼합정규분포와 혼합감마분포

여기서 $N(\mu, \sigma^2)$는 정규분포이고, $G(\lambda, r = 1)$은 형태(shape) 모수 λ와 크기(scale) 모수 r을 갖는 감마분포이다. 편의상 (4.1)의 혼합정규분포는 $Nm(p, \mu, \sigma)$로, (4.2)의 혼합감마분포는 $Gm(p, \lambda_1, \lambda_2)$로 나타내기로 한다. 각 분포의 경우 다음과 같은 모수 값을 실험에 사용한다.

[분포 1] $\mu = 5$; $\sigma = 1, 3, 5$,

[분포 2] $\lambda_1 = 1, 3, 5$; $\lambda_2 = 10$.

그림 4.1에 $p = 0.5$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 n= 30, 50, 100, 150, 200, 300, 500, 1000개 씩 생성하였다.

B. 연속형 설명변수 X와 연속형 목적변수 Y

본 논문에서는 (X, Y)가 상관계수가 ρ인 이변량 정규분포인 경우만 고려한다. 이변량 정규난수의 생성은 다음과 같이 한다.

1. $X \sim N(0, 1)$인 난수를 생성한다. $Y \sim N(0, 1)$인 난수를 생성한다. 따라서 X와 Y는 서로 독립이다.

2. $Y \leftarrow \rho X + \sqrt{(1 - \rho^2)}Y$로 변형시킨다.

실험을 위해 n= 30, 50, 100, 150, 200, 300, 500, 1000개 씩 생성한다.

4.2. 실험방법

1. k-최근접 기반 방법에서의 모수 k값 1, 2, 3, 5, 10에 대해 실험한다.
2. 연속형 설명변수와 범주형 목적변수의 경우 표본간격 방법($m = 1$)과 비교 한다.
3. 연속형 설명변수와 연속형 목적변수의 경우 ρ값이 0.9, 0.6, 0.3, 0.1인 경우에 대해 실험한다.
4. 모든 실험의 반복횟수는 500번으로 한다.
5. 붓스트랩 횟수는 1000번으로 한다.

MI 추정의 효율을 평가하기 위한 척도는 다음을 사용하였다.

$$e(k) = \frac{\hat{I}_k(X;Y)}{I(X;Y)} - 1, \tag{4.3}$$

여기서, $\hat{I}_k(X;Y)$는 k-최근접 MI 추정량을 의미하고 $I(X;Y)$는 이론상의 MI를 의미한다. $I(X;Y)$는 다음과 같이 구할 수 있다.

(a) $X \sim Nm(p, \mu, \sigma)$이고 Y가 이산형인 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 MI $I(X;Y)$는 다음과 같다.

$$I(X;Y) = h(Nm(p, \mu, \sigma)) - h_1 - h_2, \tag{4.4}$$

여기에서,

$$h_1 = \frac{1}{2}\log(2\pi e), \qquad h_2 = \frac{1}{2}\log(2\pi e\sigma^2), \tag{4.5}$$

$$h(Nm(p, \mu, \sigma)) = \int f(x) \log f(x) dx, \tag{4.6}$$

$$f(x) = pN(0, 1) + (1 - p)N(\mu, \sigma^2) \tag{4.7}$$

이다. 식 (4.5)와 (4.6)의 엔트로피를 구하기 위해서는 수치적분을 해야 하는데, 수치적분은 R의 함수 integrate를 사용한다.

(b) $X \sim Gm(p, \lambda_1, \lambda_2)$이고 Y가 이산형인 경우:

역시 Lazo와 Rathie에 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, MI 는 다음과 같이 계산할 수 있다.

$$I(X;Y) = h(Gm(p, \lambda_1, \lambda_2)) - h_1 - h_2, \tag{4.8}$$

여기에서,

$$h_1 = \log\Gamma(\lambda_1) + (1 - \lambda_1)\psi(\lambda_1) + \lambda_1, \tag{4.9}$$

