Journal of the Korean Data & Information Science Society
2015, **26**(3), 611–618

http://dx.doi.org/10.7465/jkdi.2015.26.3.611
한국데이터정보과학회지

모든 가능한 진단도구를 활용한 균형비교신뢰도의 제안

박희창[1]

[1]창원대학교 통계학과

접수 2015년 4월 13일, 수정 2015년 5월 11일, 게재확정 2015년 5월 18일

요 약

오늘날 정보 기술과 소셜미디어의 확산으로 인하여 빅 데이터에 관심이 집중되고 있다. 이를 처리하기 위한 기술 중의 하나가 데이터마이닝기법인데, 이들 중에는 연관성 규칙이 많이 활용되고 있다. 연관성 규칙은 방향에 따라 양, 음, 그리고 역의 연관성 규칙 등이 존재하며, 평가 기준을 설정하고자 하는 경우에는 이들 세 가지 연관성 규칙을 동시에 고려하는 것이 바람직하다고 할 수 있다. 이를 위해 본 논문에서는 의학진단분야에서 활용되고 있는 진단도구들 중에서 민감도, 특이도, 위양성도, 그리고 위음성도를 고려한 균형비교신뢰도를 제안하고자 한다. 또한 흥미도 측도가 가져야 할 조건들을 점검한 후, 예제를 통하여 측도의 유용성을 고찰하였다. 그 결과, 균형비교신뢰도는 비교신뢰도와 역의 비교신뢰도가 양의 값을 가지는 경우에는 양의 값을 가지며, 이들 두 값이 음인 경우에는 음으로 나타났다. 따라서 연관성 규칙의 평가 기준 관점에서 볼 때 비교신뢰도와 역의 비교신뢰도를 개별적으로 이용하기 보다는 균형비교신뢰도를 활용하는 것이 더 바람직하다고 할 수 있다.

주요용어: 균형비교신뢰도, 비교신뢰도, 빅 데이터, 역의 비교신뢰도, 연관성 규칙.

1. 서론

최근에 정보기술의 급속한 발달로 인해 빅 데이터 (big data)가 국가뿐만 아니라 사회전반에 걸쳐서 초미의 관심영역으로 부상하고 있으며, 이에 적용 가능한 데이터 마이닝 (data mining) 기술이 주목받고 있다 (Park, 2013a). 데이터 마이닝은 엄청난 크기의 데이터베이스로부터 알려져 있지 않은 패턴이나 규칙을 체계적으로 탐색하는 분석 방법으로 시장바구니 분석, 이탈고객 분류, 신용평가모형 개발, 사기탐지, 그리고 수요예측 등과 같이 현업에서 광범위하게 활용되고 있다 (Jin 등, 2011). 연관성 규칙 (association rule), 의사결정나무 (decision tree), 군집화 (clustering), 연결 분석 (link analysis), 신경망모형 (neural network) 등 여러 가지 데이터 마이닝 기법 중에서 연관성 규칙은 지지도 (support), 신뢰도 (confidence), 향상도 (lift) 등의 흥미도 측도들 (interestingness measures)에 의해 명확히 수치화한 연관성 평가 기준을 기반으로 하여 데이터베이스에 포함되어 있는 항목들 간의 관련성을 탐색하는데 활용되고 있다 (Park, 2012). 이 기법과 관련된 최근의 연구로는 Lavrac 등 (1999), Hilderman과 Hamilton (2000), Liu 등 (2000), Berzal 등 (2001), Ahn과 Kim (2003), McNicholas 등 (2008), Kuo (2009), Jin 등 (2011), 그리고 Park (2011a, 2011b, 2012, 2013a, 2013b, 2014a, 2014b, 2014c) 등이 있다.

