한국통계학회논문집
2010, 17권, 4호, 591–598

상호정보와 엔트로피를 활용한 대표문항 선택방법

최병수[1,a], 김현지[b]

[a]한성대학교 멀티미디어공학과, [b]성균관대학교 통계학과

요 약

설문문항에는 유사한 문항이나 중복문항이 있는 경우들이 있다. 본 연구에서는 전체 문항의 응답패턴을 군집분석하여 군집의 성격을 파악하고, 유사한 문항을 묶어 대표 문항을 찾는 방법을 시도하였다. 문항의 유사성 측정은 상호정보를 이용하였고 군집분석과 다차원척도법을 이용하여 대표 문항을 찾는 방법을 제안 하였으며 엔트로피를 이용하여 대표 문항을 평가하였다.

주요용어: 엔트로피, 상호정보, 군집분석, 다차원척도법(MDS).

1. 서론

설문지의 문항에는 유사한 문항이나 중복문항이 있는 경우가 있다. 의도적인 경우도 있겠지만, 의도하지 않은 경우 비슷한 문항을 파악해 볼 필요가 있다.

본 논문에서는 상호정보와 엔트로피를 활용하여 이러한 설문 자료에서 대표문항을 찾는 방안을 제안한다. 전체 문항의 응답패턴을 군집분석하여 군집의 성격을 파악하고, 유사한 문항을 묶어 대표문항을 찾는 방법을 시도하였다. 문항의 유사성 측정은 상호정보를 이용하였고 군집분석과 다차원척도법(MDS; Multi-Dimensional Scaling)을 통해 대표 문항을 찾는 방법을 제안 하였으며 엔트로피를 이용하여 대표 문항을 평가하였다.

상호정보(Mutual Information)는 Shannon (1948)의 정보이론에 기반을 두고 있다. Silvey (1964)는 선형 관계만을 나타내는 피어슨 상관계수와는 다르게 모든 형태의 연관성을 나타내기 때문에 이상적인 확률적 연관성의 측도라고 볼 수 있으며 상호정보는 변수간의 관계를 측정하는 도구로 사용될 수 있다고 하였다. 또한, Lee와 Huh (2003)는 상호정보는 아무런 가정 없이 혼합형 데이터(complex data)에서도 변수간의 관계를 파악 할 수 있다는 장점을 가지고 있다고 하였다.

엔트로피(Entropy)는 이산 확률변수 X에 대해 $H(X) = -\sum p(x)\log p(x)$으로 정의되며, 항상 $H(X) \geq 0$이다. 결합 엔트로피 $H(X,Y)$는 $H(X,Y) = H(X) + H(Y|X)$으로 정의된다. 연속 확률변수에 대한 엔트로피는 미분 엔트로피(Differential entropy)라고도 부르며 $H(X) = -\int f(x)\log f(x)dx$로 정의된다 (Thomas와 Joy, 2006).

논문의 구성은 다음과 같다. 2절에서는 문항의 패턴화와 그 문제점에 대해 알아본다. 3절에서는 응답패턴을 통한 군집분석으로 대표 문항을 찾는 방법을 제시한다. 4절에서는 대표 문항의 엔트로피를 이용하여 평가하고 마지막으로 5절에서는 본 연구의 결과에 대해 정리하고 토론한다.

본 연구는 2009학년도 한성대학교 교내 연구비의 지원을 받았음

[1] 교신저자 : (136-793) 서울시 성북구 삼성동 3가 389. 한성대학교 멀티미디어공학과, 교수.
E-mail: cbs@hansung.ac.kr

표 1: 여론조사 자료의 설명 1

문 항	유형	항목 수	비고	문 항	유형	항목 수	비고
Q_1	명목형	4개		Q_5	명목형	2개	
Q_2	명목형	5개	중복문항	Q_6	명목형	2개	
Q_3	명목형	3개		Q_7	명목형	4개	
Q_4	명목형	4개					

