Journal of the Korean Data & Information Science Society
2017, **28**(1), 49–56

http://dx.doi.org/10.7465/jkdi.2017.28.1.49
한국데이터정보과학회지

흥미도 측도 관점에서 상대적 인과 강도의 고찰

박희창[1]

[1]창원대학교 통계학과

접수 2016년 12월 28일, 수정 2017년 1월 9일, 게재확정 2017년 1월 10일

요 약

빅 데이터를 분석하기 위한 기법 중에서 연관성 규칙은 여러 가지 연관성 평가 기준을 이용하여 항목들 간에 연관성 유무를 탐색하는 기법이다. 이러한 연관성 규칙 기법은 규칙의 생성 방향에 따라 정과 부, 그리고 역의 연관성 규칙 등이 있다. 본 논문에서는 여러 가지 상대적 인과 강도를 흥미도 측도의 관점에서 어떤 유형의 연관성 규칙에 적용 가능한 지를 탐색하는 동시에 기존의 기본적인 평가 측도 증에서 여러 가지 유형의 신뢰도들과의 관계를 규명하고자 하였다. 그 결과, 후항변수가 발생할 비율이 0.5 이상이면 Good이 제안한 측도 (RCS_{IJ1})가 Lewis가 제안한 측도 (RCS_{LR1}) 보다 값의 변화폭이 더 크므로 RCS_{IJ1}이 더 바람직한 측도가 되며, 그 비율이 0.5 미만이면 RCS_{LR1}이 더 바람직하다고 할 수 있다.

주요용어: 데이터 마이닝, 상대적 인과 강도, 연관성 규칙, 인과 강도, 흥미도 측도.

1. 서론

빅 데이터는 데이터 처리와 관련된 일련의 새로운 방식을 지칭하는 한편, 이를 통한 경제 측면에서도 조망되어 왔으며, 이를 활용한 경제 가치의 창출은 국가 정책에서도 우선순위에 놓여 있다 (Kim과 Lee, 2016). 이러한 빅 데이터는 단순히 수집·축적하는 것이 중요한 것이 아니라 구조화되지 않은 대규모 데이터 속에서 숨겨진 패턴을 찾아내고 여러 변수들을 통합적으로 고려하면서 창의적으로 해석할 수 있는 분석능력이 더 중요해지고 있다. (Kang 등, 2012). 빅 데이터 분석을 위해 연관성 규칙 (association rule)을 생성할 때 이용되는 규칙의 평가 기준인 흥미도 측도 (interestingness measure)를 분류하면 크게 객관적인 측도 (objective measure)와 주관적인 측도 (subjective measure)로 나누어진다(Silberschatz와 Tuzhilin, 1996; Freitas, 1999). 이들 각각에 대한 대표 논문으로는 Hilderman과 Hamilton (2000)과 Bing 등 (2000)이 있다. 또한 Park (2015a)에 의하면 연관성 규칙의 생성 방향에 따라 여러 가지 형태로 나타나는데 먼저 양의 규칙 (positive rule)은 특정 항목 X가 발생되면 다른 항목 Y가 발생되는 규칙을 찾아내는 것이다. 음의 규칙 (negative rule)은 항목 X가 발생되면 항목 Y는 발생되지 않는 규칙을 발견하는 것이며, 역의 규칙 (inverse rule)은 항목 X가 발생되지 않으면 항목 Y가 발생되지 않는다는 규칙을 찾아내는 것이다.

