Journal of the Korean Data & Information Science Society
2014, **25**(6), 1345–1352

http://dx.doi.org/10.7465/jkdi.2014.25.6.1345
한국데이터정보과학회지

상대적 규칙 정확도의 균형화에 의한 연관성 측도의 개발

박희창[1]

[1]창원대학교 통계학과

접수 2014년 9월 11일, 수정 2014년 10월 2일, 게재확정 2014년 10월 13일

요 약

데이터마이닝 기법 중에서 연관성 규칙은 연관성 평가 기준을 기반으로 하여 데이터베이스에 포함되어 있는 항목들 간의 관련성을 탐색하는 기법이다. 일반적인 연관성 규칙 기법과는 달리 역의 연관성 규칙은 하나의 항목집합이 발생하지 않으면 다른 항목집합도 발생하지 않는다는 규칙을 찾아내는 것이다. 이러한 역의 연관성 규칙을 일반적인 연관성 규칙과 함께 생성하면 기업체에서 특정 제품을 판매하기 위해서는 그 제품만의 마케팅뿐만 아니라 더 나아가 어떤 제품의 마케팅이 필요한 지에 대한 정보를 파악할 수 있다. 이를 위해 본 논문에서는 이러한 두 종류의 연관성 규칙에 적용 가능한 균형화된 기여 상대적 규칙 정확도를 연관성 평가 기준으로 제안하고자 한다. 또한 Piatetsky-Shapiro (1991)가 제안한 흥미도 측도가 가져야 할 조건들을 점검한 후, 예제를 통하여 제안된 측도와 연관성 규칙에 적용 가능한 의학진단분야의 평가 측도들의 유용성을 비교하였다. 그 결과, 기여 상대적 정확도와 역의 기여 상대적 정확도의 크기가 다르게 나타나면 연관성의 정도를 명확하게 설명하기가 어려우므로 이들 두 측도를 동시에 고려한 균형화된 기여 상대적 규칙 정확도를 이용하는 것이 가장 바람직하다는 사실을 확인하였다.

주요용어: 규칙정확도, 균형화된 기여 상대적 규칙 정확도, 기여 상대적 규칙 정확도, 상대적 정확도, 역의 연관성 규칙.

1. 서론

오늘날 데이터의 양의 기하급수적으로 증가하고 있다. 위키 백과사전에 의하면 데이터마이닝 (data mining)은 엄청난 크기의 데이터베이스 내에서 알려져 있지 않은 규칙이나 패턴을 체계적이고도 자동적으로 찾아내는 것으로, 신용평가모형 개발, 사기탐지, 장바구니 분석 등과 같은 분야에서 광범위하게 활용되고 있다. 여러 가지 데이터마이닝 기법 중에서 연관성 규칙 (association rule)은 연관성 평가 기준을 기반으로 하여 데이터베이스에 포함되어 있는 항목들 간의 관련성을 탐색하는 데 활용되고 있다 (Park, 2012b). 이 기법은 Agrawal 등 (1993)이 처음으로 제안하였으며, 최근에는 Han 등 (2000), Pei 등 (2000), Cho와 Park (2011a, 2011b), Jin 등 (2011), 그리고 Park (2012a, 2012b, 2013) 등의 연구가 진행되었다.

일반적인 연관성 규칙은 어떤 항목집합이 발생하면 다른 항목집합도 발생한다는 규칙을 발견하는 기법인 반면에, 역의 연관성 규칙 (inverse association rule)은 하나의 항목집합이 발생하지 않으면 다른 항목집합도 발생하지 않는다는 규칙을 찾아내는 것이다 (Park, 2010). 일반적인 연관성 규칙은 전항 항목을 고정시키고 후항 항목을 마케팅 하는 반면에 역의 연관성 규칙을 추가로 생성하게 되면 후항 항목

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

을 고정시키고 전항 항목을 마케팅 하는 전략도 가능하게 된다 (Hwang과 Kim, 2003). Park (2010)에서 제시한 바와 같이 커피를 구매하는 사람은 크리머를 구매하는 경향이 많다는 정보가 의미 있는 동시에 커피를 구매하지 않는 사람은 크리머를 구매하지 않는다는 정보도 의미 있다고 할 수 있다. 따라서 크리머의 매출액을 높이기 위해서는 커피를 구매한 사람들 중에서 크리머를 구매하지 않은 사람들에게 크리머의 구매를 유도하는 것뿐만 아니라 커피를 구매하지 않은 사람들에게도 우선 커피를 구매하게 한 다음 크리머를 구매하게 하는 마케팅 전략도 필요하다. 이러한 역의 연관성 규칙을 일반적인 연관성 규칙과 함께 생성하면 어떤 제품을 판매하기 위해서는 그 제품만의 마케팅뿐만 아니라 더 나아가 어떤 제품의 마케팅이 필요한 지에 대한 정보를 파악할 수 있다.