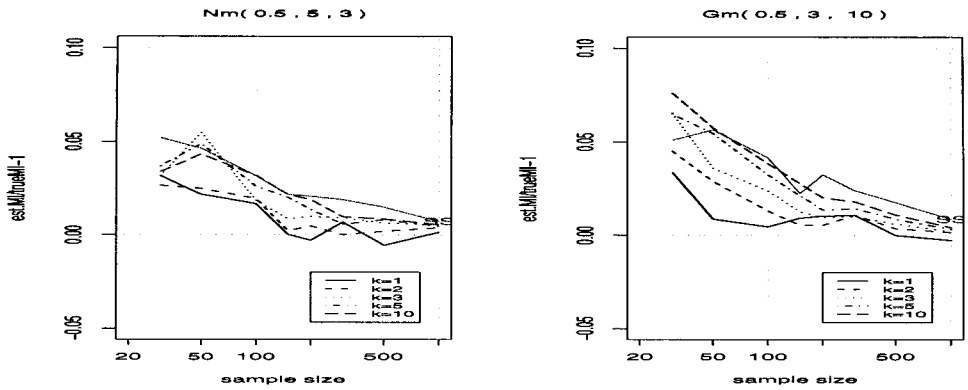

그림 4.2: $Nm(0.5, 5, 3)$와 $Gm(0.5, 3, 10)$분포에서 표본간격 방법에 의한 'MI 추정량과 이론상 MI 값의 비율 -1'과 k-최근접이웃 기반방법에 의한 $e(k)$

$$h_2 = \log\Gamma(\lambda_2) + (1-\lambda_2)\psi(\lambda_2) + \lambda_2, \tag{4.10}$$

$$h(Gm(p, \lambda_1, \lambda_2)) = \int f(x)\log f(x)dx, \tag{4.11}$$

$$f(x) = pG(\lambda_1, r=1) + (1-p)G(\lambda_2, r=1) \tag{4.12}$$

(c) $X \sim N(0,1)$, $Y \sim N(0,1)$인 경우

공분산행렬 $\sum_{XX}$와 $\sum_{YY}$를 가지는 다변량 정규 확률변수 X와 Y에 대한 MI $I(X;Y)$는 다음 식과 같다 (Cover와 Thomas, 1991).

$$I(X;Y) = -\frac{1}{2}\log\left(\frac{|\sum|}{|\sum_{XX}||\sum_{YY}|}\right) = -\frac{1}{2}\sum_i \log\left(1-\rho_i^2\right), \tag{4.13}$$

여기에서 $\sum$는 행벡터 $(X;Y)$의 공분산행렬이고 ρ_i는 상관계수이다. 따라서 이변량 X, Y의 경우 MI는 $-1/2\log(1-\rho^2)$와 같다.

4.3. 실험 결과

A. 연속형 설명변수와 범주형 목적변수인 경우

그림 4.2에는 $Nm(0.5, 5, 3)$과 $Gm(0.5, 3, 10)$ 분포에 대해 표본의 크기 (30, 50, 100, 150, 200, 300, 500, 1000)에 log를 취한 값을 x-축에 나타내고, y-축에는 'MI 추정량과 이론상 MI 값의 비율 -1' 값을 나타내었다. 여러 개의 선은 k(1, 2, 3, 5, 10)에 해당하는 효율 $e(k)$를 나타내고, 오른쪽의 SS는 $m = 1$일 때의 표본간격 방법으로 구한 'MI 추정량과 이론상 MI 값의 비율 -1'을 나타낸다. 이 그림을 보면 k값이 1과 같이 작은 값일 때 추정량이 이론상의 MI와 가까운 값을 가지는 것을 알 수 있고, k-최근접이웃 기반방법으로 MI를 추정하는 것이 $m = 1$일 때의 표본간격 방법 보다 더 좋은 결과를 가져오는 것을 알

