Journal of the Korean Data & Information Science Society
2014, **25**(3), 601–609

http://dx.doi.org/10.7465/jkdi.2014.25.3.601
한국데이터정보과학회지

연관성 규칙에서 활용 가능한 대칭적 기여 순수 신뢰도의 개발

박희창[1]

[1]창원대학교 통계학과

접수 2014년 4월 2일, 수정 2014년 4월 21일, 게재확정 2014년 5월 19일

요 약

빅 데이터 분석을 위한 데이터마이닝 기법 중의 하나인 연관성 규칙은 지지도, 신뢰도, 향상도 등의 여러 가지 연관성 평가기준을 기반으로 하여 항목집합들 간의 관련성을 찾아내는 데 활용되고 있다. 기본적인 연관성 평가기준들 중에서 가장 많이 활용되고 있는 신뢰도는 연관성의 방향 (음 또는 양)을 알 수가 없다는 단점을 가지고 있다. 이를 보완하기 위한 측도로 순수 신뢰도 기여 순수 신뢰도가 제안되었으나, 이는 전항과 후항이 바뀌면 그 값이 달라지는 문제점이 있다. 본 논문에서는 기존의 신뢰도와 순수 신뢰도, 그리고 기여 순수 신뢰도의 단점을 보완한 연관성 평가 기준으로 네 가지의 대칭적 기여 순수 신뢰도를 제안하였다. 또한 신뢰도와 기여 순수 신뢰도, 그리고 네 가지의 대칭적 기여 순수 신뢰도를 예제를 통하여 비교 분석하였다. 그 결과, 대칭적 기여 순수 신뢰도는 그 부호에 의해 연관성 규칙의 방향을 파악할 수 있는 동시에 전항과 후항이 바뀌어도 그 값이 변하지 않으므로 연관성 규칙을 생성하는 데 매우 유익한 평가 기준이라는 사실을 확인할 수 있었다. 이들 네 가지 대칭적 기여 순수 신뢰도 중에서는 두 종류의 기여 순수 신뢰도의 분자의 합과 분모의 합의 비로 나타나는 측도가 가장 바람직한 것으로 예제를 통하여 확인하였다.

주요용어: 기여 순수 신뢰도, 대칭적 기여 순수 신뢰도, 신뢰도, 연관성 평가 기준.

1. 서론

빅 데이터가 사회전반에서 새로운 관심영역으로 부상하고 있는 요즘, 엄청난 규모의 데이터 속에서 유용한 정보를 찾아내 주는 데이터마이닝 기술이 주목받고 있다 (Park, 2013a). 데이터마이닝은 특정 집단의 의미 있는 정보를 발견하기 위해 통계학 및 수학적 기술과 패턴인식 기술 등의 도구를 이용하여 방대한 양의 데이터를 탐색하고 분석하는 과정을 의미한다 (Park, 2013b). 데이터마이닝 기법 중의 하나인 연관성 규칙 (association rule)은 대용량 데이터베이스로부터 항목집합들 간에 특정한 연관성을 지지도 (support), 신뢰도 (confidence), 그리고 향상도 (lift) 등의 흥미도 측도 (interestingness measure)를 이용하여 발견하는 것으로 기업의 의사결정 문제, 유통업에서의 고객관리나 교차판매, 보험 및 의료, 생물 정보학 (Bioinformatics) 등 다양한 분야에서 활용되고 있으며, 특히 비정형 데이터로 관측되는 빅 데이터 분석에서 많이 활용되고 있다 (Park, 2011b). 이러한 연관성 규칙은 Agrawal 등 (1993)에 의해 처음 소개되었으며, 2000년 이후로 많은 학자들에 의해 연구되고 있다 (Han 등, 2000; Pei 등, 2000; Cho와 Park, 2011a; Cho와 Park, 2011b; Jin 등, 2011; Park, 2012a; Park, 2012b; Park, 2013a, Park, 2013b).