연관성 규칙은 방향에 따라 양의 연관성 규칙 (positive association rule), 음의 연관성 규칙 (negative association rule), 그리고 역의 연관성 규칙 (inverse association rule) 등이 있다. 양의 연관성 규칙은

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

하나의 항목집합이 발생하면 다른 항목집합도 발생한다는 규칙을 발견하는 기법이고, 음의 연관성 규칙은 어떤 항목이 발생하면 다른 항목은 발생하지 않는 규칙을 발견하는 기법이며, 역의 연관성 규칙은 특정 항목이 발생하지 않으면 다른 항목도 발생하지 않는다는 규칙을 찾아내는 것이다. 양과 음의 연관성 규칙은 전항 항목을 고정시키고 후항 항목을 마케팅 하는 반면에 역의 연관성 규칙을 추가로 생성하게 되면 후항 항목을 고정시키고 전항 항목을 마케팅 하는 전략도 가능하게 된다 (Hwang과 Kim, 2003). 따라서 연관성 규칙을 생성하기 위한 평가 기준을 설정하고자 하는 경우에는 이들 세 가지 연관성 규칙을 동시에 고려하는 것이 바람직하다고 할 수 있다. 이를 위해 본 논문에서는 의학진단분야에서 활용되고 있는 민감도 (sensitivity), 특이도 (specificity), 위양성도 (false positive), 그리고 위음성도 (false negative)를 고려한 균형비교신뢰도 (balanced comparative confidence)를 제안하고자 한다. 또한 Piatetsky-Shapiro (1991)가 제안한 흥미도 측도가 가져야 할 조건들을 점검한 후, 예제를 통하여 제안하는 측도의 유용성을 고찰하고자 한다.

2. 균형비교신뢰도

이 절에서는 Park (2013a)에서와 마찬가지로 다음과 같은 2×2 분할표를 활용하여 의학진단분야에서 스크리닝 검사의 정확도를 판단하는 기준으로 활용되고 있는 규칙 평가 측도들 중에서 민감도, 특이도, 위양성도, 그리고 위음성도를 정의한다. 여기서 X는 질병유무를 나타내고 Y는 진단결과를 의미한다.

Table 2.1 2×2 contingency table by proportions

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

$$\text{민감도} : Sen(X \Rightarrow Y) = P(Y|X) = \frac{a}{a+b}$$

$$\text{특이도} : Spe(X \Rightarrow Y) = P(\bar{Y}|\bar{X}) = \frac{d}{c+d}$$

$$\text{위양성도} : Fp(X \Rightarrow Y) = P(Y|\bar{X}) = \frac{c}{c+d}$$

$$\text{위음성도} : Fn(X \Rightarrow Y) = P(\bar{Y}|X) = \frac{b}{a+b}$$

스크리닝 검사 (screening test)는 주로 의학 분야에서 질병의 조기 진단이나 환자의 정밀검사 여부를 판단하는 수단으로 매우 광범위하게 사용되고 있다 (Kim, 2002). 특히 의학 분야에서는 민감도가 환자를 양성으로 판정하는 능력인 동시에 위음성도와의 합이 1이 되고, 특이도는 건강인을 음성으로 판정하는 능력으로 위양성도와의 합이 1이 된다. 따라서 민감도와 특이도의 값이 클수록 검사가 정확하다고 할 수 있다. 연관성 규칙의 평가 기준 관점에서 볼 때, 민감도는 양의 신뢰도가 되고 위양성도는 역의 신뢰도, 특이도와 위음성도는 음의 신뢰도를 나타낸다고 할 수 있어서 민감도와 특이도는 두 항목의 연관성의 방향이 동일한 반면에, 위양성도와 위음성도는 두 항목의 연관성의 방향이 동일하지 않다고 할 수 있다. 따라서 민감도와 특이도의 합과 위양성도와 위음성도의 합을 비교해봄으로써 두 항목 간의 연관성

강도를 측정할 수 있다. 본 논문에서는 다음의 식과 같이 진단도구를 고려한 균형비교신뢰도를 제안하고자 한다.

$$BCC(X \Rightarrow Y) = \frac{CC(X \Rightarrow Y)P(Y|\bar{X}) + ICC(X \Rightarrow Y)P(\bar{Y}|X)}{P(Y|\bar{X}) + P(\bar{Y}|X)}$$
$$= \frac{[P(Y|X) - P(Y|\bar{X})] + [P(\bar{Y}|\bar{X}) - P(\bar{Y}|X)]}{P(Y|\bar{X}) + P(\bar{Y}|X)}$$

여기서 $CC(X \Rightarrow Y)$와 $ICC(X \Rightarrow Y)$는 각각 비교신뢰도 (comparative confidence)와 역의 비교신뢰도 (inversely comparative confidence)를 의미한다 (Kuo, 2009; Park, 2013a).