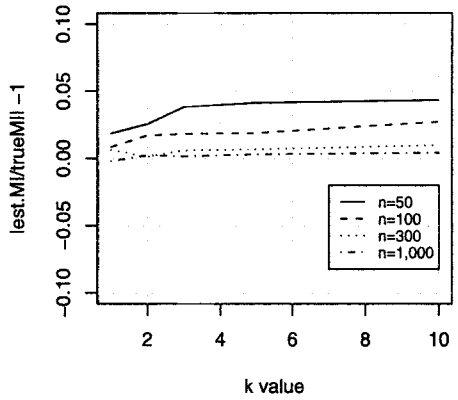

그림 4.3: $Nm(0.5, 5, 3)$ 분포에서 표본크기에 따른 $e(k)$

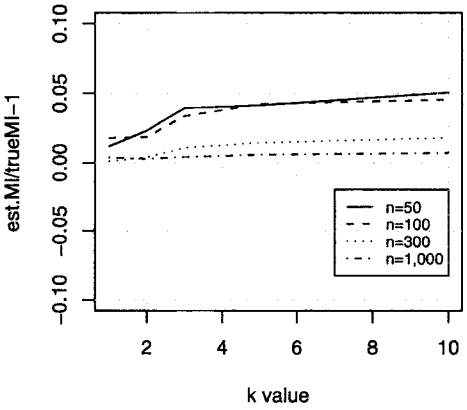

그림 4.4: $Gm(0.5, 3, 10)$ 분포에서 표본크기에 따른 $e(k)$

수 있다. 다른 모수 값을 가지는 혼합 정규분포와 혼합 감마분포에서도 비슷한 결과가 나타났다.

k-최근접이웃 기반방법으로 MI를 추정하는 경우, 표본의 크기에 따라 어떤 값을 사용하는 것이 더 좋은 추정량을 얻을 수 있는지 알아보기 위하여 각각의 표본크기에 대해 값을 변화시켜 가면서 $e(k)$를 구해 $Nm(0.5, 5, 3)$ 분포의 경우 그림 4.3에, $Gm(0.5, 3, 10)$ 분포의 경우 그림 4.4에 나타내었다. 이 그림들에는 여러 가지 표본 크기 중 n= 50, 100, 300, 1000인 경우만 나타내었다. 그림을 보면 표본의 크기에 상관없이 k값으로 1을 사용하면 좋은 결과를 가져옴을 알 수 있다.

몬테칼로 모의실험을 위해 생성한 각 분포에 대한 이론상의 MI 값을 구하면 다음과 같다.

$$Nm(0.5, 5, 1) : 0.676 \qquad Nm(0.5, 5, 3) : 0.46 \qquad Nm(0.5, 5, 5) : 0.402$$
$$Gm(0.5, 1, 10) : 0.657 \qquad Gm(0.5, 3, 10) : 0.518 \qquad Gm(0.5, 5, 10) : 0.315$$

MI 값의 크기가 k에 영향을 미치는지 알아보기 위하여 실제 MI가 가장 큰 $Nm(0.5,$

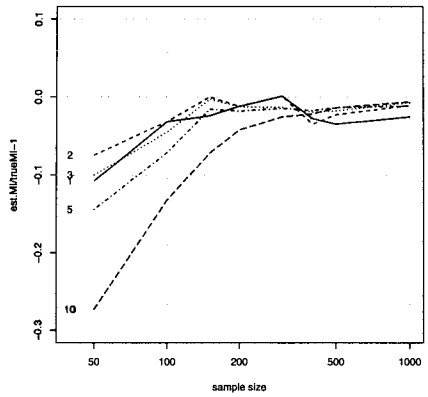

그림 4.5: ρ=0.6일 때의 $e(k)$

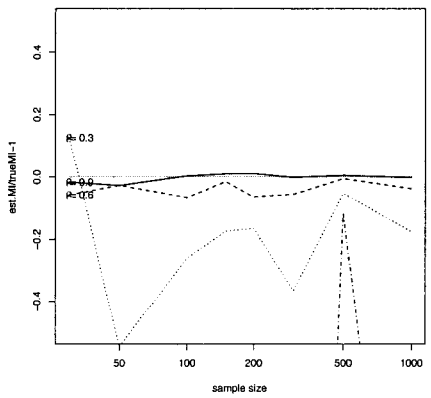

그림 4.6: $k = 1$일 때, ρ(0.9, 0.6, 0.3, 0.1) 별로 나타낸 $e(k)$

5, 1)분포와 가장 작은 Gm(0.5, 5, 10)분포에 대해 동일한 분석을 수행해 본 결과 MI 값이 가장 큰 Nm(0.5, 5, 1)분포와 MI 값이 가장 작은 Gm(0.5, 5, 10)에서도 $k = 1$일 때 가장 좋은 추정값을 가지는 것을 확인할 수 있었다. 이상의 실험에서 얻은 결과를 정리하면, 연속형 확률변수의 결합확률분포를 추정하지 않고도 MI를 추정할 수 있는 방법 중 k-최근접 이웃 기반방법이 $m = 1$인 표본간격 방법보다 더 좋은 결과를 준다. 여러 혼합분포에서 생성한 표본의 크기에 상관없이, 또 이론상의 MI 크기에 별 상관없이 $k = 1$일 때 가장 효율적인 추정 값을 얻을 수 있었다.