의미 있는 연관성 규칙을 탐색하기 위한 흥미도 측도는 Silberschatz와 Tuzhilin (1996)과 Freitas (1999)에 의해 크게 객관적 흥미도 측도와 주관적 흥미도 측도로 분류된 바 있다. 주관적 흥미도 측도는

[1] (641-773) 경상남도 창원시 의창구 창원대학로 20, 창원대학교 통계학과, 교수.
E-mail: hcpark@changwon.ac.kr

사용자 관점에서 해석 가능하도록 제안된 반면에, 객관적 흥미도 측도는 논리적인 또는 통계적인 방법에 의해 제안된 것으로 사용자에게 규칙을 정제할 수 있는 근거를 제시해준다 (Park, 2012a). 이 중에서 객관적 흥미도 측도가 더 많이 활용되고 있으며, 객관적 흥미도 측도들 중에서 사용자가 지정한 최소 지지도를 만족하는 빈발항목집합을 생성한 후 최저 신뢰도를 초과하는 규칙을 연관성 규칙으로 생성한다 (Park, 2011a). 이 때 규칙 여부를 결정하기 위해 가장 많이 사용되는 기준인 신뢰도 값의 크기로는 양의 연관성이 있는지, 아니면 음의 연관성이 있는지를 알 수 없으므로 Park (2011b)은 기존에 많이 활용되고 있는 흥미도 측도인 신뢰도와 순수 신뢰도의 단점을 보완한 기여 순수 신뢰도 (attributably pure confidence)를 제안하였다. 이 측도는 신뢰도와 순수 신뢰도의 크기를 동시에 고려한 것으로 양의 신뢰도와 음의 신뢰도의 크기를 상대적으로 비교해서 나타낸 흥미도 측도라고 할 수 있다. 그러나 기여 순수 신뢰도는 전항과 후항이 바꾸면 그 값이 달라지는 신뢰도의 단점을 그대로 가지고 있어서 바람직한 축도로 보기는 어렵다. 전항과 후항을 달리 했을 경우에 측도의 값이 달라지면 연구자들은 둘 중 어느 값을 기준으로 연관성 규칙을 생성해야 하는지에 대한 애로사항이 발생하게 된다. 본 논문에서는 이를 보완한 대칭적 기여 순수 신뢰도 (symmetrically and attributably pure confidence)를 제안하고자 한다. 본 논문의 2절에서는 Piatetsky-Shapiro (1991)가 제안한 흥미도 측도의 기준에 대한 충족여부를 파악한다. 3절에서는 예제를 통하여 기존의 신뢰도와 순수 신뢰도, 그리고 기여 순수 신뢰도와의 비교를 통해 대칭적 기여 순수 신뢰도의 유용성을 탐색한 후, 4절에서 결론을 내리고자 한다.

2. 대칭적 기여 순수 신뢰도

본 절에서는 하나의 트랜잭션에서 항목 X와 Y의 연관성의 정도를 대칭적 기여 순수 신뢰도를 이용하여 측정한 후 Park (2011b)의 연구 결과와 비교하기 위해 Table 2.1과 같은 2×2분할표를 활용하고자 한다. 먼저 연관성 규칙을 평가하기 위한 지지도 $supp(X \Rightarrow Y)$는 항목 집합 X와 항목 집합 Y가 동시에 발생하는 거래의 비율을 의미하며, a/n으로 정의된다. 신뢰도 $conf_1(X \Rightarrow Y)$는 항목 집합 X가 포함된 거래 비율 중 항목 집합 X와 항목 집합 Y가 동시에 포함된 거래의 비율을 의미하고 $a/(a+b)$로 정의되며, 전항과 후항을 바꾼 신뢰도 $conf_2(Y \Rightarrow Y)$는 $a/(a+c)$로 정의된다. 향상도 $lift(X \Rightarrow Y)$는 항목 집합 X를 구매한 경우 그 거래가 항목 집합 Y를 포함하는 경우와 항목 집합 Y가 임의로 구매되는 경우의 비를 의미하며, $na/[(a+b)(a+c)]$와 같이 정의된다.

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

한편, Ahn과 Kim (2003)은 $P(Y|X)$의 값이 최저 신뢰도의 기준을 만족하더라도 $P(Y)$의 값보다 작으면 음의 연관성을 가지는 되는데, 음의 연관성을 가진 연관성 규칙을 강한 양의 관계를 지닌 규칙으로 선택하게 하는 것이 신뢰도의 오류라고 지적한 바 있다. 그들은 이러한 신뢰도의 오류를 해결하기 위해 순수 신뢰도 (net confidence; $Nconf$)를 제안하였다. 그러나 이 측도는 양의 연관성을 의미하는 $P(Y|X)$와 음의 연관성을 나타내는 $P(Y|\bar{X})$의 값이 어떤 값을 가지더라도 두 값의 차이가 동일하면 순수 신뢰도의 값도 동일하게 되는 단점이 Park (2011b)에 의해 확인되었다. 이러한 문제를 보완하기 위

해 Park (2011b)는 다음과 같은 기여 순수 신뢰도 ($APconf$)를 제안하였다.

