Journal of the Korean Data & Information Science Society
2014, **25**(4), 857–868

http://dx.doi.org/10.7465/jkdi.2014.25.4.857
한국데이터정보과학회지

인과적 확인 측도에 의한 연관성 규칙 탐색

박희창[1]

[1]창원대학교 통계학과

접수 2014년 6월 16일, 수정 2014년 7월 10일, 게재확정 2014년 7월 15일

요 약

대량의 데이터로부터 과거에 알려지지 않았던 유용한 정보를 발견하는 기술인 데이터 마이닝 기법은 오늘날 빅 데이터 시대에 가장 대표적인 분석 기법이라고 할 수 있다. 이들 중에서도 연관성 규칙은 지지도, 신뢰도, 향상도 등의 여러 가지 흥미도 측도를 기반으로 하여 항목들 간의 관련성을 찾아내는 것이다. 그러나 기본적인 연관성 평가 기준만으로는 두 항목 간의 인과관계를 설명할 수 없을 뿐만 아니라 연관성의 방향도 파악할 수 없다. 본 논문에서는 이러한 문제를 해결하기 위해 인과적 확인 연관성 평가 기준을 제안하는 동시에, 제안한 평가 기준들이 흥미도 측도의 조건을 충족하는지의 여부를 점검하였다. 본 논문에서 제안한 인과적 확인 향상도는 세 가지 조건 모두를 만족하는 것으로 입증되었다. 인과적 확인 지지도와 인과적 확인 신뢰도는 동시 발생 확률의 값에 따라 단조 증가하는 조건과 각 항목의 주변 확률의 값에 따라 단조 감소하는 조건은 만족하였다. 또한 예제를 통해 기본적인 연관성 평가 기준과 인과적 연관성 평가 기준, 그리고 인과적 확인 연관성 평가 기준을 비교해 본 결과, 본 논문에서 제안하는 인과적 확인 측도들이 다른 평가 기준에 비해 가장 바람직한 측도라는 사실을 파악하였다.

주요용어: 데이터 마이닝, 연관성 규칙, 인과적 확인 신뢰도, 인과적 확인 지지도, 인과적 확인 향상도.

1. 서론

오늘날 정보기술이 급속도로 발전함에 따라 신속하고도 정확하게 데이터의 저장, 가공, 그리고 지식의 생성을 수행할 수 있게 됨으로 인해 데이터 마이닝 분야에서 개발된 이론적 연구의 성과들을 실제 비즈니스에 적용하여 수익으로 연결하고자 하는 시도가 활발하게 이루어지고 있다 (Kim, 2008). 데이터 마이닝은 대용량 데이터베이스로부터 그 안에 내재하고 있는 규칙, 패턴, 그리고 관계 등을 발견하여 의미 있는 지식을 창출해 내는 일련의 과정을 의미한다. 이러한 데이터 마이닝 기법은 장바구니 분석과 교차판매뿐만 아니라 금융 및 보험업, 보건 및 의학, 제조업, 이미지 분석 등 다양한 분야에서 적용되고 있다. 특히 데이터 마이닝 기법들 중에서 활발하게 연구되고 있는 연관성 규칙 기법은 항목들 간의 지지도, 신뢰도, 향상도 등의 연관성 평가 기준들을 기반으로 하여 엄청난 크기의 빅 데이터에 내재되어 있는 항목들 간의 관련성을 탐색하는 데 활용되고 있다 (Park, 2012b). Agrawal 등 (1993)에 의해 처음 제안된 연관성 규칙 기법은 그 이후로 많은 학자들에 의해 연구가 진행되어 왔다. 이들을 크게 빈발항목을 찾는 연구 (Han과 Fu, 1999; Srikant 등, 1997; Cai 등, 1998; Liu 등, 1999)와 수행속도를 향상시키기 위한 연구(Agrawal과 Srikant, 1994;Pasquier 등, 1999; Han 등, 2000; Pei 등, 2000; Saygin

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

등, 2002)로 분류된다. 또한 연관성 규칙에 대한 국내 연구에는 Lim 등 (2010), Cho와 Park (2011a, 2011b), Jin 등 (2011), Kim 등 (2013), 그리고 Park (2012a, 2012b, 2013) 등이 있다.

본 논문에서는 인과적 확인 연관성 규칙 (causally confirmed association rule)의 평가 기준을 제안하고자 한다. 이 중에서 인과적 확인 신뢰도 (causally confirmed confidence)는 Kodratoff (2000)와 Berzal 등 (2005)이 제안한 것을 활용하며, 이들의 아이디어를 확장하여 인과적 확인 지지도 (causally confirmed support)와 인과적 확인 향상도 (causally confirmed lift)를 제안함으로써 기본적인 연관성 평가 기준과 인과적 연관성 평가 기준을 대체할 수 있는 인과적 확인 연관성 규칙의 평가 기준을 제시하고자 한다. 또한 이들이 Piatetsky-Shapiro (1991)가 제안한 흥미도 측도의 조건을 충족하는지의 여부를 점검한 후, 예제를 통하여 인과적 확인 연관성 규칙의 유용성에 대해 알아보고자 한다.

2. 인과적 확인 연관성 측도의 제안

본 절에서는 Park (2012a)에서와 같이 Table 2.1과 같은 2×2분할표를 활용하여 인과적 확인 연관성 규칙 평가 기준을 제안하고, 이들과 기본적인 연관성 평가 기준 및 인과적 연관성 평가 기준에 대해 논의하고자 한다.