B. 연속형 설명변수와 연속형 목적변수

실험을 위해 생성한 설명변수 $X \sim N(0, 1)$, 목적변수 $Y \sim N(0, 1)$인 데이터에 대해 ρ값이 0.9, 0.6, 0.3, 0.1일 때 $e(k)$를 구한 결과 중 ρ가 0.6인 경우를 그림 4.5에 나타내었다. 실험에 의하면 ρ가 큰 경우 $k = 1$ 또는 2과 같이 작은 값일 때 이론적인 MI에 가까운 추정값을 얻으며, ρ가 작은 경우에는 $e(k)$ 값의 변동이 상당히 심하게 나타났다. 그림 4.6에는 $k = 1$일 때 (0.9, 0.6, 0.3, 0.1)별로 $e(k)$를 나타내었다.

표 4.1: $k = 1$, $\rho = 0.1$일 때의 $d(k)$와 $e(k)$

	$n = 50$	$n = 100$	$n = 200$	$n = 300$	$n = 500$	$n = 1000$
$d(k)$	-0.0324	-0.0172	-0.0154	-0.0113	-0.00902	-0.00901
$e(k)$	-6.44	-3.41	-3.07	-2.24	-1.8	-1.79

Kraskov 등 (2004)의 논문에서는 $\rho \to 0$인 경우 $k = 1$일 때 매우 좋은 추정값을 제공하는 것으로 되어 있으나 이는 'MI 추정량 – 이론상 MI 값'인 차이(difference)를 판단 기준으로 사용했기 때문인 것으로 결과의 해석이 잘 못 되어 있다. 표 4.1에 $k = 1$이고 ρ가 0.1일 때, 표본의 크기 n 별로 Kraskov 등 (2004)이 사용한 판단기준 $d(k) = \hat{I}_k(X;Y) - I(X;Y)$와 본 논문에서 사용하고 있는 판단기준 $e(k)$, 즉 상대오차를 정리하였다. ρ가 작을 때에는 $d(k)$가 작은 값이지만 $e(k)$는 아주 큰 값임을 확인할 수 있다. 실험결과를 정리하면 연속형 설명변수와 연속형 목적변수의 경우, ρ가 작을 때 k-최근접이웃 기반방법은 매우 불안정한 MI 추정값을 제공한다. 따라서 두 변수 사이에 상관관계가 별로 없는 경우에는 k-최근접이웃 기반방법이 효율이 떨어지는 것을 알 수 있다. 그러나 실제 문제에서 ρ가 작은 경우, 추정값도 매우 작게 추정해 주기 때문에 큰 문제는 없다. ρ가 작지 않을 경우에는 $k = 1$을 사용하여 MI 추정량을 구하면 좋은 결과를 얻을 수 있다.

5. 실제 데이터에 대한 실험

A. 연속형 설명변수와 범주형 목적변수

1. 실험에 사용한 데이터

   본 연구의 실험을 위해 모든 설명변수들이 연속형 값을 가지고, 클래스의 개수가 3인 데이터 IRIS와 클래스의 개수가 2인 데이터 WDBC를 UCI 창고(machine learning repository, Blake와 Merz, 1998)에서 구하였다.

   1) IRIS 데이터

      데이터의 크기가 150인 IRIS 데이터는 4개의 연속형 변수($X1$: Sepal Length, $X2$: Sepal Width, $X3$: Petal Length, $X4$: Petal Width)와 1개의 범주형 목적변수(Iris-setosa: 50, Iris-versicolor: 50, Iris-virginica: 50)로 구성되어 있다.