표 2: 여론조사 자료의 설명 2

문 항	항목 수	문항형태	이진변수
Q_1	4개	categorical	x_1, x_2, x_3, x_4
Q_2	5개	categorical	x_5, x_6, x_7, x_8, x_9
Q_3	3개	categorical	x_{10}, x_{11}, x_{12}
Q_4	4개	categorical	$x_{13}, x_{14}, x_{15}, x_{16}$
Q_5	2개	binary	x_{17}
Q_6	2개	binary	x_{18}
Q_7	4개	categorical	$x_{19}, x_{20}, x_{21}, x_{22}$

표 3: 이진변수 문항으로 바꾼 데이터

원래문항	Q_1				$\cdots$	Q_6	Q_7			
이진변수	X_1	X_2	X_3	X_4	$\cdots$	X_{18}	X_{19}	X_{20}	X_{21}	X_{22}
1	0	0	1	0	$\cdots$	1	1	0	1	0
2	0	0	1	0	$\cdots$	0	1	1	1	1
3	0	1	0	0	$\cdots$	0	0	1	1	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
607	1	0	0	0	$\cdots$	0	0	1	0	1
608	0	0	0	1	$\cdots$	1	0	0	0	1

2. 문항의 패턴화

명목형 변수를 이진변수로 바꾸는 다음과 같은 방법을 이용하였다 (David 등, 2002). 즉, x가 4가지 색상(red, blue, green, yellow)에 대한 응답인 명목형(categorical)자료인 경우 다음과 같은 방법으로 이진변수를 만들었다.

$$x_1 = \begin{cases} 1, & \text{red}, \\ 0, & \text{o.w}, \end{cases} \quad x_2 = \begin{cases} 1, & \text{blue}, \\ 0, & \text{o.w}, \end{cases} \quad x_3 = \begin{cases} 1, & \text{green}, \\ 0, & \text{o.w}, \end{cases} \quad x_4 = \begin{cases} 1, & \text{yellow}, \\ 0, & \text{o.w}. \end{cases}$$

본 연구에서는 대학생 이성교제에 관한 설문조사를 실시하였고 설문지 문항은 총 7개로 구성되어 있으며, 각 문항의 항목 수는 다음 표 1과 같다.

본 연구에서도 David 등 (2002)의 방법을 이용하여 7개의 명목형 변수를 22개의 이진변수로 변환하였다. 첫 번째 문항의 항목이 4개 이므로 범주가 두 개인 4개의 항목이 만들어지고 두 번째 문항에서는 5개의 항목이 만들어진다. 다섯 번째 문항과 여섯 번째 문항은 원래 범주가 두 개인 문항이므로 그대로 사용한다 (표 2). 각 문항을 범주 두 개인 항목으로 바꾼 데이터의 형태는 다음 표 3과 같다.

이론상 응답자의 응답 가능한 패턴 수는 2^{22}개 이다. 하지만 데이터에 결측값이 존재하기 때문에 결측값을 포함하여 응답 가능한 패턴 수를 보면 3^{22}개 이다. 3^{22}개 응답패턴을 고려하는 것은 불가능하고 대부분의 응답패턴의 빈도가 0이기 때문에, 빈도가 0인 응답패턴을 제외시켰다. 그 결과 다음 표 4와 같이 524개의 응답패턴만 남게 되었다.