Table 2.1 2×2contingency table

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

기본적으로 연관성 규칙을 평가하는 기준에는 지지도 ($supp(X \Rightarrow Y) = a/n$), 신뢰도 ($conf(X \Rightarrow Y) = a/(a+b)$), 향상도 ($lift(X \Rightarrow Y) = na/[(a+b)(a+c)]$) 등이 있다. 이들 기본적인 연관성 평가 기준으로는 전항과 후항의 인과관계를 설명할 수 없어서 Kodratoff (2000)와 Berzal 등 (2005)이 $Y^c \Rightarrow X^c$의 경우를 동시에 고려한 인과적 지지도와 인과적 신뢰도를 제안한 바 있다. 여기서 X^c와 Y^c의 의미는 각각 X와 Y가 일어나지 않음을 의미한다.

$$supp_{CA}(X \Rightarrow Y) = P(X \text{ and } Y) + P(X^c \text{ and } Y^c) = \frac{a+d}{n}$$

$$conf_{CA}(X \Rightarrow Y) = \frac{1}{2}[P(Y|X) + P(X^c|Y^c)] = \frac{ab + 2ad + bd}{2(a+b)(b+d)}$$

또한 Park (2013)은 이들의 아이디어를 확장하여 다음과 같은 인과적 향상도를 제안한 바 있다.

$$lift_{CA}(X \Rightarrow Y) = \frac{1}{2}\left[\frac{P(Y|X)}{P(Y)} + \frac{P(X^c|Y^c)}{P(X^c)}\right] = \frac{n}{2}\left[\frac{a}{(a+b)(a+c)} + \frac{d}{(b+d)(c+d)}\right]$$

이와 더불어 Kodratoff (2000)는 다음 식과 같이 인과적 확인 신뢰도를 제안한 바 있다.

$$conf_{CC}(X \Rightarrow Y) = \frac{1}{2}[P(Y|X) + P(X^c|Y^c)] - P(Y^c|X) = \frac{1}{2}\left[\frac{a-2b}{a+b} + \frac{d}{b+d}\right] \tag{2.1}$$

본 논문에서는 이를 토대로 하여 인과적 확인 지지도와 인과적 확인 향상도를 고안하여 인과적 확인 연관성 규칙의 평가 기준을 제안하고자 한다.

$$supp_{CC}(X \Rightarrow Y) = \frac{1}{2}[P(X \text{ and } Y) + P(X^c \text{ and } Y^c)] - P(X \text{ and } Y^c) = \frac{a - 2b + d}{2n} \tag{2.2}$$

$$lift_{CC}(X \Rightarrow Y) = \frac{1}{2}\left[\frac{P(Y|X)}{P(Y)} + \frac{P(X^c|Y^c)}{P(X^c)}\right] - \frac{P(Y^c|X)}{P(Y^c)} = \frac{n}{2}\left[\frac{(ad-bc)(a+2c+d)}{(a+b)(a+c)(b+d)(c+d)}\right] \tag{2.3}$$

이러한 인과적 확인 연관성 평가 기준은 $X \Rightarrow Y$과 $Y^c \Rightarrow X^c$에 의한 인과적 관계뿐만 아니라 $X \Rightarrow Y^c$의 경우를 동시에 고려하는 측도라고 생각할 수 있다. 따라서 이러한 평가 기준은 희귀한 사건의 발생에 대한 연관성 규칙에도 적용 가능한 동시에 이를 이용하게 되면 음의 연관성을 가지는 규칙에서 기존의 연관성 규칙 평가 기준이 가질 수 있는 오류를 미연에 방지할 수 있는 것으로 판단된다.

본 논문에서 고려하는 인과적 연관성 평가 기준들이 Piatetsky-Shapiro (1991)가 제안한 흥미도 측도의 조건을 충족하는지의 여부를 조사하면 다음과 같다.

[조건 1] 인과적 확인 측도는 $P(X)$와 $P(Y)$가 고정되어 있을 때, $P(X \text{ and } Y)$의 값에 따라 단조 증가한다.

증명: 먼저 인과적 확인 지지도 $supp_{CC}$는 $P(X)$와 $P(Y)$가 고정되어 있을 때, 식 (2.2)로부터 식 (2.4)와 같이 정리되므로 $P(X \text{ and } Y)$가 증가하면 $supp_{CC}$가 증가하는 것을 쉽게 알 수 있다.

$$supp_{CC}(X \Rightarrow Y) = \frac{1}{2}[1 - 3P(X) - P(Y) + 4P(X \text{ and } Y)]\,. \tag{2.4}$$

다음으로 인과적 확인 신뢰도 $conf_{CC}$의 조건 충족 여부를 증명하기 위해 식 (2.1)을 a에 대해 편미분하면 다음과 같이 양의 값을 가지므로 $conf_{CC}$는 $P(X \text{ and } Y)$의 값에 따라 단조 증가한다.

$$\frac{\partial conf_{CC}(X \Rightarrow Y)}{\partial a} = \frac{3b}{2(a+b)^2} \geq 0.$$

마지막으로 인과적 확인 향상도 $lift_{CC}$의 조건 충족 여부를 증명하기 위해 식 (2.3)을 재정리하면 다음의 식 (2.5)와 같이 나타낼 수 있어서 $P(X \text{ and } Y)$의 값에 따라 3개 항 모두의 값이 증가하므로 $lift_{CC}$는 단조 증가한다.

$$lift_{CC}(X \Rightarrow Y) = \frac{1}{2}\left[\frac{P(X \text{ and } Y)}{P(X)P(Y)} + \frac{1 - P(X) - P(Y) + P(X \text{ and } Y)}{(1-P(X))(1-P(Y))}\right] - \frac{P(X) - P(X \text{ and } Y)}{P(X)(1-P(Y))} \tag{2.5}$$

따라서 [조건 1]에 대한 증명은 이것으로 마무리된다. □

[조건 2] 인과적 확인 측도는 $P(X)$의 값에 따라 단조 감소한다.