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

$BCC(X \Rightarrow Y)$에 대해 Piatetsky-Shapiro가 제안한 흥미도 측도의 세 가지 조건의 충족여부를 알아보기 위해 먼저 $BCC(X \Rightarrow Y)$를 $P(Y)$에 관한 식으로 정리하면 다음과 같다.

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

이 식의 분자는 $P(Y)$가 증가함에 따라 감소하고, 분모는 $P(Y)$가 증가함에 따라 증가하므로 $BCC(X \Rightarrow Y)$는 $P(Y)$가 증가함에 따라 감소하게 되어 첫 번째 조건을 만족하는 것을 알 수 있다. 또한 이 식으로부터 $P(XY)$가 증가하면 분자는 증가하고 분모는 감소하게 되어 $BCC(X \Rightarrow Y)$는 증가하므로 두 번째 조건을 만족한다. 마지막으로 $P(XY) = P(X)P(Y)$이면 원래의 식의 값이 0이 되므로 마지막 조건을 만족하는 것을 알 수 있다.

3. 적용 예제

본 절에서는 예제를 통하여 진단도구인 민감도, 특이도, 위양성도, 그리고 위음성도의 값이 변함에 따른 비교신뢰도, 역의 비교신뢰도, 그리고 균형비교신뢰도의 변화하는 양상을 탐색하고자 한다. 이를 위해 Park (2011b)와 유사한 여러 가지 분할표를 활용하고자 한다. 먼저 Table 3.1에서는 전체 트랜잭션의 수를 100명, 항목 X의 발생빈도는 30명, 그리고 항목 Y의 발생빈도를 50명으로 하였다. 항목 X와 Y의 동시발생빈도 $n(X = 1, Y = 1)$는 a명으로 하였으며, a가 취할 수 있는 범위는 $0 \leq a \leq 30$이다.

Table 3.1 Simulation data (1)

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

Table 3.1로부터 a값에 대해 본 논문에서 고려하는 진단도구들과 비교신뢰도, 역의 비교신뢰도, 그리고 균형비교신뢰도를 계산하면 Table 3.2와 같다. 여기서 $b = n(X = 1, Y = 0)$, $c = n(X = 0, Y = 1)$,

$d = n(X = 0, Y = 0)$을 의미한다. 이 표에서 보는 바와 같이 동시발생빈도 a와 동시비발생빈도 d가 증가하고, 불일치빈도 b와 c가 감소함에 따라 민감도와 특이도는 증가하고 있고, 위양성도와 위음성도는 감소하고 있다. 이에 따라 비교신뢰도, 역의 비교신뢰도, 그리고 균형비교신뢰도는 모두 증가하고 있는 것으로 나타났다. 민감도의 값이 위양성도의 값보다 큰 경우에는 비교신뢰도의 값이 양으로 나타난 반면에 민감도의 값이 위양성도의 값보다 작은 경우에는 비교신뢰도의 값이 음으로 나타났다. 특이도의 값이 위음성도의 값보다 큰 경우에는 역의 비교신뢰도는 양의 값으로 나타난 반면에 특이도의 값이 위음성도의 값보다 작은 경우에는 역의 비교신뢰도는 음의 값으로 나타났다. 비교신뢰도와 역의 비교신뢰도가 양의 값을 가지는 경우에는 균형비교신뢰도는 양의 값을 가지며, 이들 두 값이 음인 경우에는 균형비교신뢰도도 음의 값을 가지는 것으로 나타났다. 또한 균형비교신뢰도는 위양성도와 위음성도의 값에 대한 비교신뢰도와 역의 비교신뢰도의 가중평균이므로 이들 두 측도의 사이에 위치하는 값으로 나타났다.