$$APconf_1(X \Rightarrow Y) = \frac{P(Y|X) - P(Y|\bar{X})}{P(Y|X)}$$

$$APconf_2(Y \Rightarrow X) = \frac{P(X|Y) - P(X|\bar{Y})}{P(X|Y)}$$

이 측도는 위의 식에서 보는 바와 같이 전항과 후항의 위치를 바꾸게 되면 그 값이 달라진다는 단점을 여전히 가지고 있다. 다시 말해서 이들 두 측도의 값의 차이가 크거나 이들 둘 중에서 어느 하나가 연관성 규칙 생성을 위한 최저 기준을 만족하지 않는 경우에 어느 것을 이용하여 연관성 규칙 여부를 결정할 지에 대해 판단하기가 어렵다. 이러한 문제점을 보완하기 위해 본 논문에서는 대칭적 기여 순수 신뢰도를 제안하고자 한다. 이를 위해 먼저 $APconf_1$과 $APconf_2$를 정리하면 다음과 같이 나타낼 수 있다.

$$APconf_1(X \Rightarrow Y) = \frac{P(XY) - P(X)P(Y)}{P(XY)[1 - P(X)]} \tag{2.1}$$

$$APconf_2(Y \Rightarrow X) = \frac{P(XY) - P(X)P(Y)}{P(XY)[1 - P(Y)]} \tag{2.2}$$

식 (2.1)과 (2.2)의 분모를 구성하고 있는 $1 - P(X)$와 $1 - P(Y)$을 토대로 여러 가지 대칭적인 측도를 나타낼 수 있는데, 본 논문에서는 다음과 같은 네 가지 측도를 제안하여 비교하고자 한다.

$$SAPconf_1(X \Leftrightarrow Y) = \frac{P(XY) - P(X)P(Y)}{P(XY)[1 - P(X)P(Y)]} \tag{2.3}$$

$$SAPconf_2(X \Leftrightarrow Y) = \frac{P(XY) - P(X)P(Y)}{P(XY)[1 - P(X)][1 - P(Y)]} \tag{2.4}$$

$$SAPconf_3(X \Leftrightarrow Y) = \frac{P(XY) - P(X)P(Y)}{P(XY)[1 - P(XY)]} \tag{2.5}$$

$$NSAPconf(X \Leftrightarrow Y) = \frac{P(XY) - P(X)P(Y)}{P(XY)[1 - (P(X) + P(Y))/2]} \tag{2.6}$$

위에서 보는 바와 같이 식 (2.3)의 $SAPconf_1$는 두 기여 순수 신뢰도의 분모에 있는 $P(X)$와 $P(Y)$를 이들의 곱으로 대체한 것이다. 식 (2.4)의 $SAPconf_2$는 두 식의 분모를 구성하고 있는 $1 - P(X)$및 $1 - P(Y)$대신 이들의 곱을 대입한 것이며, 식 (2.5)의 $SAPconf_3$는 $P(X)$와 $P(Y)$대신 $P(XY)$을 고려한 것이다. 마지막으로 식 (2.t)의 $NSAPconf$는 $P(X)$와 $P(Y)$을 이들의 평균으로 대체한 것으로, 이 측도는 다음의 식과 같이 나타낼 수 있어서 식 (2.1)과 (2.2)의 분자의 합과 분모의 합의 비로 나타나는 측도가 된다.

$$NSAPconf(X \Leftrightarrow Y) = \frac{2[P(XY) - P(X)P(Y)]}{P(XY)[1 - P(X)] + P(XY)[1 - P(Y)]}$$

이들 네 가지 대칭적 기여 순수 신뢰도에 대해 Piatetsky-Shapiro(1991)가 제안한 흥미도 측도의 세 가지 조건을 충족하는지의 여부를 조사하기 위해 먼저 식 (2.3)에서부터 식 (2.6)까지의 분자로부터 $P(XY) = P(X)P(Y)$이면 이들 모두는 0이 되므로 첫 번째 조건을 만족하고 있다. 또한 $SAPconf_3$는 $P(X)$의 값이 증가함에 따라 단조 감소한다는 것을 식 (2.5)로부터 바로 알 수 있으며, 나머지 측도들은 $P(X)$에 관해 편미분한 후 정리하면 음의 값을 취하므로 $P(X)$가 증가함에 따라 단조 감소한다는 것을

알 수 있다. 마지막으로 네 가지 측도들을 $P(XY)$에 관해 편미분한 후 정리하면 양의 값으로 나타나고 있으므로 $P(XY)$의 값이 증가함에 따라 단조 증가한다고 할 수 있다.