증명: 먼저 식 (2.4)로부터 인과적 확인 지지도 $supp_{CC}$는 $P(X)$의 값이 증가함에 따라 단조 감소함을 쉽게 알 수 있다.

인과적 확인 신뢰도 $conf_{CC}$의 조건 충족 여부를 증명하기 위해 먼저 식 (2.1)을 정리하면 다음과 같다.

$$conf_{CC}(X \Rightarrow Y) = \frac{1}{2}\left[\frac{3P(X \text{ and } Y)}{P(X)} - \frac{P(X) - P(X \text{ and } Y)}{1 - P(Y)} - 1\right].$$

이 식으로부터 $P(X)$의 값이 증가하면 첫 번째 항의 분모가 증가하고, 두 번째 항의 값이 감소하므로 $conf_{CC}$의 값은 감소하게 된다.

인과적 향상도 $lift_{CC}$의 조건 충족 여부를 증명하기 위해 식 (2.5)의 항을 관찰해보면 $P(X)$의 값이 증가함에 따라 첫 번째 항의 분모가 증가하게 된다. 그리고 두 번째 항을 $P(X)$에 대해 편미분하면 다음과 같은 식이 얻어지며, 음의 값을 갖는다.

$$\frac{\partial lift_{CC}(X \Rightarrow Y)}{\partial P(X)} = \frac{1}{2}\left[-\frac{P(X \text{ and } Y)}{P(X)^2} - \frac{[P(X) - P(X \text{ and } Y)]}{[(1-P(X))(1-P(Y))]^2}\right] - \frac{P(X \text{ and } Y)}{P(X)^2[1-P(Y)]}$$

따라서 $P(X)$의 값이 증가함에 따라 $lift_{CC}$는 감소하게 되며, [조건 2]에 대한 증명은 이것으로 마무리된다. □

[조건 3] $P(X \text{ and } Y) = P(X)P(Y)$이면 인과적 확인 측도는 0이 된다.

증명: $P(X \text{ and } Y) = P(X)P(Y)$이면 식 (2.3)의 3개의 항 모두 1이 되므로 인과적 확인 향상도 $lift_{CC}$는 0이 되어 이 조건을 만족하게 된다. 반면에 인과적 확인 지지도 $supp_{CC}$와 인과적 확인 신뢰도 $conf_{CC}$의 경우에는 이 조건을 만족하지 않는다. 기존의 지지도 및 신뢰도, 그리고 인과적 측도들과 확인적 측도들의 경우에도 이 조건은 충족되지 않았으나 연관성 평가 기준 중에서 가장 중심이 되는 측도로써 중요한 역할을 담당한 것과 마찬가지로 $conf_{CC}$도 연관성 규칙에서 의미 있는 역할을 수행하게 된다. □

3. 예제를 통한 고찰

본 절에서는 예제를 통하여 기본적인 연관성 규칙 평가 기준, 인과적 연관성 규칙 평가 기준, 확인적 연관성 규칙 평가 기준, 그리고 인과적 확인 연관성 규칙 평가 기준에 대해 수치적인 비교를 하고자 한다. 이를 위해 Park (2013)에서와 유사한 예제를 활용하고자 한다. 이 논문에 기술되어 있는 바와 같이 데이터베이스에 있는 총 트랜잭션의 수 (t)를 100명으로 하고, 항목 집합 X는 발생 (1) 및 비발생 (0)의 수를 각각 50명으로 하였다. 또한 항목 집합 Y는 발생 (1)의 수와 비발생 (0)의 수를 각각 30명과 70명으로 하였다. 항목 집합 X와 Y의 동시 발생 빈도수는 a명으로 하였다. 이를 정리하면 Table 3.1과 같다. 이 표에서 a가 취할 수 있는 정수 값의 범위 각각 $50 \le a \le 70$이다.

Table 3.1 Simulation data(1)

		Y		Total
		1	0	
X	1	a	$80-a$	80
	0	$70-a$	$a-50$	20
Total		70	30	100

Table 3.1로부터 a의 변화에 따른 지지도 기반 연관성 평가 기준들의 계산 결과를 Table 3.2에 제시하였다. 이 표에서 나타난 기호는 다음과 같다.

$$a = n(X=1, Y=1),\ b = n(X=1, Y=0),\ c = n(X=0, Y=1),\ d = n(X=0, Y=0),$$
$$supp_1 = P(X=1, Y=1),\ supp_2 = P(X=0, Y=0),\ supp_3 = P(X=1, Y=0),$$
$$conf_1 = P(Y=1|X=1),\ conf_2 = P(X=0|Y=0),\ conf_3 = P(Y=0|X=1),$$
$$lift_1 = \frac{P(Y=1|X=1)}{P(Y=1)},\ lift_2 = \frac{P(X=0|Y=0)}{P(X=0)},\ lift_3 = \frac{P(Y=0|X=1)}{P(Y=0)}.$$