이러한 두 종류의 연관성 규칙에 적용 가능한 의학진단분야의 평가 측도들 중에는 규칙 정확도 (rule accuracy), 상대적 정확도 (relative accuracy), 음의 신뢰도 (negative reliability), 그리고 상대적 음의 신뢰도 (relative negative reliability) 등이 있다 (Lavrac 등, 1999). 본 논문에서는 이 측도들을 이용한 기여 상대적 정확도 (attributable relative accuracy)와 역의 기여 상대적 정확도 (inversely attributable relative accuracy)를 정의한 후, 두 측도를 동시에 고려한 균형화된 기여 상대적 규칙 정확도 (balanced attributable relative accuracy)를 연관성 평가 기준으로 제안하고자 한다. 또한 Piatetsky-Shapiro (1991)가 제안한 흥미도 측도가 가져야 할 조건들을 점검한 후, 예제를 통하여 균형화된 기여 상대적 규칙 정확도의 유용성을 고찰하고자 한다.

2. 균형화된 상대적 정확도

의학진단분야에서 활용되고 있는 규칙 평가 측도들 중에서 규칙 정확도와 상대적 정확도는 양의 연관성 규칙에서 활용할 수 있고, 음의 신뢰도와 상대적 음의 신뢰도는 역의 연관성 규칙에서 활용 가능하며, 다음과 같이 표현된다.

$$\text{규칙 정확도} : AC(X \Rightarrow Y) = P(Y|X)$$
$$\text{상대적 정확도} : RA(X \Rightarrow Y) = P(Y|X) - P(Y)$$
$$\text{음의 신뢰도} : IAC(X \Rightarrow Y) = P(\overline{Y}|\overline{X})$$
$$\text{상대적 음의 신뢰도} : IRA(X \Rightarrow Y) = P(\overline{Y}|\overline{X}) - P(\overline{Y})$$

규칙 정확도는 기존의 연관성 평가 기준인 양의 신뢰도 (positive confidence)와 동일하며, 정보 검색 분야에서는 정밀도 (precision)라고 불리어진다. 하나의 항목이 발생하지 않으면 다른 항목도 발생하지 않는다는 역의 연관성 규칙 관점에서 볼 때, 음의 신뢰도는 역의 규칙 정확도 (inverse rule accuracy)라고 할 수 있다. 상대적 정확도는 항목 X의 발생에 대한 정확도 이득 (accuracy gain)을 의미하고, 정보 검색 분야에서는 정밀도 (precision)라고 불리어진다. 상대적 음의 신뢰도 역시 역의 연관성 규칙 관점에서 볼 때 역의 상대적 정확도 (inverse relative accuracy)라고 할 수 있다.

이러한 측도들에 대해 Park (2011)이 제안한 기여 순수 신뢰도 (attributably pure confidence)의 개념을 적용하여 기여 상대적 정확도 (attributable relative accuracy)와 역의 기여 상대적 정확도 (inversely attributable relative accuracy)를 정의하면 다음과 같다.

$$\text{기여 상대적 정확도} : ARA(X \Rightarrow Y) = \frac{P(Y|X) - P(Y)}{P(Y|X)}$$
$$\text{역의 기여 상대적 정확도} : IARA(X \Rightarrow Y) = \frac{P(\overline{Y}|\overline{X}) - P(\overline{Y})}{P(\overline{Y}|\overline{X})}$$

기여 상대적 정확도는 규칙 정확도에 대한 상대적 정확도의 크기를 나타내며, 역의 기여 상대적 정확도는 역의 규칙 정확도에 대한 역의 상대적 정확도의 크기를 나타낸다. 이로부터 균형화된 기여 상대적 규칙 정확도 (balanced attributable relative accuracy)를 다음과 같이 정의한다.