연관성 규칙의 탐색을 위한 기본적인 흥미도 측도에는 지지도 (support), 신뢰도 (confidence), 그리고 향상도 (lift) 등이 있다 (Park, 2016). Park (2016)에서 지적한 바와 같이 이러한 기본적인 측도를 이용하여 의미 있는 연관성 규칙을 찾는 데에는 상당한 문제점이 있으므로 그동안 이를 해결하기 위해 많은 논문이 발표되었다 (Silberschatz와 Tuzhilin, 1996; Tan 등, 2002; Ahn과 Kim, 2003; Jin 등,

[1] 교신저자: (641-773) 경상남도 창원시 의창구 사림동 9번지, 창원대학교 통계학과, 교수.
E-mail: hcpark@changwon.ac.kr

2011; Park, 2014a; Park, 2014b; Park, 2015b; Park, 2015c). 이들 연구에 이어 본 논문에서는 여러 가지 상대적 인과 강도 (relatively causal strength measures)를 흥미도 측도의 관점에서 어떤 유형의 연관성 규칙에 적용 가능한 지를 탐색하는 동시에 기존의 측도인 여러 가지 유형의 신뢰도와의 관계를 규명하고자 한다.

2. 상대적 인과 강도

Fitelson과 Hitchcock (2011)은 그동안 여러 학자들에 의해 제안된 상대적 인과 강도에 대해 확률을 이용하여 설명하였으며, 인과 관계의 관점에서 논의한 바 있다. 이 절에서는 이들 측도에 대해 연관성 규칙에서 활용할 경우 어떤 유형에 적용 가능한 지를 알아보고 기본적인 연관성 평가 기준 중의 하나인 신뢰도와의 관계를 규명하고자 한다. 이를 위해 Park (2015a)에서와 같이 연관성 규칙에서 이용하는 Table 2.1과 같은 분할표를 이용하고자 한다.

Table 2.1 2×2contingency table by proportions

		Y		Total
		1	0	
X	1	a	b	$a+b$
	0	c	d	$c+d$
Total		$a+c$	$b+d$	n

Table 2.1에서 '1'은 발생을 의미하고 '0'은 비발생을 의미하며, 본 논문에서 나타나는 수식에 포함된 조건부 확률을 기술하면 다음과 같다.

$$P(Y|X) = P(Y=1|X=1),\ P(\bar{Y}|X) = P(Y=0|X=1),$$
$$P(Y|\bar{X}) = P(Y=1|X=0),\ P(\bar{Y}|\bar{X}) = P(Y=0|X=0).$$

Table 2.1을 이용하여 상대적 인과적 강도를 나타내는 측도들을 정리하면 다음과 같다.

$$RCS_C = \frac{P(Y|X) - P(Y|\bar{X})}{P(\bar{Y}|\bar{X})} = \frac{ad-bc}{(a+b)(c+d)} \Big/ \left(\frac{d}{c+d}\right) \tag{2.1}$$

$$RCS_{LR1} = \frac{P(Y|X) - P(Y|\bar{X})}{P(Y|X) + P(Y|\bar{X})} = \frac{ad-bc}{(a+b)(c+d)} \Big/ \left(\frac{a}{a+b} + \frac{c}{c+d}\right) \tag{2.2}$$

$$RCS_{LR2} = \frac{P(Y|X) - P(Y|\bar{X})}{P(Y|X)} = \frac{ad-bc}{(a+b)(c+d)} \Big/ \left(\frac{a}{a+b}\right) \tag{2.3}$$

$$RCS_{IJ1} = \frac{P(\bar{Y}|\bar{X}) - P(\bar{Y}|X)}{P(\bar{Y}|\bar{X}) + P(\bar{Y}|X)} = \frac{ad-bc}{(a+b)(c+d)} \Big/ \left(\frac{b}{a+b} + \frac{d}{c+d}\right) \tag{2.4}$$

$$RCS_{IJ2} = \frac{P(\bar{Y}|\bar{X}) - P(\bar{Y}|X)}{P(\bar{Y}|\bar{X})} = \frac{ad-bc}{(a+b)(c+d)} \Big/ \left(\frac{d}{c+d}\right) \tag{2.5}$$

식 (2.1)의 RCS_C는 Cheng (1997)에 의해 제안된 측도이고, 식 (2.2)의 RCS_{LR1}과 식 (2.3)의 RCS_{LR2}는 Lewis (1986)에 의해 제안된 것이며, 식 (2.4)의 RCS_{IJ1}과 식 (2.5)의 RCS_{IJ2}는 Good (1961, 1962)에 의해 제안된 측도이다. 또한 $P(Y|X)$는 양의 신뢰도를 의미하고, $P(\bar{Y}|\bar{X})$는 역의 신뢰도이며, $P(Y|\bar{X})$와 $P(\bar{Y}|X)$는 음의 신뢰도를 의미한다. 식 (2.5)의 RCS_{IJ2}는 식 (2.1)의 RCS_C와 동