Table 3.2 Support based thresholds by simulation data(1)

a	b	c	d	$supp_1$	$supp_2$	$supp_3$	$supp_{CA}$	$supp_{CC}$
50	30	20	0	0.500	0.000	0.300	0.500	-0.050
51	29	19	1	0.510	0.010	0.290	0.520	-0.030
52	28	18	2	0.520	0.020	0.280	0.540	-0.010
53	27	17	3	0.530	0.030	0.270	0.560	0.010
54	26	16	4	0.540	0.040	0.260	0.580	0.030
55	25	15	5	0.550	0.050	0.250	0.600	0.050
56	24	14	6	0.560	0.060	0.240	0.620	0.070
57	23	13	7	0.570	0.070	0.230	0.640	0.090
58	22	12	8	0.580	0.080	0.220	0.660	0.110
59	21	11	9	0.590	0.090	0.210	0.680	0.130
60	20	10	10	0.600	0.100	0.200	0.700	0.150
61	19	9	11	0.610	0.110	0.190	0.720	0.170
62	18	8	12	0.620	0.120	0.180	0.740	0.190
63	17	7	13	0.630	0.130	0.170	0.760	0.210
64	16	6	14	0.640	0.140	0.160	0.780	0.230
65	15	5	15	0.650	0.150	0.150	0.800	0.250
66	14	4	16	0.660	0.160	0.140	0.820	0.270
67	13	3	17	0.670	0.170	0.130	0.840	0.290
68	12	2	18	0.680	0.180	0.120	0.860	0.310
69	11	1	19	0.690	0.190	0.110	0.880	0.330
70	10	0	40	0.700	0.400	0.100	1.100	0.450

이 표에서 보는 바와 같이 a와 d가 증가하고 b와 c가 감소함에 따라 $supp_3$를 제외한 모든 지지도 기반 측도들이 증가하는 것으로 나타났다. 이들 중에서 기존의 지지도 $supp_1$과 인과적 지지도 $supp_{CA}$를 비교해보면 $supp_{CA}$가 $supp_1$과 $supp_2$의 합으로 계산되므로 기존의 지지도보다 큰 값을 갖게 된다. 따라서 Park (2013a)가 논의한 바와 같이 특히 기존의 평가 측도인 $supp_1$을 기준으로 연관성 규칙 생성 여부를 판단했을 때 탈락되는 규칙도 인과적 평가 기준인 $supp_{CA}$를 이용하여 판단하게 되면 연관성 규칙의 후보로 생성될 수 있다. 그러나 $supp_1$뿐만 아니라 $supp_{CA}$도 항상 양의 값을 갖게 되므로 양의 연관성이 의미가 있는지 아니면 음의 연관성이 의미가 있는지를 파악하기가 곤란하다. 반면에 본 논문에서 제안하는 인과적 확인 지지도 $supp_{CC}$는 $supp_1$과 $supp_2$뿐만 아니라 음의 연관성의 크기를 나타내는 $supp_3$을 동시에 고려하기 때문에 양의 값 또는 음의 값으로 나타나게 되어 연관성의 방향을 알 수 있게 된다. 이를 좀 더 구체적으로 알아보기 위해 $a = 52$, $b = 28$, $c = 18$, $d = 2$일 때를 살펴보면 $supp_1 = 0.520$, $supp_{CA} = 0.540$, $supp_{CC} = -0.010$으로 나타나서 인과적 신뢰도 $supp_{CA}$는 $supp_1$보다 큰 값으로 나타나므로 $supp_{CA}$를 연관성 평가 기준으로 사용하게 되면 규칙이 더 잘 생성될 수 있으며, 동시에 필요 이상으로 많은 규칙이 생성될 수도 있다. 위의 경우와 $a = 54$, $b = 26$, $c = 16$, $d = 4$인 경우를 비교해보면 $supp_1$과 $supp_{CA}$는 0.520과 0.540, 그리고 0.540과 0.580으로 계산되었으므로 모두 양의 값으로 나타나는 반면에 $supp_{CA}$는 두 경우 각각 -0.010과 0.030으로 얻어져서 양 또는 음의 값으로 나타나는 것을 알 수 있다.

Table 3.1로부터 a의 변화에 따른 신뢰도를 기반으로 한 연관성 평가 기준들의 계산 결과는 Table 3.3에 제시하였다.

Table 3.3 Confidence based thresholds by simulation data(1)

a	b	c	d	$conf_1$	$conf_2$	$conf_3$	$conf_{CA}$	$conf_{CC}$
50	30	20	0	0.625	0.000	0.375	0.313	-0.063
51	29	19	1	0.638	0.033	0.363	0.335	-0.027
52	28	18	2	0.650	0.067	0.350	0.358	0.008
53	27	17	3	0.663	0.100	0.338	0.381	0.044
54	26	16	4	0.675	0.133	0.325	0.404	0.079
55	25	15	5	0.688	0.167	0.313	0.427	0.115
56	24	14	6	0.700	0.200	0.300	0.450	0.150
57	23	13	7	0.713	0.233	0.288	0.473	0.185
58	22	12	8	0.725	0.267	0.275	0.496	0.221
59	21	11	9	0.738	0.300	0.263	0.519	0.256
60	20	10	10	0.750	0.333	0.250	0.542	0.292
61	19	9	11	0.763	0.367	0.238	0.565	0.327
62	18	8	12	0.775	0.400	0.225	0.588	0.363
63	17	7	13	0.788	0.433	0.213	0.610	0.398
64	16	6	14	0.800	0.467	0.200	0.633	0.433
65	15	5	15	0.813	0.500	0.188	0.656	0.469
66	14	4	16	0.825	0.533	0.175	0.679	0.504
67	13	3	17	0.838	0.567	0.163	0.702	0.540
68	12	2	18	0.850	0.600	0.150	0.725	0.575
69	11	1	19	0.863	0.633	0.138	0.748	0.610
70	10	0	40	0.875	0.800	0.125	0.838	0.713