$$BARA(X \Rightarrow Y) = \frac{P(Y|X)ARA(X \Rightarrow Y) + P(\overline{Y}|\overline{X})IARA(X \Rightarrow Y)}{P(Y|X) + P(\overline{Y}|\overline{X})} \tag{2.1}$$

$$= \frac{[P(Y|X) - P(Y)] + [P(\overline{Y}|\overline{X}) - P(\overline{Y})]}{P(Y|X) + P(\overline{Y}|\overline{X})} \tag{2.2}$$

이 식에서 보는 바와 같이 $BARA(X \Rightarrow Y)$는 규칙 정확도와 역 규칙 정확도에 대한 $ARA(X \Rightarrow Y)$와 $IARA(X \Rightarrow Y)$의 가중 산술 평균으로 양의 연관성과 역의 연관성의 강도를 동시에 고려한 측도라고 할 수 있다. 만약 기존의 $ARA(X \Rightarrow Y)$의 값이 동일하게 나타나서 연관성의 정도를 비교하기가 곤란한 경우에 $BARA(X \Rightarrow Y)$를 이용하게 되면 연관성의 강도를 보다 정확하게 측정할 수 있다. McNicholas 등 (2008)이 제안한 표준화 방법을 고려하면 $BARA(X \Rightarrow Y)$는 다음의 구간을 갖게 된다.

$$1 - \frac{P(X)P(Y)}{\max[P(X) + P(Y) - 1,\ 1/n]} \leq BARA(X \Rightarrow Y) \leq 1 - \frac{P(X)P(Y)}{\min[P(X), P(Y), P(Y|X)]}$$

다음으로는 $BARA(X \Rightarrow Y)$에 대해 Piatetsky-Shapiro (1991)가 제안한 흥미도 측도의 조건을 충족하는지의 여부를 알아보기 위해 식 (2.1)을 정리하면 다음과 같이 표현된다.

$$BARA(X \Rightarrow Y) = \frac{P(\overline{X})[P(XY) - P(X)P(Y)] + P(X)[P(\overline{XY}) - P(\overline{X})P(\overline{Y})]}{P(\overline{X})P(XY) + P(X)P(\overline{XY})}$$

이 식으로부터 $P(XY) = P(X)P(Y)$이면 $BARA(X \Rightarrow Y)$의 분자가 0이 된다. 또한 이 식을 재정리하면 다음과 같이 나타낼 수 있다.

$$BARA(X \Rightarrow Y) = 1 - \frac{P(X)[1 - P(X)]}{P(X)[1 - P(X)] + P(XY) - P(X)P(Y)}$$

따라서 $P(XY)$의 값이 증가함에 따라 $BARA(X \Rightarrow Y)$는 단조 증가하고, $P(Y)$의 값이 증가하면 $BARA(X \Rightarrow Y)$는 단조 감소한다는 사실을 알 수 있다.

3. 적용 예제

본 절에서는 의학진단분야에서 활용되고 있는 규칙 평가 측도들 중에서 규칙 정확도, 상대적 정확도, 역의 규칙 정확도, 역의 상대적 정확도뿐만 아니라 본 논문에서 고려한 기여 상대적 정확도, 역의 기여 상대적 정확도, 그리고 균형화된 기여 상대적 정확도를 예제를 통해 비교해봄으로써 본 논문에서 고려하고 있는 측도의 유용성을 고찰하고자 한다. 이를 위해 Park (2011)과 동일한 데이터를 활용하고자 한다.

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 3.1에서 보는 바와 같이 전체 트랜잭션의 수를 100명, 항목 X의 발생빈도는 50명, 그리고 항목 Y의 발생빈도를 30명으로 하였다. 항목 집합 X와 Y가 동시에 발생한 빈도 수, 즉 동시발생빈도는 a명으로 하였으며, a가 취할 수 있는 범위는 $0 \le a \le 30$이고 $P(Y) = 0.300$이다.