Table 3.2 Comparison of CC, ICC and BCC by simulation data (1)

a	b	c	d	Sen	Fp	Spe	Fn	CC	ICC	BCC
1	29	49	21	0.033	0.700	0.300	0.967	-0.952	-0.690	-0.800
2	28	48	22	0.067	0.686	0.314	0.933	-0.903	-0.663	-0.765
3	27	47	23	0.100	0.671	0.329	0.900	-0.851	-0.635	-0.727
4	26	46	24	0.133	0.657	0.343	0.867	-0.797	-0.604	-0.688
5	25	45	25	0.167	0.643	0.357	0.833	-0.741	-0.571	-0.645
6	24	44	26	0.200	0.629	0.371	0.800	-0.682	-0.536	-0.600
7	23	43	27	0.233	0.614	0.386	0.767	-0.620	-0.497	-0.552
8	22	42	28	0.267	0.600	0.400	0.733	-0.556	-0.455	-0.500
9	21	41	29	0.300	0.586	0.414	0.700	-0.488	-0.408	-0.444
10	20	40	30	0.333	0.571	0.429	0.667	-0.417	-0.357	-0.385
11	19	39	31	0.367	0.557	0.443	0.633	-0.342	-0.301	-0.320
12	18	38	32	0.400	0.543	0.457	0.600	-0.263	-0.238	-0.250
13	17	37	33	0.433	0.529	0.471	0.567	-0.180	-0.168	-0.174
14	16	36	34	0.467	0.514	0.486	0.533	-0.093	-0.089	-0.091
15	15	35	35	0.500	0.500	0.500	0.500	0.000	0.000	0.000
16	14	34	36	0.533	0.486	0.514	0.467	0.098	0.102	0.100
17	13	33	37	0.567	0.471	0.529	0.433	0.202	0.220	0.211
18	12	32	38	0.600	0.457	0.543	0.400	0.313	0.357	0.333
19	11	31	39	0.633	0.443	0.557	0.367	0.430	0.519	0.471
20	10	30	40	0.667	0.429	0.571	0.333	0.556	0.714	0.625
21	9	29	41	0.700	0.414	0.586	0.300	0.690	0.952	0.800
22	8	28	42	0.733	0.400	0.600	0.267	0.833	1.250	1.000
23	7	27	43	0.767	0.386	0.614	0.233	0.988	1.633	1.231
24	6	26	44	0.800	0.371	0.629	0.200	1.154	2.143	1.500
25	5	25	45	0.833	0.357	0.643	0.167	1.333	2.857	1.818
26	4	24	46	0.867	0.343	0.657	0.133	1.528	3.929	2.200
27	3	23	47	0.900	0.329	0.671	0.100	1.739	5.714	2.667
28	2	22	48	0.933	0.314	0.686	0.067	1.970	9.286	3.250
29	1	21	49	0.967	0.300	0.700	0.033	2.222	20.000	4.000

이를 좀 더 구체적으로 탐색하기 위해 $a = 7$, $b = 23$, $c = 43$, $d = 27$인 경우와 $a = 19$, $b = 11$, $c = 31$,

$d = 39$인 경우를 비교해보면 전자의 경우에는 민감도와 특이도가 각각 0.233과 0,386, 위양성도와 위음성도는 각각 0.614와 0.767로 나타난 반면에 후자는 민감도와 특이도가 각각 0.633과 0.557, 위양성도와 위음성도는 각각 0.443과 0.367로 나타나서 민감도와 특이도는 증가하고 위양성도와 위음성도는 감소하는 것으로 나타났다. 또한 비교신뢰도는 각각 -0.620과 0.430, 역의 비교신뢰도는 각각 -0.497과 0.519, 그리고 균형비교신뢰도는 각각 -0.552와 0.471로 나타나서 민감도가 증가하고 위양성도가 감소하므로 비교신뢰도는 증가하는 것으로 나타났고, 특이도가 증가하고 위음성도가 감소하므로 역의 비교신뢰도 역시 증가하는 것으로 나타났다. 이러한 연유로 균형비교신뢰도가 전자의 경우에는 음의 값으로 나타난 반면에 후자의 경우에는 양의 값으로 나타났으며, 각 경우의 비교신뢰도와 역의 비교신뢰도의 사이에 있는 값을 취하고 있다. 따라서 균형비교신뢰도는 음의 연관성 평가 기준을 고려한 상태에서 양의 연관성 평가 기준과 역의 연관성 평가 기준을 절충한 측도라고 할 수 있다.