한편, 본 논문에서 제안하는 측도들은 Hilderman과 Hamilton (1999), Tan 등 (2002), Omiecinski (2003), Geng과 Hamilton (2006), 그리고 Wu 등 (2010)이 고려한 흥미도 측도들 중에서 확실성 인자 (certainty factor), Piatetsky-Shapiro 계수, 그리고 파이 계수 (phi coefficient)와 유사한 형태를 취한다고 볼 수 있다.

$$\text{확실성 인자} : \frac{P(Y|X) - P(Y)}{1 - P(Y)}$$

$$\text{Piatetsky-Shapiro 계수} : P(XY) - P(X)P(Y)$$

$$\text{파이 계수} : \frac{P(XY) - P(X)P(Y)}{\sqrt{P(X)P(Y)[1 - P(X)][1 - P(Y)]}}$$

먼저 Piatetsky-Shapiro 계수는 대칭적 기여 순수 신뢰도들의 분자만을 고려하고 있으며, 확실성 인자와 파이 계수의 분자는 이들 측도와 동일하나 분모에서는 각 항목의 주변 발생 비율을 고려하고 있다. 반면에 대칭적 기여 순수 신뢰도들의 분모는 두 항목의 주변 발생 비율뿐만 아니라 두 항목의 동시 발생 비율도 함께 고려하고 있으므로 값의 변화하는 양상이 이들 측도와는 다르다고 할 수 있다.

3. 적용 예제

이 절에서는 신뢰도와 기여 순수 신뢰도, 그리고 네 가지의 대칭적 기여 순수 신뢰도를 예제를 통하여 비교 분석함으로써 유용성을 탐색한 후 가장 바람직한 측도를 선정하기 위해 Park (2011b)에서와 같이 두 항목 X, Y에 대해 다음과 같이 가정하였다. 먼저 동시발생빈도의 변화에 따라 여러 가지 측도들이 변하는 양상을 살펴보기 위해 데이터베이스에 있는 총 트랜잭션의 수 (t)를 100명으로 하고, 항목 X는 구매한 물품의 금액을 기준으로 특정금액 이상 구매 (1)한 사람 수와 특정금액 미만을 구매 (0)한 사람 수를 각각 50명으로 하였다. 또한 항목 Y를 결제 방식을 기준으로 특정방법으로 결제 (1)한 사람 수를 30명으로 하고 그 외의 방법으로 결제 (0)한 사람의 수를 70명으로 하였다. X와 Y의 동시발생빈도인 특정금액 이상의 물품을 구매하면서 특정방법으로 결제한 빈도수는 a명으로 하였다. 이를 정리하면 Table 3.1과 같다. 이 표에서 동시발생빈도 a가 취할 수 있는 정수 값의 범위는 $0 \le a \le 30$이다.

Table 3.1 Simulation data(1)

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

Table 3.1로부터 a의 변화에 따른 지지도, 신뢰도, 기여 순수 신뢰도, 그리고 네 가지 대칭적 기여 순수 신뢰도를 계산한 결과를 Table 3.2에 제시하였다. 이 표에 나타나는 기호는 다음과 같다.

$$b = 50 - a,\ c = 30 - a,\ d = a + 20,$$
$$AP_1 = APconf_1,\ AP_2 = APconf_2,$$
$$SAP_1 = SAPconf_1,\ SAP_2 = SAPconf_2,$$
$$SAP_3 = SAPconf_3,\ NSAP = NSAPconf$$

Table 3.2 Variation of several interestingness measures by simulation data(1)