Table 3.1로부터 a값에 대해 본 논문에서 고려하는 정확도들을 계산하면 Table 3.2와 같은 결과를 얻을 수 있다. 여기서 $a = P(X = 1, Y = 1)$, $b = P(X = 1, Y = 0)$, $c = P(X = 0, Y = 1)$, $d = P(X = 0, Y = 0)$을 의미한다. 이 표에서 보는 바와 같이 동시발생빈도 a가 증가함에 따라 본 논문에서 고려하는 모든 정확도들이 증가하는 것으로 나타났다. 또한 두 항목 간에 양의 연관성의 정도가 음의 연관성 정도보다 더 강한 경우에는 $AC(X \Rightarrow Y)$와 $IAC(X \Rightarrow Y)$는 0.5 보다 크고, 다른 규칙 정확도 $RA(X \Rightarrow Y)$, $IRA(X \Rightarrow Y)$, $ARA(X \Rightarrow Y)$, $IARA(X \Rightarrow Y)$, $BARA(X \Rightarrow Y)$는 0보다 큰 값을 갖는다. 이와 반대의 경우에는 $AC(X \Rightarrow Y)$와 $IAC(X \Rightarrow Y)$는 0.5 보다 작고, 다른 정확도들은 0보다 작은 값을 갖는다. 그러나 $AC(X \Rightarrow Y)$와 $IAC(X \Rightarrow Y)$는 모두 양의 값으로만 나타나기 때문에 이들 측도로는 음의 연관성을 파악하기가 어렵다.

Table 3.2 Comparison of accuracy measures by simulation data(1)

a	b	c	d	$supp$	AC	IAC	RA	IRA	ARA	$IARA$	$BARA$
1	29	49	21	0.010	0.033	0.300	-0.467	-0.200	-14.000	-0.667	-2.000
2	28	48	22	0.020	0.067	0.314	-0.433	-0.186	-6.500	-0.591	-1.625
3	27	47	23	0.030	0.100	0.329	-0.400	-0.171	-4.000	-0.522	-1.333
4	26	46	24	0.040	0.133	0.343	-0.367	-0.157	-2.750	-0.458	-1.100
5	25	45	25	0.050	0.167	0.357	-0.333	-0.143	-2.000	-0.400	-0.909
6	24	44	26	0.060	0.200	0.371	-0.300	-0.129	-1.500	-0.346	-0.750
7	23	43	27	0.070	0.233	0.386	-0.267	-0.114	-1.143	-0.296	-0.615
8	22	42	28	0.080	0.267	0.400	-0.233	-0.100	-0.875	-0.250	-0.500
9	21	41	29	0.090	0.300	0.414	-0.200	-0.086	-0.667	-0.207	-0.400
10	20	40	30	0.100	0.333	0.429	-0.167	-0.071	-0.500	-0.167	-0.313
11	19	39	31	0.110	0.367	0.443	-0.133	-0.057	-0.364	-0.129	-0.235
12	18	38	32	0.120	0.400	0.457	-0.100	-0.043	-0.250	-0.094	-0.167
13	17	37	33	0.130	0.433	0.471	-0.067	-0.029	-0.154	-0.061	-0.105
14	16	36	34	0.140	0.467	0.486	-0.033	-0.014	-0.071	-0.029	-0.050
15	15	35	35	0.150	0.500	0.500	0.000	0.000	0.000	0.000	0.000
16	14	34	36	0.160	0.533	0.514	0.033	0.014	0.063	0.028	0.045
17	13	33	37	0.170	0.567	0.529	0.067	0.029	0.118	0.054	0.087
18	12	32	38	0.180	0.600	0.543	0.100	0.043	0.167	0.079	0.125
19	11	31	39	0.190	0.633	0.557	0.133	0.057	0.211	0.103	0.160
20	10	30	40	0.200	0.667	0.571	0.167	0.071	0.250	0.125	0.192
21	9	29	41	0.210	0.700	0.586	0.200	0.086	0.286	0.146	0.222
22	8	28	42	0.220	0.733	0.600	0.233	0.100	0.318	0.167	0.250
23	7	27	43	0.230	0.767	0.614	0.267	0.114	0.348	0.186	0.276
24	6	26	44	0.240	0.800	0.629	0.300	0.129	0.375	0.205	0.300
25	5	25	45	0.250	0.833	0.643	0.333	0.143	0.400	0.222	0.323
26	4	24	46	0.260	0.867	0.657	0.367	0.157	0.423	0.239	0.344
27	3	23	47	0.270	0.900	0.671	0.400	0.171	0.444	0.255	0.364
28	2	22	48	0.280	0.933	0.686	0.433	0.186	0.464	0.271	0.382
29	1	21	49	0.290	0.967	0.700	0.467	0.200	0.483	0.286	0.400