이번에는 Table 3.3의 분할표를 이용하여 두 항목간의 불일치빈도 c의 값의 변화에 따른 진단도구들 및 여러 가지 비교 신뢰도들의 변화하는 양상을 알아보고자 한다. 이 표에서 c가 취할 수 있는 정수 값의 범위는 $0 \le c \le 30$이다.

Table 3.3 Simulation data (2)

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

Table 3.3으로부터 c값의 변화량에 대해 진단도구들과 비교신뢰도, 역의 비교신뢰도, 그리고 균형비교신뢰도를 계산하면 Table 3.4와 같다. 이 표에서는 a와 d가 감소하고, b와 c가 증가하므로 민감도와 특이도는 감소하고 있고, 위양성도와 위음성도는 증가하고 있다. 이에 따라 비교신뢰도, 역의 비교신뢰도, 그리고 균형비교신뢰도는 모두 감소하는 것으로 나타났다. Table 3.2에서와 마찬가지로 민감도의 값이 위양성도보다 크면 비교신뢰도의 값이 양으로 나타났고, 민감도의 값이 위양성도보다 작은 경우에는 비교신뢰도는 음의 값을 가진다. 특이도의 값이 위음성도보다 크면 역의 비교신뢰도의 값은 양으로 나타난 반면에 특이도의 값이 위음성도보다 작은 경우에는 역의 비교신뢰도는 음의 값을 가진다. 이 표에서도 비교신뢰도와 역의 비교신뢰도가 양이면 균형비교신뢰도도 양의 값을 가지며, 이들 두 값이 음이면 균형비교신뢰도도 음의 값으로 나타났으며, 균형비교신뢰도의 값의 크기는 이들 두 측도의 작은 값보다는 크고 큰 값보다는 작게 나타났다. 이러한 사실을 좀 더 자세하게 알아보기 위해 $a = 37$, $b = 33$, $c = 13$, $d = 17$인 경우와 $a = 27$, $b = 43$, $c = 23$, $d = 7$인 경우를 비교해보면 전자의 경우에는 민감도와 특이도가 각각 0.529와 0,567, 위양성도와 위음성도는 각각 0.433과 0.471로 나타난 반면에 후자는 민감도와 특이도가 각각 0.386과 0.233, 위양성도와 위음성도는 각각 0.767과 0.614로 나타나서 민감도와 특이도는 각각 감소하는 반면에 위양성도와 위음성도는 각각 증가하는 것으로 나타났다. 또한 비교신뢰도는 전자와 후자의 경우 각각 0.220과 -0.497, 역의 비교신뢰도는 각각 0.202와 -0.620, 그리고 균형비교신뢰도는 각각 0.211과 -0.552로 나타나서 민감도가 감소하고 위양성도가 증가하므로 비교신뢰도는 감소하고, 특이도가 감소하고 위음성도가 증가하므로 역의 비교신뢰도도 감소하는 것으로 나타났다. 따라서 균형비교신뢰도는 전자의 경우에는 양으로 나타난 반면에 후자의 경우에는 음으로 나타났으며, 각 경우의 비교신뢰도와 역의 비교신뢰도의 가운데 값을 취하고 있다. 또 다른 불일치빈도 b와 동시 비 발생 빈도 d에 대해 여러 가지 진단도 측도들과 비교신뢰도들의 변화하는 양상을 살펴보았는데 위에서 논의한 결과와 유사한 결과를 얻을 수 있었다.