일한 것으로 나타났다. 이들 측도들은 절대적 인과 강도를 나타내는 측도들 중의 하나인 Eells (1991)가 제안한 CS_E의 수식으로 나타낼 수도 있다.

$$CS_E = P(Y|X) - P(Y|\bar{X}) = \frac{ad - bc}{(a+b)(c+d)}$$

측도 CS_E는 의학 분야에서 이용되고 있는 기여위험률 (attributable risk)과 같은 측도이며, Ahn과 Kim (2003)이 신뢰도의 단점을 극복하기 위해 이를 순수 신뢰도 (net confidence)로 명명하여 연관성 규칙 분야에 제안한 바 있다. 이 측도의 의미는 항목 X가 발생되지 않은 데이터베이스 내에서도 항목 Y가 발생되는 경우가 많으면 이들 항목간의 관계는 적은 것을 의미한다. 각 상대적 인과 강도를 신뢰도와 CS_E와의 관계식으로 나타낼 수 있다.

$$RCS_C = \frac{CS_E}{P(\bar{Y}|\bar{X})}$$

$$RCS_{LR1} = \frac{CS_E}{P(Y|X) + P(Y|\bar{X})}$$

$$RCS_{LR2} = \frac{CS_E}{P(Y|X)}$$

$$RCS_{IJ1} = \frac{CS_E}{P(\bar{Y}|X) + P(\bar{Y}|\bar{X})}$$

$$RCS_{IJ2} = \frac{CS_E}{P(\bar{Y}|\bar{X})}$$

이 식에서 보는 바와 같이 RCS_C는 역의 신뢰도에 대한 CS_E의 상대적 크기, RCS_{LR1}은 양과 음의 신뢰도의 합에 대한 CS_E의 상대적 크기, RCS_{LR2}는 양의 신뢰도에 대한 CS_E의 상대적 크기, 그리고 RCS_{IJ1}은 역과 음의 신뢰도의 합에 대한 CS_E의 상대적 크기를 나타내는 측도이다. 또한 RCS_{IJ2}는 역의 신뢰도에 대한 역의 순수 신뢰도의 상대적 크기를 나타내지만 이는 결국 역의 신뢰도에 대한 CS_E의 상대적 크기를 의미하므로 RCS_C와 동일한 측도이다.

3. 예제에 의한 고찰

본 절에서는 상대적 인과 강도를 나타내는 측도들의 변화하는 양상과 여러 유형의 신뢰도와의 비교, 그리고 가장 바람직한 연관성 평가 기준을 선정하기 위해 Park (2015a)의 실험과 유사한 자료를 활용하고자 한다. 먼저 i의 값이 증가하는 경우, 즉 동시 발생 빈도의 값과 동시 비 발생 빈도의 값이 증가하고 불일치빈도의 값이 감소하는 경우인 Table 3.1의 데이터에 대해 고려한다.

Table 3.1 Simulation data (1)

		Y		Total
		1	0	
X	1	i	$50-i$	50
	0	$30-i$	$i+20$	50
Total		30	70	100

Table 3.1을 이용하여 여러 형태의 신뢰도와 상대적 인과 강도를 계산하면 Table 3.2의 결과를 얻는다. 이 표에서 a는 동시 발생 빈도 $n(X=1, Y=1)$, d는 동시 비 발생 빈도 $n(X=0, Y=0)$, 그리고