이 표를 좀 더 구체적으로 살펴보기 위해 $a = 8$, $b = 22$, $c = 42$, $d = 28$인 경우와 $a = 19$, $b = 11$, $c = 31$, $d = 39$인 경우를 비교해보면 먼저 음의 연관성 정도가 양의 연관성 정도보다 큰 전자의 경우에는 $AC(X \Rightarrow Y)$와 $IAC(X \Rightarrow Y)$는 각각 (0.267, 0.400)로 나타나서 0.5보다 작으며, 이와 반대인 후자의 경우에는 이들 두 측도가 각각 (0.633, 0.557)로 나타나서 0.5보다 큰 값으로 나타났다. 그러나 모두 양의 값으로 나타나고 있어서 연관성의 방향을 알 수가 없다. 또 다른 정확도 $RA(X \Rightarrow Y)$,

$IRA(X \Rightarrow Y)$, $ARA(X \Rightarrow Y)$, $IARA(X \Rightarrow Y)$, $BARA(X \Rightarrow Y)$는 각각 (-0.233, -0.100, -0.875, -0.250, -0.500)과 (0.133, 0.057, 0.211, 0.103, 0.160)으로 나타나서 음의 연관성 강도가 강한 전자의 경우에는 이들 측도 모두 음의 값을 가지며, 양의 연관성의 강도가 강한 후자의 경우에는 이들 모두 0보다 큰 값으로 나타났다. 따라서 $AC(X \Rightarrow Y)$와 $IAC(X \Rightarrow Y)$보다는 연관성의 방향을 알게 해주는 $RA(X \Rightarrow Y)$, $IRA(X \Rightarrow Y)$, $ARA(X \Rightarrow Y)$, $IARA(X \Rightarrow Y)$, 그리고 $BARA(X \Rightarrow Y)$의 측도들이 더 바람직하다고 할 수 있다.

연관성 규칙의 관점에서는 $AC(X \Rightarrow Y)$보다 $RA(X \Rightarrow Y)$가 연관성의 정도를 좀 더 바람직하게 나타낸다고 할 수 있다. 그 이유는 $P(Y|X)$가 상당히 큰 값을 갖는다고 해도 $P(Y)$가 큰 경우에는 두 항목이 연관성의 정도가 강하다고 할 수 없기 때문이다. 이와 마찬가지로 $P(\overline{Y})$가 큰 경우에는 $P(\overline{Y}|\overline{X})$가 큰 값을 갖는다고 할지라도 두 항목 간에 역의 연관성이 있다고 하기에는 무리가 따르므로 $IAC(X \Rightarrow Y)$보다는 $IRA(X \Rightarrow Y)$를 이용하는 것이 더 바람직하다. 또한 $AC(X \Rightarrow Y)$와 $P(Y)$, $IAC(X \Rightarrow Y)$와 $P(\overline{Y})$가 같은 값을 갖는 경우에는 $RA(X \Rightarrow Y)$와 $IRA(X \Rightarrow Y)$의 값은 0이 된다. 또한 $P(Y|X)$와 $P(Y)$의 차이 및 $P(\overline{Y}|\overline{X})$와 $P(\overline{Y})$의 차이가 동일하면 $RA(X \Rightarrow Y)$와 $IRA(X \Rightarrow Y)$가 같은 값으로 나타나므로 이 경우에는 $ARA(X \Rightarrow Y)$와 $IARA(X \Rightarrow Y)$를 사용하는 것이 더 바람직하다. 그러나 이들 두 측도는 각각 양의 연관성 정도와 역의 연관성 정도만을 나타내므로 이 둘을 동시에 고려한 $BARA(X \Rightarrow Y)$를 이용하는 것이 가장 바람직한 것으로 판단된다.

이번에는 두 항목간의 불일치빈도 b의 값의 변화에 따라 정확도들의 변화하는 양상을 파악하기 위해 Table 3.3과 같은 분할표를 이용하고자 한다. 이 표에서 b가 취할 수 있는 정수 값의 범위는 $0 \leq b \leq 20$이고, $P(Y) = 0.800$이다.

