Journal of the Korean Data & Information Science Society
2014, **25**(2), 365–371

http://dx.doi.org/10.7465/jkdi.2014.25.2.365
한국데이터정보과학회지

분류 모형 구축에 유용한 신뢰도 측도 간의 비교

박희창[1]

[1]창원대학교 통계학과

접수 2014년 2월 4일, 수정 2014년 2월 24일, 게재확정 2014년 3월 3일

요 약

데이터 마이닝 기법 중에서 연관성 규칙은 하나의 거래나 사건에 포함되어 있는 항목들의 관련성을 파악하기 위한 탐색적 자료 분석 방법이다. 이 기법은 지지도, 신뢰도, 향상도 등과 같은 흥미도 측도들을 이용하여 연관성 규칙을 생성한다. 일반적인 연관성 규칙에서는 최소 지지도를 만족하는 빈발항목집합을 생성한 후 최저 신뢰도를 만족하는 것을 연관성 규칙으로 채택하게 된다. 이 때 규칙 여부를 결정하기 위해 가장 많이 사용되는 신뢰도는 고려하는 항목의 순서가 바뀌게 되면 그 값이 달라지는 비대칭적 측도가 되는 동시에 항상 양의 값을 가진다. 따라서 신뢰도 값의 크기로는 양의 연관성이 있는지, 아니면 음의 연관성이 있는지를 알 수 없다. 본 논문에서는 이러한 문제를 극복하기 위해 분류 모형 구축에 유용한 신뢰도 측도들을 소개하고, 신뢰도들 간의 비교 분석을 통해 유용성을 평가하였다. 그 결과, 인과적 확인 신뢰도가 연관성의 방향을 보다 정확하게 나타내고 있다는 사실을 확인하였다.

주요용어: 연관성 규칙, 인과적 신뢰도, 인과적 확인 신뢰도, 확인적 신뢰도, 흥미도 측도.

1. 서론

최근 컴퓨터 및 처리기술이 발달하고 스마트폰의 활성화와 SNS (social network service) 서비스 등으로 인하여 데이터가 폭발적으로 증가하게 되어 빅데이터 (big data) 분석이 화두로 등장하였다. 위키백과사전에 의하면 빅 데이터는 기존 데이터베이스 관리도구로 데이터를 관리하고 분석할 수 있는 역량을 넘어서는 대량의 정형 또는 비정형 데이터 집합 및 이를 처리하는 기술을 의미하며, 국가에서는 이를 창조경제 및 정부3.0의 핵심동력으로 육성하는 방안을 마련하고 있다. 빅데이터를 분석하여 보다 가치있는 정보를 찾기 위한 기술은 지금도 많은 학자들에 의해 연구되고 있으며, 대표적인 기법으로는 데이터 마이닝 (data mining), 텍스트 마이닝 (text mining), 평판 분석 (opinion mining) 사회 연결망분석 (social network analytics), 그리고 클러스터 분석 (cluster analysis) 등이 있다. 이들 중 데이터 마이닝은 대량의 정형 또는 비정형 데이터 집합에서 숨겨진 지식이나 새로운 법칙과 관계를 발견하고 이를 바탕으로 의사결정 등을 위한 정보로 활용하고자 하는 것이다.

데이터 마이닝 기법 중의 하나인 연관성 규칙은 하나의 트랜잭션에 포함되어 있는 항목들의 관련성을 도출하는 탐색적 자료 분석 방법으로 두 항목간의 관계를 지지도 (support), 신뢰도 (confidence), 그리고 향상도 (lift) 등의 흥미도 측도 (interestingness measure)에 의해 명확히 수치화함으로써 항목간의 관련성을 나타내기 때문에 현업에서 많이 활용되고 있다 (Park, 2013). 이 기법은 Agrawal 등 (1993)이 최초로 제안하였으며, 그 이후로 이들 중에는 Han과 Fu (1999), Liu 등 (1999), Pasquier 등

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

(1999), Han 등 (2000), Pei 등 (2000), Saygin 등 (2002), Cho와 Park (2011a, 2011b), Jin 등 (2011), Park (2011a, 2011b, 2012a, 2012b, 2013) 등 많은 학자들에 의해 연구되고 있다. 일반적인 연관성 규칙은 사용자가 지정한 최소 지지도를 만족하는 빈발항목집합을 생성한 후 최저 신뢰도를 만족하는 규칙을 연관성 규칙으로 생성한다 (Park, 2011a). 이 때 규칙 여부를 결정하기 위해 가장 많이 사용되는 신뢰도는 고려하는 항목의 순서가 바뀌게 되면 그 값이 달라지는 비대칭적 측도가 되는 동시에 항상 양의 값을 가진다. 따라서 신뢰도 값의 크기로는 양의 연관성이 있는지, 아니면 음의 연관성이 있는지를 알 수 없다. 이러한 신뢰도의 단점을 보완하기 위해 본 논문에서는 연관성 평가 기준의 관점에서 Kodratoff (2000)가 제안하고 Berzal 등 (2005)이 논의한 분류 모형 구축에 유용한 신뢰도를 비교 분석한 후, 그들의 유용성에 대해 고찰하고자 한다.