b와 c는 각각 불일치 빈도 $n(X=0,Y=1)$와 $n(X=1,Y=0)$을 의미한다. a와 d가 증가하고 b와 c가 감소하게 되면 양의 신뢰도 $P(Y|X)$와 역의 신뢰도 $P(\bar{Y}|\bar{X})$, 그리고 CS_E는 증가하고 있는 반면에 음의 신뢰도인 $P(Y|\bar{X})$와 $P(\bar{Y}|X)$는 감소하고 있는 것으로 나타났다. 그리고 상대적 인과 강도를 나타내는 측도들은 이 경우에 모두 증가하였다. 이를 좀 더 구체적으로 살펴보면 RCS_{LR1}은 분모인 양과 음의 신뢰도의 합이 모든 사례에서 0.6으로 계산되었으므로 CS_E가 증가하면 증가하는 것으로 나타났다. RCS_{LR2}는 분모인 양의 신뢰도와 분자인 CS_E의 값이 모두 증가하고 있으나 증가하는 폭의 차이로 인하여 증가하였다. RCS_C는 분모인 역의 신뢰도가 증가하고 분자인 CS_E도 동시에 증가하나 이 또한 증가하는 폭이 분자가 더 크므로 증가하는 것으로 나타났다. RCS_{IJ1}은 분모인 역과 음의 신뢰도의 합이 모든 사례에서 1.4로 고정된 값을 가지므로 CS_E가 증가함에 따라 증가하였다. 또한 상대적 측도들은 모두 양, 0, 그리고 음의 값 모두를 취하는 것으로 나타났으며, $ad-bc$의 값이 0이면 항목 간에는 연관성이 없다고 할 수 있다. 만약 $ad-bc$의 값이 음이면 이들은 음의 값을 갖게 되어 항목 간에 음의 연관성이 있으며, 그 값이 양이면 이들 측도들은 모두 양의 값을 갖게 되어 양의 연관성이 존재한다고 할 수 있다. 상대적 인과 강도를 나타내는 측도 간들의 비교를 위해 각 측도가 취할 수 있는 값의 구간을 살펴보면 RCS_{LR1}과 RCS_{IJ1}은 $[-1,\ 1]$의 범위를 가지는 반면에 다른 측도들은 이 범위를 벗어나는 값도 취하는 것으로 나타났다. 이를 좀 더 구체적으로 살펴보면 RCS_{LR1}과 RCS_{IJ1}은 식 (2.2)와 식 (2.3)에 의해 그 절대값의 크기가 항상 1보다 작은 값을 갖게 되는 반면에 RCS_{LR2}는 $P(Y|X) \geq P(Y|\bar{X})$이면 $[0,\ 1]$의 값을 취하고 그 반대이면 $[-\infty,\ 0]$의 값을 취한다. 또한 RCS_{IJ2}는 $P(\bar{Y}|\bar{X}) \geq P(\bar{Y}|X)$이면 $[0,\ 1]$의 값을 취하고 그 반대이면 $[-\infty,\ 0]$의 값을 취한다. 따라서 RCS_{LR1}과 RCS_{IJ1}의 두 측도가 연관성 평가 기준으로 바람직하다고 볼 수 있다. 또한 변화하는 폭을 살펴보면 RCS_{IJ1}에 비해 RCS_{LR1}이 더 크므로 연관성 정도를 더 잘 구분할 수 있으므로 상대적 인과 강도를 나타내는 측도들 중에서는 RCS_{LR1}이 보다 바람직한 측도라고 생각된다.

Table 3.2 Comparison of relatively causal strength measures by data (1)