Table 3.3 Simulation data(2)

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

이 표로부터 불일치빈도 b값의 변화에 따른 여러 가지 정확도를 계산하면 다음 Table 3.4와 같은 결과를 얻을 수 있다. 이 표에서는 불일치빈도 b가 증가함에 따라 고려 대상 정확도들이 모두 감소하는 것으로 나타났다. $AC(X \Rightarrow Y)$와 $IAC(X \Rightarrow Y)$는 모든 경우에 양의 값으로 나타나기 때문에 연관성의 방향을 파악할 수가 없다. 앞의 결과에서와 마찬가지로 두 항목 간에 양의 연관성이 음의 연관성보다 더 크면 $RA(X \Rightarrow Y)$, $IRA(X \Rightarrow Y)$, $ARA(X \Rightarrow Y)$, $IARA(X \Rightarrow Y)$, $BARA(X \Rightarrow Y)$는 양의 값을 갖는다. 좀 더 구체적으로 알아보기 위해 $a = 27$, $b = 3$, $c = 53$, $d = 17$인 경우와 $a = 18$, $b = 12$, $c = 62$, $d = 8$인 경우를 비교해보면 먼저 양의 연관성 정도가 음의 연관성 정도보다 강한 전자의 경우에는 $AC(X \Rightarrow Y)$와 $IAC(X \Rightarrow Y)$는 각각 (0.900, 0.243)로 나타나고 이와 반대인 후자의 경우에는 이들 두 측도가 각각 (0.600, 0.114)로 나타났다. 따라서 이 두 측도의 값을 비교해본 결과, 값들의 차이는 확인할 수 있었으나 연관성의 방향은 확인하기가 곤란하였다. 또 다른 정확도 $RA(X \Rightarrow Y)$, $IRA(X \Rightarrow Y)$, $ARA(X \Rightarrow Y)$, $IARA(X \Rightarrow Y)$, $BARA(X \Rightarrow Y)$는 각각 (0.100, 0.043, 0.111, 0.176, 0.125)와 (-0.200, -0.086, -0.333, -0.750, -0.400)으로 나타나서 양의 연관성 강도가 강한 전자의 경우에는 이들 측도 모두가 0보다 큰 값으로 나타난 반면에, 음의 연관성의 강도가 강한 후자의 경우에는 이들 모두 0보다 작은 값으로 나타났다. 따라서 $AC(X \Rightarrow Y)$와 $IAC(X \Rightarrow Y)$보다는 연관성의 방향을 알 수 있는 $RA(X \Rightarrow Y)$, $IRA(X \Rightarrow Y)$, $ARA(X \Rightarrow Y)$, $IARA(X \Rightarrow Y)$, 그리고 $BARA(X \Rightarrow Y)$의 측도들이 더 바람직하다고 할 수 있다.

Table 3.4 Comparison of accuracy measures by simulation data(2)

a	b	c	d	$supp$	AC	IAC	RA	IRA	ARA	$IARA$	$BARA$
30	0	50	20	0.300	1.000	0.286	0.200	0.086	0.200	0.300	0.222
29	1	51	19	0.290	0.967	0.271	0.167	0.071	0.172	0.263	0.192
28	2	52	18	0.280	0.933	0.257	0.133	0.057	0.143	0.222	0.160
27	3	53	17	0.270	0.900	0.243	0.100	0.043	0.111	0.176	0.125
26	4	54	16	0.260	0.867	0.229	0.067	0.029	0.077	0.125	0.087
25	5	55	15	0.250	0.833	0.214	0.033	0.014	0.040	0.067	0.045
24	6	56	14	0.240	0.800	0.200	0.000	0.000	0.000	0.000	0.000
23	7	57	13	0.230	0.767	0.186	-0.033	-0.014	-0.043	-0.077	-0.050
22	8	58	12	0.220	0.733	0.171	-0.067	-0.029	-0.091	-0.167	-0.105
21	9	59	11	0.210	0.700	0.157	-0.100	-0.043	-0.143	-0.273	-0.167
20	10	60	10	0.200	0.667	0.143	-0.133	-0.057	-0.200	-0.400	-0.235
19	11	61	9	0.190	0.633	0.129	-0.167	-0.071	-0.263	-0.556	-0.313
18	12	62	8	0.180	0.600	0.114	-0.200	-0.086	-0.333	-0.750	-0.400
17	13	63	7	0.170	0.567	0.100	-0.233	-0.100	-0.412	-1.000	-0.500
16	14	64	6	0.160	0.533	0.086	-0.267	-0.114	-0.500	-1.333	-0.615
15	15	65	5	0.150	0.500	0.071	-0.300	-0.129	-0.600	-1.800	-0.750
14	16	66	4	0.140	0.467	0.057	-0.333	-0.143	-0.714	-2.500	-0.909
13	17	67	3	0.130	0.433	0.043	-0.367	-0.157	-0.846	-3.667	-1.100
12	18	68	2	0.120	0.400	0.029	-0.400	-0.171	-1.000	-6.000	-1.333
11	19	69	1	0.110	0.367	0.014	-0.433	-0.186	-1.182	-13.000	-1.625