표 4: 응답패턴

patterns	X_1	X_2	X_3	$\cdots$	X_{20}	X_{21}	X_{22}	Freq.
1	0	0	1	$\cdots$	0	1	1	1
2	0	0	1	$\cdots$	0	0	1	1
3	0	0	1	$\cdots$	0	0	1	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
522	−1	−1	−1	$\cdots$	0	1	1	1
523	−1	−1	−1	$\cdots$	0	0	1	1
524	−1	−1	−1	$\cdots$	0	0	−1	1

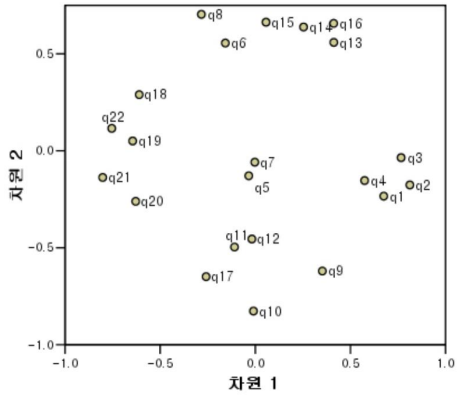

그림 1: *다차원척도법*

이 결과로부터 524개의 응답패턴을 군집분석하여 나눠진 군집의 내재된 성격을 파악한다. 표 4에서 1은 긍정적 응답(Yes), 0은 부정적 응답(No), −1은 결측값이다.

군집분석 결과 군집은 4개 또는 5개로 나눠졌지만 응답자 수와 비슷한 응답패턴 수로 인해 군집의 해석이 용이하지 않을 뿐만 아니라 패턴을 통해 군집의 성격을 파악하기는 쉽지 않았다. 그러므로 응답패턴을 이용하여 항목의 수를 줄이는 방법을 시도해야 하며, 본 논문에서는 상호정보를 이용하여 유사문항을 찾아 응답패턴을 통해 대표 문항을 찾는 방법을 제안하도록 한다. 이는 다음절에서 소개하도록 한다.

3. 응답패턴을 통한 군집분석

문항 수가 많으면 설문자로부터 성실한 답변을 얻기 어려울 뿐 아니라 유사한 문항이 존재할 가능성이 높아진다. 본 연구에서는 효과적이며 체계적인 설문지를 작성하고, 설문지 전체를 포괄할 수 있는 대표문항을 찾는 방법을 제안하고자 한다.

개체를 군집분석 할 경우 각 군집에서 중요한 문항이 무엇인지, 군집을 묶은 요인이 무엇인지 알기 어렵다. 그러나 응답패턴으로 군집분석을 하면 집단의 성격을 파악할 수 있고, 군집을 묶는 요인을 쉽게 파악할 수 있다. 또한, 유사문항이나 중복문항에 대해 대표 문항을 찾을 수 있게 된다. 본 연구에서는 전체 22개 문항으로부터 유사한 문항을 찾기 위하여 상호정보를 이용한 다차원척도법을 실시하였고, 그 결과는 그림 1과 같다.

문항 q_1, q_2, q_3, q_4은 오른쪽 중앙에서 그룹 지어졌고, $q_{18}, q_{19}, q_{20}, q_{21}, q_{22}$는 왼쪽 중앙에서 그룹 지어졌다. 유사한 문항끼리 그룹 지어진 결과는 그림 1과 같다.

표 5: 다차원척도법 시행 결과

그룹 명	그룹 지어진 문항	그룹 설명
그룹 1	q_1, q_2, q_3, q_4	Experience
그룹 2	q_5, q_7	Importance 1
그룹 3	q_6, q_8	Importance 2
그룹 4	q_9, q_{10}, q_{11}, q_{12}, q_{17}	Skinship
그룹 5	q_{18}, q_{19}, q_{20}, q_{21}, q_{22}	Baby
그룹 6	q_{13}, q_{14}, q_{15}, q_{16},	Ideal

표 6: 그룹 1 문항에 대한 빈도표

Item	Response			Total
	−1	0	1	
q_1: expri_x	4	498	106	608
q_2: expri_1	4	106	498	608
q_3: expri_3	4	338	266	608
q_4: expri_5	4	490	114	608

표 7: 그룹 1 문항의 응답패턴

	X_1	X_2	X_3	X_4	Frequency
pattern1	1	0	0	0	106
pattern2	0	1	0	0	232
pattern3	0	1	1	0	152
pattern4	0	1	1	1	114
pattern5	−1	−1	−1	−1	4

표 5의 결과로 부터 응답패턴을 통해 군집분석 해보자. 그룹 1: q_1, q_2, q_3, q_4의 응답패턴을 고려하여 대표문항을 찾았다.