앞의 표에서와 마찬가지로 a와 d가 증가함에 따라 음의 신뢰도를 나타내는 $conf_3$을 제외한 모든 신뢰도 기반 측도들이 증가하는 것으로 나타났다. 이들 중에서 기존의 신뢰도 $conf_1$과 인과적 신뢰도 $conf_{CA}$를 비교해보면 $conf_{CA}$가 $conf_1$과 $conf_2$의 평균에 의해 계산되므로 $conf_2$가 $conf_1$에 비해 상대적으로 작은 경우에는 기존의 신뢰도보다 작은 값을 갖게 된다. 따라서 기존의 평가 측도인 $conf_1$을 기준으로 연관성 규칙 생성 여부를 판단했을 때 탈락되는 규칙도 인과적 신뢰도 $conf_{CA}$를 이용하여 판단하게 되면 연관성 규칙의 후보로 생성될 수 있다. 하지만 기존의 신뢰도 $conf_1$과 인과적 신뢰도 $conf_{CA}$가 항상 양의 값을 가지므로 양의 연관성이 있는지 아니면 음의 연관성이 있는지, 즉 연관성의 방향을 파악하기가 어렵다. 반면에 본 논문에서 제안하는 인과적 확인 신뢰도 $conf_{CC}$는 양의 연관성을 나타내는 $conf_1$과 $conf_2$의 평균값과 음의 연관성 여부를 나타내는 $conf_3$의 값에 의해 부호 및 그 크기가 결정되므로 연관성의 방향을 알 수 있다. 이러한 사실들에 대해 좀 더 구체적으로 알아보기 위해 $a = 51$, $b = 29$, $c = 19$, $d = 1$인 경우와 $a = 54$, $b = 26$, $c = 16$, $d = 4$인 경우를 비교해보면, 전자는 $conf_1 = 0.638$, $conf_{CA} = 0.335$, $conf_{CC} = -0.027$, 그리고 후자는 $conf_1 = 0.675$, $conf_{CA} = 0.404$, $conf_{CC} = 0.079$으로 얻어졌다. 따라서 두 경우 모두 $conf_{CA}$가 $conf_1$보다 작게 나타나고 있는데 그 이유는 $conf_2$가 $conf_1$에 비해 매우 작은 값으로 나타나고 있기 때문이다. 또한 $conf_{CA}$와 $conf_{CC}$의 크기를 비교해보면 각각 0.335와 -0.027, 0.404와 0.079로 나타나서 $conf_{CA}$는 모두 양으로 나타난 반면에 $conf_{CC}$는 양 또는 음의 값으로 나타나서 연관성의 방향을 나타내는 측도라고 할 수 있다.

Table 3.1로부터 동시발생빈도 a의 변화에 따른 향상도 기반 연관성 평가 기준들의 계산 결과는 Table 3.4에 제시하였다.

Table 3.4 Lift based thresholds by simulation data(1)

a	b	c	d	$lift_1$	$lift_2$	$lift_3$	$lift_{CA}$	$lift_{CC}$
50	30	20	0	0.893	0.000	1.250	0.446	-0.804
51	29	19	1	0.911	0.167	1.208	0.539	-0.670
52	28	18	2	0.929	0.333	1.167	0.631	-0.536
53	27	17	3	0.946	0.500	1.125	0.723	-0.402
54	26	16	4	0.964	0.667	1.083	0.815	-0.268
55	25	15	5	0.982	0.833	1.042	0.908	-0.134
56	24	14	6	1.000	1.000	1.000	1.000	0.000
57	23	13	7	1.018	1.167	0.958	1.092	0.134
58	22	12	8	1.036	1.333	0.917	1.185	0.268
59	21	11	9	1.054	1.500	0.875	1.277	0.402
60	20	10	10	1.071	1.667	0.833	1.369	0.536
61	19	9	11	1.089	1.833	0.792	1.461	0.670
62	18	8	12	1.107	2.000	0.750	1.554	0.804
63	17	7	13	1.125	2.167	0.708	1.646	0.938
64	16	6	14	1.143	2.333	0.667	1.738	1.071
65	15	5	15	1.161	2.500	0.625	1.830	1.205
66	14	4	16	1.179	2.667	0.583	1.923	1.339
67	13	3	17	1.196	2.833	0.542	2.015	1.473
68	12	2	18	1.214	3.000	0.500	2.107	1.607
69	11	1	19	1.232	3.167	0.458	2.199	1.741
70	10	0	40	1.250	2.000	0.250	1.625	1.375

이 표에서도 a와 d가 증가하고 b와 c가 감소함에 따라 향상도 기반 측도들 중에서는 $lift_3$을 제외하고는 모두가 증가하는 것으로 나타났다. 기존의 향상도 $lift_1$와 인과적 향상도 $lift_{CA}$를 비교해보면 1의 값을 기준으로 1 보다 큰 경우에는 $lift_{CA}$이 $lift_1$보다 큰 값을 갖게 되고, 1 보다 작은 경우에는 이와 반대의 결과가 나타난다. 양의 연관성 규칙을 생성하는 경우에는 향상도의 값이 1 보다 작으면 의미가 없으므로 1 보다 큰 경우를 의미가 있는 것으로 간주하는데 이 경우에는 $lift_{CA}$가 $lift_1$보다 더 바람직한 측도라고 할 수 있다. 그러나 두 측도 모두 양의 값만을 취하고 있는 반면에 인과적 확인 향상도 $lift_{CC}$는 양 또는 음의 값을 취할 수 있는 동시에 대칭적인 값으로 나타나고 있어서 $lift_{CC}$가 $lift_1$또는 $lift_{CA}$에 비해 좀 더 바람직한 측도라고 할 수 있다. 좀 더 구체적으로 살펴보기 위해 $a = 53$, $b = 27$, $c = 17$, $d = 3$인 경우와 $a = 59$, $b = 21$, $c = 11$, $d = 9$인 경우를 비교해보면, 전자는 $lift_1 = 0.946$, $lift_2 = 0.500$, $lift_3 = 1.125$, $lift_{CA} = 0.723$, $lift_{CC} = -0.402$, 그리고 후자는 $lift_1 = 1.054$, $lift_2 = 1.500$, $lift_3 = 0.875$, $lift_{CA} = 1.277$, $lift_{CC} = 0.402$로 계산되었다. 다시 말해서 $lift_{CA}$와 $lift_1$의 값이 공히 1 보다 큰 경우에는 $lift_{CA}$이 1.277로 1.054인 $lift_1$보다 더 큰 값으로 나타났고, 1 보다 작은 경우에는 이와 반대로 나타났다. 또한 $lift_3$의 값이 $lift_1$과 $lift_2$보다 큰 경우에 $lift_{CC}$는 -0.402로 음의 값으로 계산된 반면에 그 반대인 경우에는 $lift_{CC}$가 0.402로 양의 값으로 계산되었으며, 이 두 값은 0을 중심으로 대칭적인 형태로 나타났다. 따라서 $lift_{CA}$와 $lift_1$중에서는 $lift_{CA}$가 더 바람직하다고 할 수 있으나 $lift_1$와 $lift_{CA}$, 그리고 $lift_{CC}$중에서는 $lift_{CC}$가 가장 바람직한 측도라고 할 수 있다.