2. 여러 가지 신뢰도의 고찰

본 절에서는 하나의 트랜잭션에서 항목집합 X와 Y의 연관성의 정도를 측정하기 위해 Table 2.1과 같은 2×2분할표를 활용하여 분류 모형 구축에 유용한 여러 가지 신뢰도에 대해 논의하고자 한다.

Table 2.1 2×2 contingency table

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

먼저 기존의 연관성 규칙에서 일반적으로 적용하고 있는 지지도 $supp(X \Rightarrow Y)$는 항목 집합 X와 항목 집합 Y가 동시에 발생하는 거래의 비율을 의미하며, Table 2.1로부터 a/n으로 계산된다. 신뢰도 $conf(X \Rightarrow Y)$는 항목 집합 X가 포함된 거래 비율 중 항목 집합 X와 항목 집합 Y가 동시에 포함된 거래의 비율을 의미하며, 다음과 같이 표현된다 (Park, 2012a).

$$conf(X \Rightarrow Y) = P(Y|X) = \frac{a}{a+b}$$

기존의 연관성 평가 기준인 신뢰도는 항목 집합이 발생하는 경우만을 고려한 측도이다. Kodratoff (2000)는 기본적인 연관성 평가 기준만으로는 전항과 후항의 인과관계를 설명할 수 없으므로 $X \Rightarrow Y^c$의 경우를 동시에 고려하여야 한다는 의미에서 식 (2.1)과 같은 확인적 신뢰도 (confirmed confidence)를 분류 목적을 위해 제안한 바 있다.

$$conf_{CO}(X \Rightarrow Y) = [P(Y|X) - P(Y^c|X)] = \frac{a-b}{a+b} \tag{2.1}$$

여기서 Y^c의 의미는 Y가 일어나지 않음을 의미한다. 식 (2.1)에서 보는 바와 같이 확인적 신뢰도 $conf_{CO}$는 기존의 양의 신뢰도에서 음의 신뢰도를 뺀 값으로 정의된다. 또한 그는 $Y^c \Rightarrow X^c$의 경우를 동시에 고려하여야 한다는 의미에서 식 (2.2)와 같은 인과적 신뢰도 (causal confidence)를 분류 목적을 위해 제안한 바 있다.

$$conf_{CA}(X \Rightarrow Y) = \frac{1}{2}[P(Y|X) + P(X^c|Y^c)] = \frac{ab+2ad+bd}{2(a+b)(b+d)} \tag{2.2}$$

여기서 X^c의 의미는 X가 일어나지 않음을 의미한다. 인과적 신뢰도는 하나의 항목이 발생하면 다른 항목이 발생하거나, 하나의 항목이 발생하지 않으면 다른 항목도 발생하지 않는, 즉 방향이 동일한

확률의 산술평균을 나타낸 것으로 볼 수 있다. Park (2013)에서 밝힌 바와 같이 인과적 신뢰도는 기존의 양의 신뢰도의 단점을 보완한 것으로 양의 신뢰도와 역의 신뢰도의 산술평균으로도 해석될 수 있다. 특히 Berzal 등 (2005)은 이들을 동시에 고려하면 잠재적으로 유용한 규칙을 생성할 수 있는 것으로 생각하였다. 또한 Kodratoff (2000)는 흡연과 암발생 유무와의 인과관계를 파악하기 위해서는 흡연(X)하였을 때 암(Y)이 발생하는 경우와 암이 발생하지 않았을 때 흡연하지 않은 경우의 두 가지를 동시에 고려하는 것이 바람직하다고 판단하였다 (Park, 2013). Kodratoff (2000)는 이와 같은 확인적 신뢰도 $conf_{CO}$및 인과적 신뢰도 $conf_{CA}$를 결합한 식 (2.3)의 인과적 확인 신뢰도 (causal confirmed confidence)를 동시에 제안하였다.

$$conf_{CC}(X \Rightarrow Y) = \frac{1}{2}[P(Y|X) + P(X^c|Y^c)] - P(Y^c|X) = \frac{b(a-2b) - d(b-2a)}{2(a+b)(b+d)} \tag{2.3}$$

이러한 인과적 확인 신뢰도는 확인적 신뢰도의 첫 번째 항에 기존의 신뢰도 대신 인과적 신뢰도를 대입한 것으로 인과관계를 고려한 확인적 측도라고 할 수 있다.