표 6은 그룹 1 문항에 대한 빈도표이다. 이 표에서 긍정적 응답(Yes)은 1, 부정적 응답(No)은 0, 결측값은 −1이다.

표 7은 608명 응답자에 대한 응답패턴이다. 이론적으로 $3^4 = 81$개의 응답패턴이 나오지만, 본 논문에서는 빈도가 0인 패턴은 무시했기 때문에 608명의 응답자는 5개의 응답패턴으로 그룹화 되었다.

첫 번째 응답패턴의 빈도 106은 106명의 응답자가 '이성교제 경험이 없다'라고 대답했음을 의미한다. 군집분석을 사용하여 응답패턴을 그룹화 하였고, 그 결과의 나무구조그림은 그림 2와 같다.

위의 결과에서 표 8처럼 응답자를 두 그룹으로 나누었을 때, 군집 1의 모든 $X_1 = 0$이고, $X_2 = 1$로 나타나고 있다. 이는 군집 1이 이성교제 경험이 없는 군집, 군집 2는 이성교제 경험이 있는 군집임을 의미한다. 군집 2는 결측값을 포함하고 있으므로 결측값은 $X_1 = 1$, 즉 '이성교제 경험이 있다'로 처리할 수 있다.

이와 같은 방법을 통해 응답자들을 전체 문항의 응답패턴으로 그룹화 할 수 없는 한계점을 해결할 수 있었다.

다른 문항그룹에 대해서도 위와 동일한 방법으로 대표문항을 찾는 시도를 하였다. 그 결과 범주가 두 개인 6개의 대표문항을 찾았다. 다음은 6개의 대표문항이다.

S_1 : experience　　S_3 : importance of character　　S_5 : ba (baby_bear)
S_2 : importance of face　　S_4 : ideal_together　　S_6 : sk (skin_kiss)

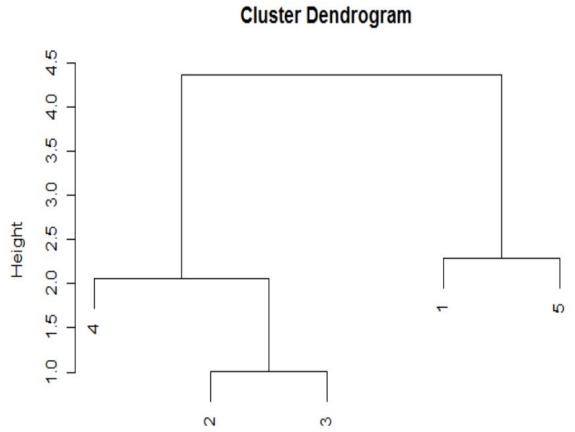

그림 2: *그룹 1 문항의 응답패턴을 이용하여 군집분석한 나무구조그림*

표 8: 그룹 1 문항의 응답패턴을 이용하여 군집분석한 결과표

군집1	(0, 1, 0, 0), (0, 1, 1, 0), (0, 1, 1, 1)
군집2	(1, 0, 0, 0), (−1, −1, −1, −1)

표 9: 각 문항의 엔트로피

	S_1	S_2	S_3	S_4	S_5	S_6
$H(X)$	0.472	0.472	0.597	0.367	0.521	0.665

4. 엔트로피를 이용한 항목 평가

원래 문항들의 집합을 $\boldsymbol{X}$, 대표 문항들의 집합을 $\boldsymbol{S} = \{S_1, \ldots, S_6\}$이라고 한다면 상호정보 $I(\boldsymbol{X};\boldsymbol{S})$는 다음과 같이 정의된다.