이번에는 확인적 연관성 평가 기준들의 유용성을 불일치빈도의 변화 양상과 함께 알아보기 위해 Table 3.5를 이용하여 불일치빈도 b의 변화에 따른 기존의 평가 기준과 인과적 확인 측도들의 계산하였으며, 지지도 기반 연관성 평가 기준들의 계산 결과를 Table 3.6에 제시하였다. 이 표에서 b가 취할 수 있는 정수 값의 범위는 a와 마찬가지로 $50 \leq b \leq 70$이다.

Table 3.5 Simulation data(2)

		Y		Total
		1	0	
X	1	$80-b$	b	80
	0	$b-50$	$70-b$	20
Total		30	70	100

이 표에서 보는 바와 같이 b와 c가 증가하고 a와 d가 감소함에 따라 $supp_3$는 증가하는 반면에 이를 제외한 모든 지지도 기반 측도들은 감소하는 것으로 나타났다. 이들 중에서 기존의 지지도 $supp_1$과 인과적 지지도 $supp_{CA}$를 비교해보면 앞에서와 마찬가지로 $supp_{CA}$가 $supp_1$보다 큰 값을 갖게 된다. 그런데 $supp_1$와 $supp_{CA}$는 항상 양의 값을 갖게 되는 반면에 $supp_{CC}$는 $supp_1$과 $supp_2$뿐만 아니라 음의 연관성의 크기를 나타내는 $supp_3$을 동시에 고려하기 때문에 양의 값 또는 음의 값으로 나타나게 된다. 이를 좀 더 구체적으로 알아보기 위해 $a=25$, $b=55$, $c=5$, $d=15$인 경우와 $a=20$, $b=60$, $c=0$, $d=10$인 경우를 살펴보면 전자에서는 $supp_1=0.250$, $supp_{CA}=0.400$, $supp_{CC}=-0.350$으로 나타났고, 후자인 경우에는 $supp_1=0.200$, $supp_{CA}=0.300$, $supp_{CC}=-0.450$으로 나타나서 두 경우 공히 인과적 지지도 $supp_{CA}$가 기존의 지지도 $supp_1$보다 큰 값으로 나타나므로 $supp_{CA}$가 더 바람직하다고 할 수 있다. 그러나 두 경우 모두 $supp_1$과 $supp_{CA}$가 양의 값만으로 나타나므로 연관성의 방향을 나타내지 못한다고 할 수 있으며, 이 예제에서는 $supp_{CC}$가 모두 음의 값을 취하기는 하나, 일반적으로는 양의 연관성의 크기뿐만 아니라 음의 연관성의 크기를 동시에 고려하기 때문에 양의 값 또는 음의 값으로 나타나게 된다.

Table 3.6 Support based thresholds by simulation data(2)

a	b	c	d	$supp_1$	$supp_2$	$supp_3$	$supp_{CA}$	$supp_{CC}$
30	50	0	20	0.300	0.200	0.500	0.500	-0.250
29	51	1	19	0.290	0.190	0.510	0.480	-0.270
28	52	2	18	0.280	0.180	0.520	0.460	-0.290
27	53	3	17	0.270	0.170	0.530	0.440	-0.310
26	54	4	16	0.260	0.160	0.540	0.420	-0.330
25	55	5	15	0.250	0.150	0.550	0.400	-0.350
24	56	6	14	0.240	0.140	0.560	0.380	-0.370
23	57	7	13	0.230	0.130	0.570	0.360	-0.390
22	58	8	12	0.220	0.120	0.580	0.340	-0.410
21	59	9	11	0.210	0.110	0.590	0.320	-0.430
20	60	10	10	0.200	0.100	0.600	0.300	-0.450
19	61	11	9	0.190	0.090	0.610	0.280	-0.470
18	62	12	8	0.180	0.080	0.620	0.260	-0.490
17	62	13	7	0.170	0.070	0.620	0.240	-0.500
16	64	14	6	0.160	0.060	0.640	0.220	-0.530
15	65	15	5	0.150	0.050	0.650	0.200	-0.550
14	66	16	4	0.140	0.040	0.660	0.180	-0.570
13	67	17	3	0.130	0.030	0.670	0.160	-0.590
12	68	18	2	0.120	0.020	0.680	0.140	-0.610
11	69	19	1	0.110	0.010	0.690	0.120	-0.630
10	70	20	0	0.100	0.000	0.700	0.100	-0.650