3. 예제를 통한 탐색

본 절에서는 예제를 통하여 분류 모형 구축에 유용한 여러 가지 신뢰도들의 변화하는 양상에 대해 탐색하고자 한다. 이를 위해 Park (2011a)에서와 같이 항목 X, Y에 대해 다음과 같이 가정하였다. 먼저 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 신용카드로 결제 (1)한 사람 수를 30명으로 하고 그 외의 방법으로 결제 (0)한 사람의 수를 70명으로 하였다. 두 항목 X와 Y가 동시에 발생한 빈도 수, 즉 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, 분류를 위한 신뢰도를 계산한 결과를 Table 3.2에 제시하였다. 여기서 $b = 50-a$, $c = 30-a$, 그리고 $d = a+20$이다. 이 표에서 보는 바와 같이 a가 증가함에 따라 $P(Y^c|X)$을 제외하고는 모든 평가 기준의 값들이 증가하고 있으며, 인과적 신뢰도 $conf_{CA}$는 양의 값을 갖는 반면에 확인적 신뢰도 $conf_{CO}$와 인과적 확인 신뢰도 $conf_{CC}$는 양과 음의 값을 갖는다. 따라서 연관성 규칙의 평가 기준 관점에서는 $conf_{CO}$와 $conf_{CC}$가 $conf_{CA}$에 비해 더 바람직한 측도라고 할 수 있다. $conf_{CO}$와 $conf_{CC}$중에서는 후자인 인과적 확인 신뢰도가 양의 신뢰도 크기와 음의 신뢰도 크기의 영향을 좀 더 합리적으로 나타내주므로 더 바람직한 측도라고 할 수 있다. 이를 좀 더 구체적으로 알아보기 위해 $a = 24$, $b = 26$, $c = 6$, $d = 44$일 때를 살펴보면 $P(Y|X) = 0.480$이고, $P(Y|X^c) = 0.120$이므로 연관성 평가 기준의 값은 양의 값으로 나타나는 것이 더 바람직하다고 할 수 있는데, 이 경우에는 $conf_{CO}$는 -0.040으로 계산된 반면에 $conf_{CC}$는 0.034라는 양의 값으로 계산되었다. Table 3.3을 동시 비발생빈도 d의 관점에서 탐색해보아도 위와 같은 결과를 얻을 수 있다. 따라서 분류 모형 구축을 위한 신뢰도 중에서는 인과적 확인 신뢰도 $conf_{CC}$가 연관성 평가 기준으로 가장 바람직한 측도라고 할 수 있다.

Table 3.2 Variation of several confidences by simulation data(1)

a	b	c	d	$supp$	$conf$	$P(X^c\|Y^c)$	$P(Y^c\|X)$	$conf_{CA}$	$conf_{CO}$	$conf_{CC}$
1	49	29	21	0.010	0.020	0.300	0.980	0.160	-0.960	-0.820
2	48	28	22	0.020	0.040	0.314	0.960	0.177	-0.920	-0.783
3	47	27	23	0.030	0.060	0.329	0.940	0.194	-0.880	-0.746
4	46	26	24	0.040	0.080	0.343	0.920	0.211	-0.840	-0.709
5	45	25	25	0.050	0.100	0.357	0.900	0.229	-0.800	-0.671
6	44	24	26	0.060	0.120	0.371	0.880	0.246	-0.760	-0.634
7	43	23	27	0.070	0.140	0.386	0.860	0.263	-0.720	-0.597
8	42	22	28	0.080	0.160	0.400	0.840	0.280	-0.680	-0.560
9	41	21	29	0.090	0.180	0.414	0.820	0.297	-0.640	-0.523
10	40	20	30	0.100	0.200	0.429	0.800	0.314	-0.600	-0.486
11	39	19	31	0.110	0.220	0.443	0.780	0.331	-0.560	-0.449
12	38	18	32	0.120	0.240	0.457	0.760	0.349	-0.520	-0.411
13	37	17	33	0.130	0.260	0.471	0.740	0.366	-0.480	-0.374
14	36	16	34	0.140	0.280	0.486	0.720	0.383	-0.440	-0.337
15	35	15	35	0.150	0.300	0.500	0.700	0.400	-0.400	-0.300
16	34	14	36	0.160	0.320	0.514	0.680	0.417	-0.360	-0.263
17	33	13	37	0.170	0.340	0.529	0.660	0.434	-0.320	-0.226
18	32	12	38	0.180	0.360	0.543	0.640	0.451	-0.280	-0.189
19	31	11	39	0.190	0.380	0.557	0.620	0.469	-0.240	-0.151
20	30	10	40	0.200	0.400	0.571	0.600	0.486	-0.200	-0.114
21	29	9	41	0.210	0.420	0.586	0.580	0.503	-0.160	-0.077
22	28	8	42	0.220	0.440	0.600	0.560	0.520	-0.120	-0.040
23	27	7	43	0.230	0.460	0.614	0.540	0.537	-0.080	-0.003
24	26	6	44	0.240	0.480	0.629	0.520	0.554	-0.040	0.034
25	25	5	45	0.250	0.500	0.643	0.500	0.571	0.000	0.071
26	24	4	46	0.260	0.520	0.657	0.480	0.589	0.040	0.109
27	23	3	47	0.270	0.540	0.671	0.460	0.606	0.080	0.146
28	22	2	48	0.280	0.560	0.686	0.440	0.623	0.120	0.183
29	21	1	49	0.290	0.580	0.700	0.420	0.640	0.160	0.220