a	b	c	d	$supp$	$conf_1$	$conf_2$	AP_1	AP_2	SAP_1	SAP_2	SAP_3	$NSAP$
1	49	29	21	0.010	0.020	0.033	-28.000	-20.000	-16.471	-40.000	-14.141	-23.333
2	48	28	22	0.020	0.040	0.067	-13.000	-9.286	-7.647	-18.571	-6.633	-10.833
3	47	27	23	0.030	0.060	0.100	-8.000	-5.714	-4.706	-11.429	-4.124	-6.667
4	46	26	24	0.040	0.080	0.133	-5.500	-3.929	-3.235	-7.857	-2.865	-4.583
5	45	25	25	0.050	0.100	0.167	-4.000	-2.857	-2.353	-5.714	-2.105	-3.333
6	44	24	26	0.060	0.120	0.200	-3.000	-2.143	-1.765	-4.286	-1.596	-2.500
7	43	23	27	0.070	0.140	0.233	-2.286	-1.633	-1.345	-3.265	-1.229	-1.905
8	42	22	28	0.080	0.160	0.267	-1.750	-1.250	-1.029	-2.500	-0.951	-1.458
9	41	21	29	0.090	0.180	0.300	-1.333	-0.952	-0.784	-1.905	-0.733	-1.111
10	40	20	30	0.100	0.200	0.333	-1.000	-0.714	-0.588	-1.429	-0.556	-0.833
11	39	19	31	0.110	0.220	0.367	-0.727	-0.519	-0.428	-1.039	-0.409	-0.606
12	38	18	32	0.120	0.240	0.400	-0.500	-0.357	-0.294	-0.714	-0.284	-0.417
13	37	17	33	0.130	0.260	0.433	-0.308	-0.220	-0.181	-0.440	-0.177	-0.256
14	36	16	34	0.140	0.280	0.467	-0.143	-0.102	-0.084	-0.204	-0.083	-0.119
15	35	15	35	0.150	0.300	0.500	0.000	0.000	0.000	0.000	0.000	0.000
16	34	14	36	0.160	0.320	0.533	0.125	0.089	0.074	0.179	0.074	0.104
17	33	13	37	0.170	0.340	0.567	0.235	0.168	0.138	0.336	0.142	0.196
18	32	12	38	0.180	0.360	0.600	0.333	0.238	0.196	0.476	0.203	0.278
19	31	11	39	0.190	0.380	0.633	0.421	0.301	0.248	0.602	0.260	0.351
20	30	10	40	0.200	0.400	0.667	0.500	0.357	0.294	0.714	0.313	0.417
21	29	9	41	0.210	0.420	0.700	0.571	0.408	0.336	0.816	0.362	0.476
22	28	8	42	0.220	0.440	0.733	0.636	0.455	0.374	0.909	0.408	0.530
23	27	7	43	0.230	0.460	0.767	0.696	0.497	0.409	0.994	0.452	0.580
24	26	6	44	0.240	0.480	0.800	0.750	0.536	0.441	1.071	0.493	0.625
25	25	5	45	0.250	0.500	0.833	0.800	0.571	0.471	1.143	0.533	0.667
26	24	4	46	0.260	0.520	0.867	0.846	0.604	0.498	1.209	0.572	0.705
27	23	3	47	0.270	0.540	0.900	0.889	0.635	0.523	1.270	0.609	0.741
28	22	2	48	0.280	0.560	0.933	0.929	0.663	0.546	1.327	0.645	0.774
29	21	1	49	0.290	0.580	0.967	0.966	0.690	0.568	1.379	0.680	0.805

Table 3.2에서 $P(X) = 0.500$이고, $P(Y) = 0.300$이다. 이 표에서 보는 바와 같이 동시발생빈도 a가 증가함에 따라 본 논문에서 고려하고 있는 네 가지의 대칭적 기여 순수 신뢰도를 포함한 모든 평가 기준의 값들이 증가하고 있으며, 지지도와 신뢰도를 제외한 모든 측도들이 양의 값과 음의 값을 동시에 나타내고 있다. 따라서 기여 순수 신뢰도와 대칭적 기여 순수 신뢰도는 모두 연관성의 방향을 나타내는 측도라고 할 수 있어서 기본적인 신뢰도에 비해 바람직한 연관성 측도라고 할 수 있다. 또한 신뢰도 $conf_1$과 $conf_2$, 그리고 기여 순수 신뢰도 AP_1과 AP_2는 모든 경우에 다른 값을 취하고 있으므로 두 항목 X, Y가 바뀌면 이들 두 측도는 값이 달라진다는 사실을 알 수 있다. 반면에 대칭적 기여 순수 신뢰도 모두는 전항과 후항이 바뀌더라도 값이 달라지지 않는다. 네 가지의 대칭적 기여 순수 신뢰도 중에서는 $NSAP$만이 유일하게 AP_1과 AP_2의 사이 값을 갖는 반면에 다른 측도들은 둘 중 작은 것보다 작거나 큰 것보다는 큰 값을 취하고 있다. 이를 좀 더 구체적으로 확인하기 위해 $a = 17$, $b = 33$, $c = 13$, $d = 37$인 경우를 살펴보면 $conf_1 = 0.340$, $conf_2 = 0.567$로 나타나므로 두 항목의 발생확률이 다른 경우에 전항과 후항을 바꾸면 신뢰도의 값이 달라진다. 또한 기여 순수 신뢰도 역시 $AP_1 = 0.235$와 $AP_2 = 0.168$로 나타나므로 전항과 후항이 바뀌게 되면 값이 달라진다는 것을 알 수 있다. 대칭적 기여 순수 신뢰도들은 $SAP_1 = 0.138$, $SAP_2 = 0.336$, $SAP_3 = 0.142$, 그리고 $NSAP = 0.196$으로 계산되었다. 따라서 SAP_1과 SAP_3은 AP_1과 AP_2 중에서 작은 값인 AP_2보다 작게 나타났고 이 두 측도 중에는 SAP_1이 더 작은 값을 취하고 있다. SAP_2는 AP_1과 AP_2 중에서 큰 값인 AP_1보다 크게 나타난 반면에 $NSAP$는 AP_1보다는 작고 AP_2보다는 큰 값으로 계산되었다.