   2) WDBC 데이터

      WDBC는 569명의 유방암 환자에 대한 데이터로서 30개($X1 \sim X30$)의 연속형 설명변수와 1개의 범주형 목적변수(양성환자: 357, 악성환자: 212)로 구성되어 있다.

2. 실험결과

   IRIS 데이터와 WDBC 데이터에서 목적변수에 영향을 미치는 중요한 설명변수 10개의 MI 추정 값 크기의 순서는 $k = 1, 2, 3, 5, 10$에 대하여 다음 표 5.1과 같다.

표 5.1: k-최근접이웃 기반방법에 의한 중요변수 순위

	IRIS	WDBC
$k=1$	4, 3, 1, 2	24, 8, 28, 3, 21, 23, 4, 7, 14, 27
$k=2$	4, 3, 1, 2	23, 24, 8, 28, 21, 3, 4, 7, 1, 14
$k=3$	4, 3, 1, 2	23, 21, 24, 28, 8, 3, 4, 7, 1, 14
$k=5$	4, 3, 1, 2	23, 21, 24, 28, 8, 3, 4, 7, 1, 14
$k=10$	4, 3, 1, 2	23, 24, 21, 28, 8, 3, 4, 7, 1, 27

표 5.2: 다른 방법에 의한 중요변수 순위

방법	IRIS	WDBC
표본간격($m=1$)	3, 4, 1, 2	24, 23, 21, 28, 8, 3, 7, 1, 4, 27
이산화(4-구간방법)	4, 3, 1, 2	21, 23, 24, 28, 8, 3, 1, 4, 7, 14
ReliefF	4, 3, 1, 2	21, 28, 23, 22, 1, 3, 8, 24, 4, 7

동일한 데이터에 대해 DAVIS (Huh, 2005)를 사용하여 표본간격, 이산화, ReliefF (Kononenko, 1994)방법으로 구한 중요변수 순위는 다음 표 5.2와 같다. 본 실험에서 사용한 이산화방법은 상자도형에 기반을 둔 4-구간 방법 (Cha와 Huh, 2005)이다.

표에서 알 수 있는 바와 같이, $k=1$과 같이 작은 값으로 k-최근접이웃 기반방법에 의해 MI를 추정하는 경우 변수 순서는 약간 다르더라도 다른 방법에서와 유사한 중요변수를 구할 수 있다.

B. 연속형 설명변수와 연속형 목적변수

1. 실험에 사용한 데이터

   연속형 설명변수와 범주형 목적변수 경우에 사용한 IRIS 데이터를 다시 사용하고, 설명변수와 목적변수가 모두 연속형인 CHEESE 데이터를 CMU 데이터베이스 (http://lib. stat.cmu.edu/DASL/alltopics.html)에서 구하였다.

   1) IRIS 데이터

      $X1 \sim X4$의 연속형 설명변수 중 $X1$을 목적변수로 사용하고, $X2, X3, X4$를 설명변수로 사용한다.

   2) CHEESE 데이터

      치즈는 생산된 후 시간이 지날수록 화학반응이 일어나서 치즈의 맛이 정해진다. 이 데이터는 연속형 변수 $X1 \sim X4$로 이루어져 있고 크기는 30이다. $X1$(Taste)은 치즈의 맛을 수치로 나타낸 변수이며 본 실험에서 목적변수로 사용하였다. $X2$(Acetic)는 초산의 농축, $X3$(H2S)는 황화수소의 농축, $X4$(Lactic)는 유산의 농축을 나타내며 설명변수로 사용하였다.

2. 실험결과

   IRIS, CHEESE 데이터에서 $k=1, 2, 3, 5, 10$일 때, 목적변수 $X1$에 대해 MI가 큰 변수 순으로 정리하면 다음 표 5.3과 같다.

표 5.3: $X1$에 대해 MI가 큰 변수 순위

	IRIS	CHEESE
$k=1$	3, 4, 2	3, 4, 2
$k=2$	3, 4, 2	3, 4, 2
$k=3$	3, 4, 2	3, 4, 2
$k=5$	3, 4, 2	3, 4, 2
$k=10$	3, 4, 2	3, 4, 2

IRIS 데이터의 경우 $X1$과 $X2$, $X3$, $X4$와의 상관계수는 -0.109, 0.872, 0.818이고, CHEESE 데이터의 경우는 0.550, 0.756, 0.704이다. 따라서 $k=1$과 같이 작은 값을 사용하더라도 큰 값을 사용하는 경우와 마찬가지로 목적변수와 관계가 깊은 변수 순서를 잘 구해주는 것을 확인할 수 있다.