$$\begin{aligned} I(\boldsymbol{X};\boldsymbol{S}) &= \int f(\boldsymbol{X},\boldsymbol{S}) \log \frac{f(\boldsymbol{X},\boldsymbol{S})}{f(\boldsymbol{X})f(\boldsymbol{S})} \\ &= \int f(\boldsymbol{X},\boldsymbol{S}) \log \frac{f(\boldsymbol{X})}{f(\boldsymbol{X})f(\boldsymbol{S})} \\ &= H(\boldsymbol{S}), \end{aligned} \tag{4.1}$$

여기서, $\boldsymbol{S} \in \boldsymbol{X}$이므로 $f(\boldsymbol{X},\boldsymbol{S}) = f(\boldsymbol{X})$가 성립한다. $H(\boldsymbol{S})$는 대표 문항들의 결합 엔트로피이다. 즉, $I(\boldsymbol{X};\boldsymbol{S})$을 크게 하는 것은 $H(\boldsymbol{S})$를 크게 하는 문항들의 집합을 찾는 것과 같은 문제가 된다.

엔트로피를 이용하여 3절에서의 대표 문항을 평가해 보자. 엔트로피는 확률변수(random variable)에 대한 정보의 양으로 정의 되며 James와 Constance (2003)는 이것을 분산의 측도로 사용할 수 있다고 하였다. 각 문항의 엔트로피를 계산한 결과 S_6의 엔트로피가 가장 크다. 즉 S_6의 설명력이 가장 크다는 것을 의미한다. 6개 대표 문항의 엔트로피는 표 9와 같다.

두 개 문항의 결합 엔트로피를 고려해보자. 6개의 문항 중 설명력이 가장 큰 문항 S_6과 다른 문항과의 결합 엔트로피를 비교해 본 결과 문항 S_3, S_6의 결합 엔트로피가 가장 크게 나왔다. 여기서, 결합 엔트로피는 식 (4.2)에 의해 계산된다.

$$H(X,Y) = H(X) + H(Y|X). \tag{4.2}$$

표 10: 문항 S_6과 나머지 문항과의 결합 엔트로피

	S_1, S_6	S_2, S_6	S_3, S_6	S_4, S_6	S_5, S_6
$H(X, Y)$	1.135	1.138	1.262	1.033	1.184

표 11: 세 개 문항의 결합 엔트로피

	X	Y	Z	$H(X, Y, Z)$		X	Y	Z	$H(X, Y, Z)$
a	S_1	S_2	S_3	2.52	k	S_2	S_3	S_4	2.03
b	S_1	S_2	S_4	2.56	l	S_2	S_3	S_5	3.14
c	S_1	S_2	S_5	3.32	m	S_2	S_3	S_6	2.91
d	S_1	S_2	S_6	3.55	n	S_2	S_4	S_5	3.27
e	S_1	S_3	S_4	2.69	o	S_2	S_4	S_6	3.59
f	S_1	S_3	S_5	3.17	p	S_2	S_5	S_6	3.52
g	S_1	S_3	S_6	3.54	q	S_3	S_4	S_5	3.34
h	S_1	S_4	S_5	3.26	r	S_3	S_4	S_6	3.86
i	S_1	S_4	S_6	3.35	s	S_3	S_5	S_6	3.81
j	S_1	S_5	S_6	3.47	t	S_4	S_5	S_6	3.35

다음으로 세 개 문항에 대한 결합 엔트로피 $H(X, Y, Z)$를 고려해 보자. 세 개 문항의 조건부 엔트로피와 결합 엔트로피는 다음과 같이 계산되어진다.

$$\begin{aligned} H(X, Y, Z) &= H(X) + H(Y, Z|X) \\ &= H(X) + H(Y|X) + H(Z|Y, X). \end{aligned} \tag{4.3}$$

S_3, S_5, S_6의 결합 엔트로피보다 S_3, S_4, S_6의 결합 엔트로피가 더 크다. 즉 S_5의 엔트로피는 크지만 S_5가 S_3, S_6에 의해 설명되므로 $H(S_3, S_4, S_6)$는 $H(S_3, S_5, S_6)$보다 크다. 따라서 S_3, S_4, S_6이 데이터를 설명하는 설명력이 가장 크다고 할 수 있다.