따라서 네 가지의 대칭적 기여 순수 신뢰도 중에서는 $NSAP$가 기여 순수 신뢰도 AP_1과 AP_2를 적절히 조정한 측도라고 볼 수 있다. Table 3.2를 동시 비 발생 빈도 d의 관점에서 탐색해보아도 이와 유사한 결과를 얻을 수 있다.

이번에는 Table 3.3을 활용하여 두 항목간의 불일치빈도 b의 값의 변화에 따른 지지도 및 각종 신뢰도들의 변화하는 양상을 파악하고자 한다.

Table 3.3 Simulation data(2)

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

이 표에서 b가 취할 수 있는 정수 값의 범위는 $0 \le b \le 30$이며, 이 표를 이용하여 계산한 결과의 일부를 Table 3.4에 제시하였다. 여기서 $a = 50-b$, $c = b+20$, 그리고 $d = 30-b$이다. 이 표에서 $P(X)$와 $P(Y)$를 계산하면 각각 0.500과 0.700로 나타난다. 이 표에서 보는 바와 같이 불일치빈도 b가 증가함에 따라 본 논문에서 고려하고 있는 네 가지의 대칭적 기여 순수 신뢰도를 포함한 모든 평가 기준의 값들이 감소하고 있으며, 위의 표와 마찬가지로 지지도와 신뢰도를 제외한 모든 측도들이 양의 값과 음의 값을 동시에 나타내고 있다. 따라서 기여 순수 신뢰도와 대칭적 기여 순수 신뢰도는 기본적인 신뢰도에 비해 바람직한 연관성 측도라고 할 수 있다. 또한 Table 3.2의 결과와 마찬가지로 $conf_1$과 $conf_2$, 그리고 AP_1과 AP_2는 각기 다른 값을 취하므로 X, Y가 바뀌면 이들 두 측도는 값이 달라지는 반면에 네 가지 대칭적 기여 순수 신뢰도는 전항과 후항이 바뀌더라도 값이 달라지지 않는다. 여기서도 $NSAP$는 AP_1과 AP_2의 사이 값을 갖는 반면에 다른 대칭적 기여 순수 신뢰도들은 둘 중 작은 것보다 작거나 큰 것보다는 큰 값으로 나타나고 있다. 이들에 대해 좀 더 구체적으로 탐색하기 위해 $a = 43$, $b = 7$, $c = 27$, $d = 23$인 경우를 살펴보면 $conf_1 = 0.860$, $conf_2 = 0.614$로 나타나므로 전항과 후항을 바꾸면 신뢰도의 값이 달라진다는 것을 알 수 있다. 또한 $AP_1 = 0.372$와 $AP_2 = 0.620$으로 기여 순수 신뢰도의 값이 계산되므로 전항과 후항이 바뀌게 되면 이 또한 값이 달라진다는 것을 알 수 있다. 대칭적 기여 순수 신뢰도들은 $SAP_1 = 0.286$, $SAP_2 = 1.240$, $SAP_3 = 0.326$, 그리고 $NSAP = 0.465$로 계산되었다. 따라서 SAP_1과 SAP_3은 AP_1과 AP_2 중에서 작은 값인 AP_1보다 작게 나타났고 이 두 측도 중에는 SAP_1이 더 작은 값을 취하고 있으며, SAP_2는 AP_1과 AP_2 중에서 큰 값인 AP_2보다 크게 나타났다. 반면에 $NSAP$는 AP_1보다는 크고 AP_2보다는 작은 값으로 나타났다. 따라서 여기서도 $NSAP$가 AP_1과 AP_2를 적절히 절충한 측도라고 볼 수 있다. Table 3.3을 또 다른 불일치 빈도 c의 관점에서 살펴보아도 이와 유사한 결과를 얻게 된다.