Table 3.5로부터 b의 변화에 따른 신뢰도 기반 연관성 평가 기준들의 계산 결과는 Table 3.7에 제시하였다. 이 표에서 보는 바와 같이 b와 c가 증가하고 a와 d가 감소함에 따라 $conf_3$는 증가하는 반면에 이를 제외한 모든 신뢰도 기반 측도들은 감소하는 것으로 나타났다. 이들 중에서 기존의 신뢰도 $conf_1$과 인과적 신뢰도 $conf_{CA}$를 비교해보면 $conf_{CA}$가 $conf_1$과 $conf_2$의 평균에 의해 계산되므로 $conf_2$가 $conf_1$에 비해 상대적으로 작은 경우에는 기존의 지지도보다 작은 값을 갖게 된다. 앞의 예제에서와 마찬가지로 기존의 평가 측도인 $conf_1$을 기준으로 연관성 규칙 생성 여부를 판단했을 때 탈락되는 규칙도 인과적 신뢰도 $conf_{CA}$를 이용하여 판단하게 되면 연관성 규칙의 후보로 생성될 수 있다. 또한 $conf_1$과 $conf_{CA}$는 항상 양의 값을 취하는 반면에 $conf_{CC}$는 여기서는 비록 모두 음의 값을 취하고는 있으나 양의 신뢰도를 측정하는 $conf_1$과 $conf_2$, 그리고 음의 신뢰도를 나타내는 측도 $conf_3$를 모두 고려하기 때문에 양 또는 음의 값을 취하는 측도라고 할 수 있다. Table 3.5로부터 불일치빈도 b의 변화에 따른 향상도 기반 연관성 평가 기준들을 계산해보았는데 이 경우에도 앞의 예제에서와 유사한 결과를 얻을 수 있었다.

Table 3.7 Confidence based thresholds by simulation data(2)

a	b	c	d	$conf_1$	$conf_2$	$conf_3$	$conf_{CA}$	$conf_{CC}$
30	50	0	20	0.375	0.286	0.625	0.330	-0.295
29	51	1	19	0.363	0.271	0.638	0.317	-0.321
28	52	2	18	0.350	0.257	0.650	0.304	-0.346
27	53	3	17	0.338	0.243	0.663	0.290	-0.372
26	54	4	16	0.325	0.229	0.675	0.277	-0.398
25	55	5	15	0.313	0.214	0.688	0.263	-0.424
24	56	6	14	0.300	0.200	0.700	0.250	-0.450
23	57	7	13	0.288	0.186	0.713	0.237	-0.476
22	58	8	12	0.275	0.171	0.725	0.223	-0.502
21	59	9	11	0.263	0.157	0.738	0.210	-0.528
20	60	10	10	0.250	0.143	0.750	0.196	-0.554
19	61	11	9	0.238	0.129	0.763	0.183	-0.579
18	62	12	8	0.225	0.114	0.775	0.170	-0.605
17	62	13	7	0.215	0.101	0.785	0.158	-0.626
16	64	14	6	0.200	0.086	0.800	0.143	-0.657
15	65	15	5	0.188	0.071	0.813	0.129	-0.683
14	66	16	4	0.175	0.057	0.825	0.116	-0.709
13	67	17	3	0.163	0.043	0.838	0.103	-0.735
12	68	18	2	0.150	0.029	0.850	0.089	-0.761
11	69	19	1	0.138	0.014	0.863	0.076	-0.787
10	70	20	0	0.125	0.000	0.875	0.063	-0.813

4. 결론

데이터 마이닝 기법들 중에서 연관성 규칙을 생성하고자 하는 경우 기본적인 연관성 평가 기준만으로는 연관성의 방향을 파악할 수 없는 동시에 전항과 후항의 인과관계를 설명할 수 없다. 이러한 문제를 해결하기 위해 본 논문에서는 인과적 확인 연관성 평가 기준인 인과적 확인 지지도, 인과적 확인 신뢰도, 그리고 인과적 확인 향상도를 제안하였다. 또한 이들이 Piatetsky-Shapiro (1991)가 제안한 흥미도 측

도의 조건 충족 여부를 조사하였다. 그 결과, 세 가지 인과적 확인 측도 모두 동시 발생 확률의 값에 따라 단조 증가하는 조건과 각 항목의 주변 확률의 값에 따라 단조 감소하는 조건은 만족하였다. 반면에 두 항목이 독립이면 연관성 평가 기준의 값이 0이 되는 조건에 대해서는 인과적 지지도와 신뢰도는 기존의 지지도와 신뢰도와 같이 이 조건을 충족하지 않는다. 그러나 기존의 연관성 규칙에서의 지지도와 신뢰도와 마찬가지로 중심이 되는 측도이므로 중요한 역할을 담당하는 측도로 볼 수 있다. 또한 예제를 통해 기본적인 연관성 평가 기준과 인과적 연관성 평가 기준, 그리고 인과적 확인 연관성 평가 기준을 비교해 본 결과, 본 논문에서 제안하는 인과적 확인 측도들이 다른 평가 기준에 비해 가장 바람직한 측도라는 사실을 파악하였다. 특히 동시 발생 빈도와 동시 비 발생빈도가 증가하고 불일치빈도가 감소하는 경우에는 음의 신뢰도를 제외한 모든 신뢰도 기반 측도들이 증가하는 것으로 나타났다. 이들 중에서 기존의 신뢰도와 인과적 신뢰도를 비교해보면 인과적 신뢰도가 양의 신뢰도와 역의 신뢰도의 평균에 의해 계산되므로 역의 신뢰도가 양의 신뢰도에 비해 상대적으로 작은 경우에는 기존의 신뢰도보다 작은 값을 갖게 된다. 따라서 기존의 평가 측도인 신뢰도를 기준으로 연관성 규칙 생성 여부를 판단했을 때 탈락되는 규칙도 인과적 신뢰도를 이용하여 판단하게 되면 연관성 규칙의 후보로 생성될 수 있다. 하지만 기존의 신뢰도와 인과적 신뢰도가 항상 양의 값을 가지므로 연관성의 방향을 파악하기가 어렵다. 반면에 본 논문에서 제안하는 인과적 확인 신뢰도는 양의 신뢰도와 역의 신뢰도의 평균값과 음의 신뢰도의 값에 의해 부호 및 그 크기가 결정되므로 연관성의 방향을 알 수 있다.