6. 결론

본 논문에서는 연속형 변수에 대한 결합확률분포를 추정하지 않고도 MI 추정량을 구할 수 있는 k-최근접이웃 기반방법에 대하여 연구하였다. Kraskov 등 (2004)의 k-최근접이웃 기반방법을 지터링과 붓스트랩 방법으로 조정한 알고리즘을 사용하였다. 연속형 설명변수와 범주형 목적변수의 경우, 몬테칼로 모의실험에서는 $k=1$을 사용한 k-최근접이웃 기반방법이 좋은 MI 추정량을 구해주고, 표본간격 방법보다 더 좋은 결과를 가져옴을 확인할 수 있었다. 실제 데이터에 대한 실험에서도 같은 결과를 얻었다. 설명변수와 목적변수가 둘 다 연속형인 경우에는 ρ가 작지 않은 경우에 k-최근접이웃 기반방법에서 $k=1$을 사용하면 MI 추정량을 잘 구해 주었다. 실제 데이터에 대한 실험에서는 연속형 목적변수와 상관관계가 높은 변수 순으로 MI 추정량의 값이 크게 나타났다. ρ가 작은 경우, 비록 상대오차가 매우 불안정하다고 하더라도 추정 값이 매우 작기 때문에 실제 문제를 다룰 때는 큰 문제가 없을 것이다. 표본간격 방법은 목적변수가 범주형인 경우에만 사용할 수 있는 반면에 k-최근접이웃 기반방법은 목적변수가 범주형뿐만 아니라 연속형인 경우에도 적용할 수 있다. 또한 본 연구에서는 두 변수 사이의 MI 추정량을 구해 변수의 중요도를 구하는 것만 고려하였으나, k-최근접이웃 기반방법은 다차원으로 확장이 가능하기 때문에 중요변수 집합을 구하는 변수 선택(feature subset selection) 문제에도 적용할 수 있다.

참고문헌

허문열, 차운옥 (2008). Sample-spacing 방법에 의한 상호정보의 추정, <응용통계연구>, **21**, 301–312.

Beirlant, J., Dudewicz, E. J., Gyorfi, 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. (1998). UCI machine learning repository, http:// www.ics.uci.edu/ mlearn/MLRepository

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

Cha, W. O. and Huh, M. Y. (2005). Discretization method based on quantiles for variable selection using mutual information, *Communications of the Korean Statistical Society,* **12**, 659–672.

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

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

Kononenko, I. (1994). Estimating attributes: Analysis and extensions of RELIEF, In *Proceedings of European Conference on Machine Learning,* 171–182.

Kraskov, A., Staugbauer, H. and Grassberger, P. (2004). Estimating Mutual Information, *Physical Review E 69,* 066138.

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 estimation of entropy, *The Journal of Machine Learning Research,* **4**, 1271–1295.

Staugbauer, H., Kraskov, A., Astakhov, S. A. and Grassberger, P. (2004). Least dependent component analysis based on mutual information, *Physical Review* E 70, 066123.

[2008년 8월 접수, 2008년 10월 채택]

k-Nearest Neighbor-Based Approach for the Estimation of Mutual Information†

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

Abstract

This study is about the k-nearest neighbor-based approach for the estimation of mutual information when the type of target variable is categorical and continuous. The results of Monte-Carlo simulation and experiments with real-world data show that $k = 1$ is preferable. In practical application with real world data, our study shows that jittering and bootstrapping is needed.

Keywords: Mutual information; k-nearest neighbor; jittering; bootstrap.

† This research was financially supported by Hansung University in the year of 2008.

1) *Professor, Department of Mutimedia Engineering, Hansung University, Seoul 136-792, Korea.* Correspondence: wcha@hansung.ac.kr

2) Professor, Department of Statistics, Sungkyunkwan University, Seoul 110-745, Korea.