Table 3.4 Variation of several interestingness measures by simulation data(2)

a	b	c	d	$supp$	$conf_1$	$conf_2$	AP_1	AP_2	SAP_1	SAP_2	SAP_3	$NSAP$
45	5	25	25	0.450	0.900	0.643	0.444	0.741	0.342	1.481	0.404	0.556
44	6	26	24	0.440	0.880	0.629	0.409	0.682	0.315	1.364	0.365	0.511
43	7	27	23	0.430	0.860	0.614	0.372	0.620	0.286	1.240	0.326	0.465
42	8	28	22	0.420	0.840	0.600	0.333	0.556	0.256	1.111	0.287	0.417
41	9	29	21	0.410	0.820	0.586	0.293	0.488	0.225	0.976	0.248	0.366
40	10	30	20	0.400	0.800	0.571	0.250	0.417	0.192	0.833	0.208	0.313
39	11	31	19	0.390	0.780	0.557	0.205	0.342	0.158	0.684	0.168	0.256
38	12	32	18	0.380	0.760	0.543	0.158	0.263	0.121	0.526	0.127	0.197
37	13	33	17	0.370	0.740	0.529	0.108	0.180	0.083	0.360	0.086	0.135
36	14	34	16	0.360	0.720	0.514	0.056	0.093	0.043	0.185	0.043	0.069
35	15	35	15	0.350	0.700	0.500	0.000	0.000	0.000	0.000	0.000	0.000
34	16	36	14	0.340	0.680	0.486	-0.059	-0.098	-0.045	-0.196	-0.045	-0.074
33	17	37	13	0.330	0.660	0.471	-0.121	-0.202	-0.093	-0.404	-0.090	-0.152
32	18	38	12	0.320	0.640	0.457	-0.188	-0.313	-0.144	-0.625	-0.138	-0.234
31	19	39	11	0.310	0.620	0.443	-0.258	-0.430	-0.199	-0.860	-0.187	-0.323
30	20	40	10	0.300	0.600	0.429	-0.333	-0.556	-0.256	-1.111	-0.238	-0.417
29	21	41	9	0.290	0.580	0.414	-0.414	-0.690	-0.318	-1.379	-0.291	-0.517
28	22	42	8	0.280	0.560	0.400	-0.500	-0.833	-0.385	-1.667	-0.347	-0.625
27	23	43	7	0.270	0.540	0.386	-0.593	-0.988	-0.456	-1.975	-0.406	-0.741
26	24	44	6	0.260	0.520	0.371	-0.692	-1.154	-0.533	-2.308	-0.468	-0.865
25	25	45	5	0.250	0.500	0.357	-0.800	-1.333	-0.615	-2.667	-0.533	-1.000
24	26	46	4	0.240	0.480	0.343	-0.917	-1.528	-0.705	-3.056	-0.603	-1.146
23	27	47	3	0.230	0.460	0.329	-1.043	-1.739	-0.803	-3.478	-0.678	-1.304
22	28	48	2	0.220	0.440	0.314	-1.182	-1.970	-0.909	-3.939	-0.758	-1.477
21	29	49	1	0.210	0.420	0.300	-1.333	-2.222	-1.026	-4.444	-0.844	-1.667
20	30	50	0	0.200	0.400	0.286	-1.500	-2.500	-1.154	-5.000	-0.938	-1.875