위의 결과들을 종합해볼 때, 본 논문에서 제안하는 인과적 확인 연관성 평가 기준은 음의 연관성을 가지는 규칙에서 기존의 연관성 규칙 평가 기준이 가질 수 있는 오류를 미연에 방지할 수 있을 뿐만 아니라 희귀한 사건의 발생에 대한 연관성 규칙에도 적용 가능할 것으로 판단된다.

References

Agrawal, R., Imielinski, R. and Swami, A. (1993). Mining association rules between sets of items in large databases. *Proceedings of the ACM SIGMOD Conference on Management of Data*, 207-216.

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. *Proceedings of the 20th VLDB Conference*, 487-499.

Berzal, F., Cubero, J., Marin, N., Sanchez, D., Serrano, J. and Vila, A. (2005). Association rule evaluation for classification purposes. *Actas del III Taller Nacional de Minería de Datosy Aprendizaje, TAMIDA2005*, 135-144.

Cai, C. H., Fu, A. W. C., Cheng, C. H. and Kwong, W. W. (1998). Mining association rules with weighted items. *Proceedings of International Database Engineering and Applications Symposium*, 68-77.

Cho, K. H. and Park, H. C. (2011a). Study on the multi intervening relation in association rules. *Journal of the Korean Data Analysis Society*, **13**, 297-306.

Cho, K. H. and Park, H. C. (2011b). A study on insignificant rules discovery in association rule mining. *Journal of the Korean Data & Information Science Society*, **22**, 81-88.

Han, J. and Fu, Y. (1999). Mining multiple-level association rules in large databases. *IEEE Transactions on Knowledge and Data Engineering*, **11**, 68-77.

Han, J., Pei, J. and Yin, Y. (2000). Mining frequent patterns without candidate generation. *Proceedings of ACM SIGMOD Conference on Management of Data*, 1-12.

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.

Kim, B., Park, Y. and Jang, N. (2013). Study for independence of hits in professional baseball games. *Journal of the Korean Data & Information Science Society*, **24**, 1421-1428.

Kim, N. (2008). Effect of market basket size on the accuracy of association rule measures. *Asia Pacific Journal of Information Systems*, **18**, 95-114.

Kodratoff, Y. (2000). Comparing machine learning and knowledge discovery in databases: An application to knowledge discovery in texts. *Proceeding of Machine Learning and its Applications: Advanced Lectures*, 1-21.

Lim, J., Lee, K. and Cho, Y. (2010). A study of association rule by considering the frequency. *Journal of the Korean Data & Information Science Society*, **21**, 1061-1069.

Liu, B., Hsu, W. and Ma, Y. (1999). Mining association rules with multiple minimum supports. *Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining*, 337-241.

Park, H. C. (2012a). Negatively attributable and pure confidence for generation of negative association rules. *Journal of the Korean Data & Information Science Society*, **23**, 707-716.

Park, H. C. (2012b). Exploration of PIM based similarity measures as association rule thresholds. *Journal of the Korean Data & Information Science Society*, **23**, 1127-1135.

Park, H. C. (2013). Proposition of causal association rule thresholds. *Journal of the Korean Data & Information Science Society*, **24**, 1189-1197.

Pasquier, N., Bastide, Y., Taouil, R. and Lakhal, L. (1999). Discovering frequent closed itemsets for association rules. *Proceedings of the 7th International Conference on Database Theory*, 398-416.

Pei, J., Han, J. and Mao, R. (2000). CLOSET: An efficient algorithm for mining frequent closed itemsets. *Proceedings of ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, 21-30.

Piatetsky-Shapiro, G. (1991). Discovery, analysis and presentation of strong rules. *Knowledge Discovery in Databases*, AAAI/MIT Press, 229-248.

Saygin, Y., Vassilios, S. V. and Clifton, C. (2002). Using unknowns to prevent discovery of association rules. *Proceedings of 2002 Conference on Research Issues in Data Engineering*, 45-54.

Srinkant, R., Vu, Q. and Agrawal, R. (1997). Mining association rules with item constraints. *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, 67-73.

Journal of the Korean Data & Information Science Society
2014, **25**(4), 857–868

http://dx.doi.org/10.7465/jkdi.2014.25.4.857
한국데이터정보과학회지

Proposition of causally confirmed measures in association rule mining

Hee Chang Park[1]

[1]Department of Statistics, Changwon National University

Received 16 June 2014, revised 10 July 2014, accepted 15 July 2014

Abstract

Data mining is the representative analysis methodology in the era of big data, and is the process to analyze a massive volume database and summarize it into meaningful information. Association rule technique finds the relationship among several items in huge database using the interestingness measures such as support, confidence, lift, etc. But these interestingness measures cannot be used to establish a causality relationship between antecedent and consequent item sets. Moreover, we can not know association direction by them. This paper propose causally confirmed association thresholds to compensate for these problems, and then check the three conditions of interestingness measures. The comparative studies with basic association thresholds, causal association thresholds, and causally confirmed association thresholds are shown by simulation studies. The results show that causally confirmed association thresholds are better than basic and causal association thresholds.

Keywords: Association rule, causally confirmed confidence, causally confirmed lift, causally confirmed support, data mining.

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