한편, Table 3.2의 후자의 경우와 Table 3.4의 전자의 경우를 비교해보면 $ARA(X \Rightarrow Y)$는 Table 3.2의 경우가 Table 3.4의 경우에 비해 더 큰 반면에 $IARA(X \Rightarrow Y)$는 그 반대로 나타나고 있다. 이러한 경우에는 두 항목 간에 연관성의 강도가 어느 경우가 더 큰지를 알 수 없으므로 이들 두 측도의 가중평균인 $BARA(X \Rightarrow Y)$를 사용하는 것이 연관성 규칙 생성의 관점에서는 보다 합리적이라고 할 수 있다. $P(Y|X)$와 $P(\overline{Y}|\overline{X})$의 합이 1보다 큰 경우에는 $ARA(X \Rightarrow Y)$와 $IARA(X \Rightarrow Y)$의 가중평균인 $BARA(X \Rightarrow Y)$의 값이 0보다 큰 값으로 나타났으며, $P(Y|X)$가 $P(\overline{Y}|\overline{X})$의 값보다 크면 $BARA(X \Rightarrow Y)$는 $IARA(X \Rightarrow Y)$보다는 $ARA(X \Rightarrow Y)$에 가까운 값으로 나타났다. 이 경우에는 양의 연관성 정도가 역의 연관성 정도보다 더 크게 나타났다. 결국 $BARA(X \Rightarrow Y)$는 $ARA(X \Rightarrow Y)$와 $IARA(X \Rightarrow Y)$를 절충한 측도라고 볼 수 있으며, $ARA(X \Rightarrow Y)$ 또는 $IARA(X \Rightarrow Y)$의 값이 동일하다고 할지라도 $BARA(X \Rightarrow Y)$의 값에 의해 연관성의 강도를 좀 더 명확하게 표현할 수 있다. 또 다른 불일치빈도 c와 동시 비 발생 빈도 d에 대해 정확도들의 변화하는 양상을 살펴보았는데, 이 경우에도 위에서 논의한 결과와 유사한 결과를 얻을 수 있었다.

4. 결론

본 논문에서는 양의 연관성 규칙과 역의 연관성 규칙을 동시에 고려한 균형화된 기여 상대적 규칙 정확도를 제안하였으며, 이 측도가 흥미도 측도의 조건에 충족한다는 사실을 확인하였다. 연관성 규칙에 적용 가능한 의학진단분야의 평가 측도들 중에서 규칙 정확도는 기존의 연관성 평가 기준인 양의 신뢰도와 동일하며, 역의 연관성 규칙 관점에서 볼 때, 음의 신뢰도는 역의 규칙 정확도라고 할 수 있다. 또한 상대적 정확도는 특정 항목의 발생에 대한 정확도 이득을 의미하고, 정보 상대적 음의 신뢰도 역시 역의 연관성 규칙 관점에서 볼 때 역의 상대적 정확도라고 할 수 있다. 기여 상대적 정확도는 규칙 정확도에 대한 상대적 정확도의 크기를 나타내며, 역의 기여 상대적 정확도는 역의 규칙 정확도에 대한 역의 상대적 정확도의 크기를 나타낸다. 본 논문에서 제안한 균형화된 기여 상대적 규칙 정확도는 규칙 정확도와 역 규칙 정확도에 대한 기여 상대적 정확도와 역의 기여 상대적 정확도의 가중 산술 평균으로 양의 연관

성과 역의 연관성의 강도를 동시에 고려한 측도라고 할 수 있다.