a	b	c	d	$P(Y\|X)$	$P(Y\|\bar{X})$	$P(\bar{Y}\|X)$	$P(\bar{Y}\|\bar{X})$	CS_E	RCS_{LR1}	RCS_{LR2}	RCS_C	RCS_{IJ1}
1	49	29	21	0.020	0.580	0.980	0.420	-0.560	-0.933	-28.000	-1.333	-0.400
2	48	28	22	0.040	0.560	0.960	0.440	-0.520	-0.867	-13.000	-1.182	-0.371
3	47	27	23	0.060	0.540	0.940	0.460	-0.480	-0.800	-8.000	-1.043	-0.343
4	46	26	24	0.080	0.520	0.920	0.480	-0.440	-0.733	-5.500	-0.917	-0.314
5	45	25	25	0.100	0.500	0.900	0.500	-0.400	-0.667	-4.000	-0.800	-0.286
6	44	24	26	0.120	0.480	0.880	0.520	-0.360	-0.600	-3.000	-0.692	-0.257
7	43	23	27	0.140	0.460	0.860	0.540	-0.320	-0.533	-2.286	-0.593	-0.229
8	42	22	28	0.160	0.440	0.840	0.560	-0.280	-0.467	-1.750	-0.500	-0.200
9	41	21	29	0.180	0.420	0.820	0.580	-0.240	-0.400	-1.333	-0.414	-0.171
10	40	20	30	0.200	0.400	0.800	0.600	-0.200	-0.333	-1.000	-0.333	-0.143
11	39	19	31	0.220	0.380	0.780	0.620	-0.160	-0.267	-0.727	-0.258	-0.114
12	38	18	32	0.240	0.360	0.760	0.640	-0.120	-0.200	-0.500	-0.188	-0.086
13	37	17	33	0.260	0.340	0.740	0.660	-0.080	-0.133	-0.308	-0.121	-0.057
14	36	16	34	0.280	0.320	0.720	0.680	-0.040	-0.067	-0.143	-0.059	-0.029
15	35	15	35	0.300	0.300	0.700	0.700	0.000	0.000	0.000	0.000	0.000
16	34	14	36	0.320	0.280	0.680	0.720	0.040	0.067	0.125	0.056	0.029
17	33	13	37	0.340	0.260	0.660	0.740	0.080	0.133	0.235	0.108	0.057
18	32	12	38	0.360	0.240	0.640	0.760	0.120	0.200	0.333	0.158	0.086
19	31	11	39	0.380	0.220	0.620	0.780	0.160	0.267	0.421	0.205	0.114
20	30	10	40	0.400	0.200	0.600	0.800	0.200	0.333	0.500	0.250	0.143
21	29	9	41	0.420	0.180	0.580	0.820	0.240	0.400	0.571	0.293	0.171
22	28	8	42	0.440	0.160	0.560	0.840	0.280	0.467	0.636	0.333	0.200
23	27	7	43	0.460	0.140	0.540	0.860	0.320	0.533	0.696	0.372	0.229
24	26	6	44	0.480	0.120	0.520	0.880	0.360	0.600	0.750	0.409	0.257
25	25	5	45	0.500	0.100	0.500	0.900	0.400	0.667	0.800	0.444	0.286
26	24	4	46	0.520	0.080	0.480	0.920	0.440	0.733	0.846	0.478	0.314
27	23	3	47	0.540	0.060	0.460	0.940	0.480	0.800	0.889	0.511	0.343
28	22	2	48	0.560	0.040	0.440	0.960	0.520	0.867	0.929	0.542	0.371
29	21	1	49	0.580	0.020	0.420	0.980	0.560	0.933	0.966	0.571	0.400

이번에는 j의 값이 증가하는 경우, 즉 두 종류의 불일치 빈도의 값은 커지나 동시 발생 빈도 및 동시 비 발생 빈도의 값은 줄어드는 경우를 고려하고자 한다.

Table 3.3 Simulation data (2)

		Y		Total
		1	0	
X	1	$50-j$	j	50
	0	$j+20$	$30-j$	50
Total		70	30	100

앞에서와 마찬가지로 Table 3.3 이용하여 신뢰도와 상대적 인과 강도를 나타내는 측도들을 계산하면 Table 3.4의 결과를 얻는다.