Table 3.4 Comparison of CC, ICC and BCC by simulation data (2)

a	b	c	d	Sen	Fp	Spe	Fn	CC	ICC	BCC
49	21	1	29	0.700	0.033	0.967	0.300	20.000	2.222	4.000
48	22	2	28	0.686	0.067	0.933	0.314	9.286	1.970	3.250
47	23	3	27	0.671	0.100	0.900	0.329	5.714	1.739	2.667
46	24	4	26	0.657	0.133	0.867	0.343	3.929	1.528	2.200
45	25	5	25	0.643	0.167	0.833	0.357	2.857	1.333	1.818
44	26	6	24	0.629	0.200	0.800	0.371	2.143	1.154	1.500
43	27	7	23	0.614	0.233	0.767	0.386	1.633	0.988	1.231
42	28	8	22	0.600	0.267	0.733	0.400	1.250	0.833	1.000
41	29	9	21	0.586	0.300	0.700	0.414	0.952	0.690	0.800
40	30	10	20	0.571	0.333	0.667	0.429	0.714	0.556	0.625
39	31	11	19	0.557	0.367	0.633	0.443	0.519	0.430	0.471
38	32	12	18	0.543	0.400	0.600	0.457	0.357	0.313	0.333
37	33	13	17	0.529	0.433	0.567	0.471	0.220	0.202	0.211
36	34	14	16	0.514	0.467	0.533	0.486	0.102	0.098	0.100
35	35	15	15	0.500	0.500	0.500	0.500	0.000	0.000	0.000
34	36	16	14	0.486	0.533	0.467	0.514	-0.089	-0.093	-0.091
33	37	17	13	0.471	0.567	0.433	0.529	-0.168	-0.180	-0.174
32	38	18	12	0.457	0.600	0.400	0.543	-0.238	-0.263	-0.250
31	39	19	11	0.443	0.633	0.367	0.557	-0.301	-0.342	-0.320
30	40	20	10	0.429	0.667	0.333	0.571	-0.357	-0.417	-0.385
29	41	21	9	0.414	0.700	0.300	0.586	-0.408	-0.488	-0.444
28	42	22	8	0.400	0.733	0.267	0.600	-0.455	-0.556	-0.500
27	43	23	7	0.386	0.767	0.233	0.614	-0.497	-0.620	-0.552
26	44	24	6	0.371	0.800	0.200	0.629	-0.536	-0.682	-0.600
25	45	25	5	0.357	0.833	0.167	0.643	-0.571	-0.741	-0.645
24	46	26	4	0.343	0.867	0.133	0.657	-0.604	-0.797	-0.688
23	47	27	3	0.329	0.900	0.100	0.671	-0.635	-0.851	-0.727
22	48	28	2	0.314	0.933	0.067	0.686	-0.663	-0.903	-0.765
21	49	29	1	0.300	0.967	0.033	0.700	-0.690	-0.952	-0.800
20	50	30	0	0.286	1.000	0.000	0.714	-0.714	-1.000	-0.833

4. 결론

일반적으로 제품이나 시스템은 모집단으로부터 일부를 시료로 추출하여 불량여부를 검사한 후 불량률을 검사하게 된다. 하지만 제품의 불량을 검사하기가 경비나 시간이 많이 소요되는 경우에는 간편하게 검사 절차를 간소화한 스크리닝 검사를 통해서 불량 여부를 판단하는 경우도 있다 (Kim, 2002). 의학 분야에서는 민감도와 특이도, 그리고 위양성도와 위음성도를 고려하여 검사의 정확성 여부를 판단한다. 본 논문에서는 이러한 진단도구를 고려한 균형비교신뢰도를 제안하였으며, 흥미도 측도의 조건을 만족한다는 사실을 확인하였다. 또한 예제를 이용하여 고찰해본 결과, 동시발생빈도와 동시비발생빈도가 증가하고, 불일치빈도들이 감소함에 따라 민감도와 특이도는 증가하고 있고, 위양성도와 위음성도는 감소하고 있어서 비교신뢰도, 역의 비교신뢰도, 그리고 균형비교신뢰도는 모두 증가하고 있는 것으로 나타났다. 반면에 동시발생빈도와 동시비발생빈도가 감소하고, 불일치빈도들이 증가함에 따라 민감도와 특