또한 예제를 통해 제안된 측도와 연관성 규칙에 적용 가능한 의학진단분야의 평가 측도들의 유용성을 비교하였다. 먼저 동시발생빈도가 증가하면 본 논문에서 고려하는 모든 정확도들이 증가하고, 불일치빈도가 증가하면 모든 정확도들이 감소하는 것으로 나타났다. 또한 두 항목 간에 양의 연관성의 정도가 음의 연관성 정도보다 더 강한 경우에는 규칙 정확도와 역의 규칙 정확도는 0.5 보다 크고, 다른 규칙 정확도들은 0보다 큰 값으로 나타났다. 이와 반대의 경우에는 규칙 정확도와 역의 규칙 정확도는 0.5 보다 작고, 다른 정확도들은 0보다 작은 값으로 나타났다. 따라서 규칙 정확도와 역의 규칙 정확도는 모두 양의 값으로만 나타나기 때문에 이들 측도로는 연관성의 방향을 파악하기가 어려운 반면에 다른 정확도들에 의해서는 연관성의 방향을 파악할 수 있었다. $P(Y|X)$가 상당히 큰 값을 갖는다고 해도 $P(Y)$가 큰 경우에는 두 항목이 연관성의 정도가 강하다고 할 수 없기 때문에 연관성 규칙의 관점에서는 규칙 정확도보다 상대적 정확도가 연관성의 정도를 좀 더 바람직하게 나타낸다고 할 수 있다. 이와 마찬가지로 역의 규칙 정확도보다는 역의 상대적 정확도를 이용하는 것이 더 바람직한 것으로 나타났다. 기여 상대적 정확도와 역의 기여 상대적 정확도의 크기가 다르게 나타나면 연관성의 정도를 명확하게 설명하기가 어려우므로 이들의 가중평균인 균형화된 기여 상대적 규칙 정확도를 이용하는 것이 가장 바람직하다는 사실을 확인하였다.

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.

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

Hwang, J. and Kim, J. (2003). Target marketing using inverse association rule. *Journal of Intelligence and Information Systems*, **9**, 195-209.

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

Lavrac, N., Flach, P. and Zupan, B. (1999). Rule evaluation measures: A unifying view. *Proceedings of the 9th International Workshop on Inductive Logic Programming*, 174-185.

McNicholas, P. D., Murphy, T. B. and O'Regan, O. (2008). Standardising the lift of an association rule. *Computational Statistics and Data Analysis*, **52**, 4712-4721.

Park, H. C. (2010). Proposition of inverse pure association rule. *Journal of the Korean Data Analysis Society*, **12**, 3305-3315.

Park, H. C. (2011). The proposition of attributably pure confidence in association rule mining. *Journal of the Korean Data and 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. (2013). 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.

Journal of the Korean Data & Information Science Society
2014, **25**(6), 1345–1352

http://dx.doi.org/10.7465/jkdi.2014.25.6.1345
한국데이터정보과학회지

Development of association rule threshold by balancing of relative rule accuracy

Hee Chang Park[1]

[1]Department of Statistics, Changwon National University

Received 11 September 2014, revised 2 October 2014, accepted 13 October 2014

Abstract

Data mining is the representative methodology to obtain meaningful information in the era of big data.By Wikipedia, association rule learning is a popular and well researched method for discovering interesting relationship between itemsets in large databases using association thresholds. It is intended to identify strong rules discovered in databases using different interestingness measures. Unlike general association rule, inverse association rule mining finds the rules that a special item does not occur if an item does not occur. If two types of association rule can be simultaneously considered, we can obtain the marketing information for some related products as well as the information of specific product marketing. In this paper, we propose a balanced attributable relative accuracy applicable to these association rule techniques, and then check the three conditions of interestingness measures by Piatetsky-Shapiro (1991). The comparative studies with rule accuracy, relative accuracy, attributable relative accuracy, and balanced attributable relative accuracy are shown by numerical example. The results show that balanced attributable relative accuracy is better than any other accuracy measures.

Keywords: Attributable relative accuracy, balanced attributable relative accuracy, inverse association rule, relative accuracy, rule accuracy.

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