Table 3.4 Comparison of relatively causal strength measures by data (2)

a	b	c	d	$P(Y\|X)$	$P(Y\|\bar{X})$	$P(\bar{Y}\|X)$	$P(\bar{Y}\|\bar{X})$	CS_E	RCS_{LR1}	RCS_{LR2}	RCS_C	RCS_{IJ1}
49	1	21	29	0.980	0.420	0.020	0.580	0.560	0.400	0.571	0.966	0.933
48	2	22	28	0.960	0.440	0.040	0.560	0.520	0.371	0.542	0.929	0.867
47	3	23	27	0.940	0.460	0.060	0.540	0.480	0.343	0.511	0.889	0.800
46	4	24	26	0.920	0.480	0.080	0.520	0.440	0.314	0.478	0.846	0.733
45	5	25	25	0.900	0.500	0.100	0.500	0.400	0.286	0.444	0.800	0.667
44	6	26	24	0.880	0.520	0.120	0.480	0.360	0.257	0.409	0.750	0.600
43	7	27	23	0.860	0.540	0.140	0.460	0.320	0.229	0.372	0.696	0.533
42	8	28	22	0.840	0.560	0.160	0.440	0.280	0.200	0.333	0.636	0.467
41	9	29	21	0.820	0.580	0.180	0.420	0.240	0.171	0.293	0.571	0.400
40	10	30	20	0.800	0.600	0.200	0.400	0.200	0.143	0.250	0.500	0.333
39	11	31	19	0.780	0.620	0.220	0.380	0.160	0.114	0.205	0.421	0.267
38	12	32	18	0.760	0.640	0.240	0.360	0.120	0.086	0.158	0.333	0.200
37	13	33	17	0.740	0.660	0.260	0.340	0.080	0.057	0.108	0.235	0.133
36	14	34	16	0.720	0.680	0.280	0.320	0.040	0.029	0.056	0.125	0.067
35	15	35	15	0.700	0.700	0.300	0.300	0.000	0.000	0.000	0.000	0.000
34	16	36	14	0.680	0.720	0.320	0.280	-0.040	-0.029	-0.059	-0.143	-0.067
33	17	37	13	0.660	0.740	0.340	0.260	-0.080	-0.057	-0.121	-0.308	-0.133
32	18	38	12	0.640	0.760	0.360	0.240	-0.120	-0.086	-0.188	-0.500	-0.200
31	19	39	11	0.620	0.780	0.380	0.220	-0.160	-0.114	-0.258	-0.727	-0.267
30	20	40	10	0.600	0.800	0.400	0.200	-0.200	-0.143	-0.333	-1.000	-0.333
29	21	41	9	0.580	0.820	0.420	0.180	-0.240	-0.171	-0.414	-1.333	-0.400
28	22	42	8	0.560	0.840	0.440	0.160	-0.280	-0.200	-0.500	-1.750	-0.467
27	23	43	7	0.540	0.860	0.460	0.140	-0.320	-0.229	-0.593	-2.286	-0.533
26	24	44	6	0.520	0.880	0.480	0.120	-0.360	-0.257	-0.692	-3.000	-0.600
25	25	45	5	0.500	0.900	0.500	0.100	-0.400	-0.286	-0.800	-4.000	-0.667
24	26	46	4	0.480	0.920	0.520	0.080	-0.440	-0.314	-0.917	-5.500	-0.733
23	27	47	3	0.460	0.940	0.540	0.060	-0.480	-0.343	-1.043	-8.000	-0.800
22	28	48	2	0.440	0.960	0.560	0.040	-0.520	-0.371	-1.182	-13.000	-0.867
21	29	49	1	0.420	0.980	0.580	0.020	-0.560	-0.400	-1.333	-28.000	-0.933

Table 3.4에서 a, b, c, d는 Table 3.2에서와 동일하게 동시 발생 빈도, 불일치 빈도, 동시 비 발생 빈도를 나타낸다. Table 3.4에서 보는 바와 같이 불일치 빈도인 b와 c가 증가하고 동시 발생 빈도 a와 동시 비 발생 빈도 d가 감소하게 되면 앞의 결과와 반대로 음의 신뢰도 $P(Y|\bar{X})$와 $P(\bar{Y}|X)$는 증가하고 양의 신뢰도 $P(Y|X)$와 역의 신뢰도 $P(\bar{Y}|\bar{X})$, 그리고 CS_E는 감소하였다. 그리고 상대적 인과 강도를 나타내는 측도들은 이 경우에 모두 감소하는 것으로 나타났다. 이를 좀 더 구체적으로 살펴보면 RCS_{LR1}은 분모인 양과 음의 신뢰도의 합이 모든 경우가 1.4로 계산되었으므로 CS_E가 감소함에 따라 감소하는 것으로 나타났다. RCS_{LR2}는 분모인 양의 신뢰도와 분자인 CS_E의 값이 모두 감소하고 있으나 감소폭의 차이로 인하여 감소하였다. RCS_C는 분모인 역의 신뢰도가 감소하고 분자인 CS_E도 동시에 감소하나