4. 결론

오늘날 빅 데이터가 신 성장 동력으로 주목받고 있는 동시에 이에 대한 장밋빛 전망이 쏟아지고 있다. 빅 데이터 분석을 위한 데이터마이닝 기법 중의 하나인 연관성 규칙은 탐색적이고 비목적성 분석 기법인 동시에 계산이 용이하며, 지지도, 신뢰도, 향상도 등의 여러 가지 연관성 평가기준을 기반으로 하여 항목집합들 간의 관련성을 찾아내는 데 활용되고 있다. 기본적인 연관성 평가기준들 중에서 신뢰도가 가장 많이 활용되고 있으나 신뢰도는 연관성의 방향을 알 수가 없다는 단점을 가지고 있다. 이를 보완하기 위한 측도로 순수 신뢰도가 개발되었으나. 이 또한 양의 신뢰도의 값과 음의 신뢰도의 값이 동일한 경우에는 순수 신뢰도의 값이 같아지므로 이러한 경우에는 순수 신뢰도로는 차이를 알 수 없다. 기존의 신뢰도와 순수 신뢰도의 단점을 보완하기 위한 측도로 기여 순수 신뢰도가 제안되었으나, 이는 전항과 후항이 바뀌면 그 값이 달라지는 문제점이 있다. 이에 본 논문에서는 기존의 신뢰도와 순수 신뢰도, 그리고 기여 순수 신뢰도의 단점을 보완한 연관성 평가 기준으로 네 가지의 대칭적 기여 순수 신뢰도를 제안하였다. 이들 측도에 대해 Piatetsky-Shapiro (1991)가 제안한 흥미도 측도의 조건을 증명한 결과, 모두 충족하는 것으로 나타났다. 또한 신뢰도와 기여 순수 신뢰도, 그리고 네 가지의 대칭적 기여 순수 신뢰도를 예제를 통하여 유용성을 비교해 본 결과, 대칭적 기여 순수 신뢰도들은 그 부호에 의해 연관성 규칙의 방향을 파악할 수 있는 동시에 전항과 후항이 바뀌어도 그 값이 변하지 않으므로 연관성 규칙을 생성하는 데 매우 유익한 평가 기준이라는 사실을 확인할 수 있었다. 이들 네 가지 대칭적 기여 순수 신뢰도 중에서는 측도 $NSAP$가 두 가지의 기여 순수 신뢰도의 분자의 합과 분모의 합의 비로 나타나는 측도인 동시에 두 값을 잘 조정하여 나타내는 사실이 예제를 통해 확인되었다. 따라서 본 논문에서 고려한 측도들 중에서는 측도 $NSAP$가 가장 바람직한 측도라고 할 수 있다. 그러나 이 측도의 범위가 $[-1, +1]$을 초과하고 있어서 행태적인 해석이 곤란하다는 한계점을 가지고 있으므로 향후에는 이를 보완할 수 있는 측도가 개발되어야 할 것으로 사료된다.

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.

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.

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.

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

Geng, L. and Hamilton, H. J. (2006). Interestingness measures for data mining: A survey. *ACM Computing Surveys*, **38**, 1-32.

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.

Hilderman, R. J. and Hamilton, H. J. (1999). *Knowledge discovery and interestingness measures: A survey*, Technical Report CS 99-04, Department of Computer Science, University of Regina, 1-27.

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.

Omiecinski, E. R. (2003). Alternative interest measures for mining associations in databases. *IEEE Transactions on Knowledge and Data Engineering*, **15**, 57-69.

Park, H. C. (2011a). Association rule ranking function by decreased lift influence. *Journal of the Korean Data & Information Science Society*, **22**, 179-188.

Park, H. C. (2011b). The proposition of attributably pure confidence in association rule mining. *Journal of the Korean Data & Information Science Society*, **22**, 235-243.

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. (2013a). The proposition of compared and attributably pure confidence in association rule mining. *Journal of the Korean Data & Information Science Society*, **24**, 523-532.

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

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.

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 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 32-41.

Wu, T., Chen, Y. and Han, J. (2010). Re-examination of interestingness measures in pattern mining: A unified framework. *Data Mining and Knowledge Discovery*, **21**, 371-397.

Journal of the Korean Data & Information Science Society
2014, **25**(3), 601–609

http://dx.doi.org/10.7465/jkdi.2014.25.3.601
한국데이터정보과학회지

The development of symmetrically and attributably pure confidence in association rule mining

Hee Chang Park[1]

[1]Department of Statistics, Changwon National University

Received 2 April 2014, revised 21 April 2014, accepted 19 May 2014

Abstract

The most widely used data mining technique for big data analysis is to generate meaningful association rules. This method has been used to find the relationship between set of items based on the association criteria such as support, confidence, lift, etc. Among them, confidence is the most frequently used, but it has the drawback that we can not know the direction of association by it. The attributably pure confidence was developed to compensate for this drawback, but the value was changed by the position of two item sets. In this paper, we propose four symmetrically and attributably pure confidence measures to compensate the shortcomings of confidence and the attributably pure confidence. And then we prove three conditions of interestingness measure by Piatetsky-Shapiro, and comparative studies with confidence, attributably pure confidence, and four symmetrically and attributably pure confidence measures are shown by numerical examples. The results show that the symmetrically and attributably pure confidence measures are better than confidence and the attributably pure confidence. Also the measure $NSAP$is found to be the best among these four symmetrically and attributably pure confidence measures.

Keywords: Association criteria, attributably pure confidence, confidence, symmetrically and attributably pure confidence.

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