이도는 감소하고 위양성도와 위음성도는 증가하므로 이들 비교신뢰도들은 모두 감소하는 것으로 나타났다. 또한 비교신뢰도의 값은 민감도가 위양성도보다 큰 경우에는 양으로 나타난 반면에 민감도가 위양성도보다 작은 경우에는 음의 값으로 나타났다. 역의 비교신뢰도의 경우에는 특이도가 위음성도보다 큰 경우에는 양의 값으로 나타난 반면에 특이도가 위음성도보다 작은 경우에는 음의 값으로 나타났다. 균형비교신뢰도는 비교신뢰도와 역의 비교신뢰도가 양의 값을 가지는 경우에는 양의 값을 가지며, 이들 두 값이 음인 경우에는 음으로 나타났다. 이러한 균형비교신뢰도는 위양성도와 위음성도에 대한 비교신뢰도와 역의 비교신뢰도의 가중평균이므로 음의 연관성 평가 기준을 고려한 상태에서 양의 연관성 평가 기준과 역의 연관성 평가 기준을 절충한 측도라고 할 수 있다. 따라서 연관성 규칙의 평가 기준 관점에서 볼 때 비교신뢰도와 역의 비교신뢰도를 개별적으로 이용하기 보다는 균형비교신뢰도를 활용하는 것이 더 바람직하다고 할 수 있다.

References

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

Berzal, F., Blanco, I., Sanchez, D. and Vila, M. (2001). A new framework to assess association rules. *Proceedings of the 4th International Conference on Intelligent Data Analysis*, 95-104.

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*, 432-439.

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 association rules. *Journal of the Korean Data Analysis Society*, **13**, 2945-2952.

Kim, T. (2002). Estimation of defect rate from the screening test - The case of unknown sensitivity and specificity. *Journal of the Korean Society for Quality Management*, **30**, 144-151.

Kuo, Y. T. (2009) *Mining surprising patterns*, The doctoral paper of Melbourne university, Australia.

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.

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

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

Park, H. C. (2011b). Proposition of symmetrically pure confidence in association rule discovery. *Journal of the Korean Data Analysis Society*, **13**, 879-890.

Park, H. C. (2012). Exploration of symmetric similarity measures by conditional probabilities as association rule thresholds. *Journal of the Korean Data Analysis Society*, **14**, 707-716.

Park, H. C. (2013a). The proposition of c ompared and a ttributably pure confidence in association rule mining. *Journal of the Korean Data & Information Science Society*, **24**, 523-532.

Park, H. C. (2013b). A proposition of association rule thresholds considering relative occurrence/ nonoccurrence. *Journal of the Korean Data Analysis Society*, **15**, 1841-1850.

Park, H. C. (2014a). Comparison of confidence measures useful for classification model building. *Journal of the Korean Data & Information Science Society*, **25**, 365-371.

Park, H. C. (2014b). Proposition of causally confirmed measures in association rule mining. *Journal of the Korean Data & Information Science Society*, **25**, 857-868.

Park, H. C. (2014c). Development of association rule threshold by balancing of relative rule accuracy. *Journal of the Korean Data & Information Science Society*, **25**, 1345-1352.

Piatetsky-Shapiro, G. (1991). *Knowledge discovery in databases*, MIT Press, Cambridge.

Journal of the Korean Data & Information Science Society
2015, **26**(3), 611–618

http://dx.doi.org/10.7465/jkdi.2015.26.3.611
한국데이터정보과학회지

Proposition of balanced comparative confidence considering all available diagnostic tools

Hee Chang Park[1]

[1]Department of Statistics, Changwon National University

Received 13 April 2015, revised 11 May 2015, accepted 18 May 2015

Abstract

By Wikipedia, big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate. Data mining is the computational process of discovering patterns in huge data sets involving methods at the intersection of association rule, decision tree, clustering, artificial intelligence, machine learning. Association rule is a well researched method for discovering interesting relationships between itemsets in huge databases and has been applied in various fields. There are positive, negative, and inverse association rules according to the direction of association. If you want to set the evaluation criteria of association rule, it may be desirable to consider three types of association rules at the same time. To this end, we proposed a balanced comparative confidence considering sensitivity, specificity, false positive, and false negative, checked the conditions for association threshold by Piatetsky-Shapiro, and compared it with comparative confidence and inversely comparative confidence through a few experiments.

Keywords: Association rule, balanced comparative confidence, big data, comparative confidence, inversely comparative confidence.

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