감소폭이 분자가 더 크므로 감소하는 것으로 나타났다. RCS_{IJ1}은 분모인 역과 음의 신뢰도의 합이 모든 경우에 0.6의 고정된 값으로 나타났으므로 CS_E가 감소함에 따라 감소하는 것으로 나타났다. 여기서도 상대적 인과강도들은 모두 양, 0, 그리고 음의 값 모두를 취하는 것으로 나타났으며, RCS_{LR1}과 RCS_{IJ1}은 $[-1,\ 1]$의 범위를 가지는 반면에 다른 측도들은 이 범위를 벗어나는 값도 취하는 것으로 나타났다. 따라서 여러 측도 중에서 RCS_{LR1}과 RCS_{IJ1}이 연관성 평가 기준으로 바람직하다고 볼 수 있는데, 이들의 변화하는 폭을 비교해보면 RCS_{LR1}에 비해 RCS_{IJ1}이 더 크므로 연관성 정도를 더 잘 구분할 수 있으므로 상대적 인과 강도를 나타내는 측도들 중에서는 RCS_{IJ1}이 더 바람직한 측도라고 생각된다.

이에 추가하여 항목 X의 발생과 비 발생의 비, 그리고 Y의 발생과 비 발생의 비를 여러 가지로 다양하게 실험해보았는데, 그 결과는 위에서 기술한 것과 동일하게 나타났다.

4. 결론

최근 IT 분야의 화두는 단연 빅 데이터라고 할 수 있을 것이다. 이러한 빅 데이터를 단순히 수집하고 축적하는 것이 중요한 것이 아니라 구조화되지 않은 대규모 데이터 속에서 숨겨진 패턴을 찾아내고 여러 변수들을 통합적으로 고려하면서 창의적으로 해석할 수 있는 분석능력이 더 중요해지고 있다 (Kang 등, 2012). 본 논문에서는 이러한 패턴을 찾아내기 위한 기법 중의 하나인 연관성 규칙과 관련된 흥미도 측도의 관점에서 상대적 인과 강도의 측도를 탐색하였다. 그 결과, 동시 발생 빈도와 동시 비 발생 빈도가 증가하고, 불일치 빈도가 감소함에 따라 양의 신뢰도와 역의 신뢰도, 증가하는 반면에 음의 신뢰도는 감소하는 것으로 나타났다. 그리고 상대적 인과 강도를 나타내는 측도들은 이 경우에 모두 증가하였다. 또한 동시 발생 빈도와 동시 비 발생 빈도가 감소하고, 불일치 빈도가 증가하는 경우에는 이와 반대로 나타났다. 상대적 인과 강도들을 전체적으로 살펴보면 후항변수가 발생할 비율이 0.5 이상이면 RCS_{IJ1}이 RCS_{LR1}보다 값의 변화폭이 더 크므로 RCS_{IJ1}이 더 바람직한 측도가 되며, 그 비율이 0.5 미만이면 RCS_{LR1}이 더 바람직하다고 할 수 있다.

References

Ahn, K. and Kim, S. (2003). A new interstingness measure in association rules mining. *Journal of the Korean Institute of Industrial Engineers*, **29**, 41-48.

Bing Liu, B., Hsu, W., Chen, S. and Ma, Y. (2000). Analyzing the subjective interestingness of association rules. *IEEE Intelligent Systems*, **15**, 47-55.

Cheng, P. (1997). From covariation to causation: A causal power theory. *Psychological Review*, **104**, 367-405.