결합 엔트로피를 이론적으로 구할 수는 있다. 하지만 세 개 문항 이상의 경우 분할표의 빈도수가 0인 경우가 발생하기 때문에 분포가 제대로 추정 되지 않는 문제가 발생한다. 따라서 결합 엔트로피를 이용하여 항목을 평가하는데 한계점을 가지고 있다.

5. 결론

본 논문은 상호정보와 엔트로피를 설문 자료에 적용하여 설문지 항목으로부터 대표문항을 찾는 방법과 평가하는 방법을 제안하였다.

범주가 여러 개인 설문문항을 범주가 두 개인 문항으로 만들어, 대표문항을 찾았으며 모든 문항의 응답패턴을 고려한 군집분석을 실시하였다. 유사문항을 찾는데 상호정보를 이용하여 다차원척도법을 실시하였고, 가까운 문항들의 응답패턴을 군집분석하여 응답자들을 응답패턴으로 그룹화 시켰다. 이 경우 원래 집단의 문항과 대표문항들 사이의 관련이 클수록 문항의 대표성이 잘 표현되었다고 할 수 있다.

응답패턴을 군집분석한 결과 결측값을 포함한 응답패턴들을 처리할 수 있었다. 이는 결측값의 보완 문제에도 이용할 수 있을 것으로 보인다.

대표문항을 평가하기 위해 엔트로피를 이용하였다. 두 문항의 결합엔트로피 결과는 만족스러웠지만 세 개 이상의 결합 엔트로피는 교차표에서 칸 빈도수가 0인 경우가 발생하기 때문에 분포를 추정하는데 문제가 발생하게 된다. 따라서 세 개 변수 이상에서 결합 엔트로피를 구하는 방법에 대한 더 많은 연구가 필요하며, 여러 변수의 결합 엔트로피와 상호정보 추정에 대한 연구가 요구된다.

참고 문헌

David, J. B., Fiona, S., Irini, M. and Jane, I. G. (2002). *The Analysis and Interpretation of Multivariate Data for Social Sicentists*, Chapman & Hall/CRC.

James, V. Z. and Constance, V. E. (2003). Uncertainty, entropy, variance and the effect of partial information, *Mathematical Statistics and Applications: Lecture Notes-Monograph Series*, **42**, 155–167.

Lee, S. C. and Huh, M. Y. (2003). A measure of association for complex data, *Computational Statistics and Data Analysis*, **44**, 211–222.

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

Silvey, S. D. (1964). On a measure of association, *The Annals of Mathematical Statistics*, **35**, 1157–1166.

Thomas, M. C. and Joy, A. T. (2006). *Elements of Information Theory*, Wiley-Interscience.

2010년 4월 접수; 2010년 6월 채택

A Method Finding Representative Questionare for Mutual Information and Entropy

Byong Su Choi[1,a], HyunJi Kim[b]

[a]Department of Multimedia Engineering, Hansung University
[b]Department of Statistics, Sungkyunkwan University

Abstract

A questionnaire may consist of duplicated or similar items. This study finds the duplicated or similar items by using the MDS and the cluster analysis of response patterns. By identifying the characteristics of the cluster, those items are combined into a representative item. The similarity of items is measured by the mutual information.

Keywords: Entropy, mutual information, cluster analysis, multidimensional scaling.

This research was financially supported by Hansung University in the year of 2009.

[1] Corresponding author: Professor, Department of Multimedia Engineering, Hansung University, 289, 2-Ga, Samseon-Dong, Seongbuk-Gu 136-792, Korea. E-mail: cbsroot@hansung.ac.kr