Eells, E. (1991). *Probabilistic causality*, Cambridge University Press, U.K.

Fitelson, B. and Hitchcock, C. (2011). Probabilistic measures of causal strength. *Causality in the sciences*, Oxford University Press, Oxford, 600-627.

Freitas, A (1999). On rule interestingness measures. *Knowledge-based System*, **12**, 309-315.

Good, I. J. (1961). A causal calculus I. *British Journal for the Philosophy of Science*, **11**, 305-18.

Good, I. J. (1962). A causal calculus II. *British Journal for the Philosophy of Science*, **12**, 43-51.

Hilderman, R. J. and Hamilton H. J. (2000). Applying objective interestingness measures in data mining systems. *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, Springer-Verlag, London, UK, 432-439.

Jin, D. S., Kang, C., Kim, K. K. and Choi, S. B. (2011). CRM on travel agency using association rules. *Journal of the Korean Data Analysis Society*, **13**, 2945-2952.

Kang, M., Kim, S. and Park, S. (2012). Analysis and utilization of big data. *Communications of the Korean Institute of Information Scientists and Engineers*, **30**, 25-32.

Kim, H. and Lee, M. (2016). Big data and entertainment content: Case studies and prospects. *Journal of Internet Computing and Services*, **17**, 109-118.

Lewis, D. (1986). Postscripts to causation. *Philosophical Papers*, **2**, 173-213.

Park, H. C. (2014a). Comparison of cosine family similarity measures in the aspect of association rule. *Journal of the Korean Data Analysis Society*, **16**, 729-737.

Park, H. C. (2014b). Comparison of confidence measures useful for classification model building. *Journal of the Korean Data & Information Science Society*, **25**, 1-7.

Park, H. C. (2015a). Proposition of balanced comparative confidence considering all available diagnostic tools. *Journal of the Korean Data & Information Science Society*, **26**, 611-618.

Park, H. C. (2015b). Comparison study of symmetric confirmation measures and probabilistic interestingness measure. *Journal of the Korean Data Analysis Society*, 17, 749-758.

Park, H. C. (2015c). A study on the bounds of PIM based similarity measures with AMP. *Journal of the Korean Data Analysis Society*, **17**, 1839-1847.

Park, H. C. (2016). Signed Hellinger measure for directional association. *Journal of the Korean Data & Information Science Society*, **27**, 353-362.

Silberschatz, A. and Tuzhilin, A. (1996). What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge Data Engineering*, **8**, 970-974.

Tan, P. N., Kumar, V. and Srivastava, J. (2002). Selecting the right interestingness measure for association patterns. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, New York, USA, 32-41.

Journal of the Korean Data & Information Science Society
2017, **28**(1), 49–56

http://dx.doi.org/10.7465/jkdi.2017.28.1.49
한국데이터정보과학회지

A study on the relatively causal strength measures in a viewpoint of interestingness measure

Hee Chang Park[1]

[1]Department of Statistics, Changwon National University

Received 28 December 2016, revised 9 January 2017, accepted 10 January 2017

Abstract

Among the techniques for analyzing big data, the association rule mining is a technique for searching for relationship between some items using various relevance evaluation criteria. This associative rule scheme is based on the direction of rule creation, and there are positive, negative, and inverse association rules. The purpose of this paper is to investigate the applicability of various types of relatively causal strength measures to the types of association rules from the point of view of interestingness measure. We also clarify the relationship between various types of confidence measures. As a result, if the rate of occurrence of the posterior item is more than 0.5, the first measure (RCS_{IJ1}) proposed by Good (1961) is more preferable to the first measure (RCS_{LR1}) proposed by Lewis (1986) because the variation of the value is larger than that of RCS_{LR1}, and if the ratio is less than 0.5, $RCS_{LR1}$is more preferable to RCS_{IJ1}.

Keywords: Association rule, causal strength, data mining, interestingness measure, relatively causal strength measure.

[1] Professor, Department of Statistics, Changwon National University, Gyeongnam 641-773, Korea. E-mail: hcpark@changwon.ac.kr