이번에는 두 항목간의 불일치빈도 b의 값의 변화에 따라 지지도, 신뢰도, 분류를 위한 신뢰도들의 변화하는 양상을 파악하기 위해 Table 3.3을 활용하고자 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.

Table 3.4에서 보는 바와 같이 b가 증가함에 따라 $P(Y^c|X)$을 제외하고는 모든 평가 기준의 값들이 감소하고 있으며, Table 3.2에서와 같이 $conf_{CA}$는 양의 값을 갖는 반면에 $conf_{CO}$와 $conf_{CC}$는 양과 음의 값을 갖는다. 따라서 연관성 규칙의 평가 기준 관점에서는 $conf_{CA}$보다는 연관성의 방향을 알 수 있는 $conf_{CO}$와 $conf_{CC}$가 더 바람직한 측도라고 할 수 있다. $conf_{CO}$와 $conf_{CC}$중에서는 후자가 양의 신뢰도 크기와 음의 신뢰도 크기의 영향을 좀 더 합리적으로 나타내주므로 더 바람직한 측도라고 할 수 있다. 이를 좀 더 구체적으로 알아보기 위해 $a = 26$, $b = 24$, $c = 44$, $d = 6$일 때를 살펴보면 $P(Y|X) = 0.520$이고, $P(Y|X^c) = 0.880$이므로 연관성 평가 기준의 값은 음의 값으로 나타나는 것이 더 바람직하다고 할 수 있는데, 이 경우에는 $conf_{CO}$는 0.040으로 계산된 반면에 $conf_{CC}$는 -0.120으

로 음의 값이 되었다. 또한 이 표를 불일치 빈도 c의 관점에서 여러 가지 신뢰도들의 변화하는 양상을 살펴보아도 이와 동일한 결과를 얻는다. 따라서 이 경우에도 분류 모형 구축을 위한 신뢰도 중에서도 $conf_{CC}$가 연관성 평가 기준으로 가장 바람직한 측도라는 결론을 내릴 수 있다.

Table 3.4 Variation of several confidences by simulation data(2)

a	b	c	d	$supp$	$conf$	$P(X^c\|Y^c)$	$P(Y^c\|X)$	$conf_{CA}$	$conf_{CO}$	$conf_{CC}$
50	0	20	30	0.500	1.000	1.000	0.000	1.000	1.000	1.000
49	1	21	29	0.490	0.980	0.967	0.020	0.973	0.960	0.953
48	2	22	28	0.480	0.960	0.933	0.040	0.947	0.920	0.907
47	3	23	27	0.470	0.940	0.900	0.060	0.920	0.880	0.860
46	4	24	26	0.460	0.920	0.867	0.080	0.893	0.840	0.813
45	5	25	25	0.450	0.900	0.833	0.100	0.867	0.800	0.767
44	6	26	24	0.440	0.880	0.800	0.120	0.840	0.760	0.720
43	7	27	23	0.430	0.860	0.767	0.140	0.813	0.720	0.673
42	8	28	22	0.420	0.840	0.733	0.160	0.787	0.680	0.627
41	9	29	21	0.410	0.820	0.700	0.180	0.760	0.640	0.580
40	10	30	20	0.400	0.800	0.667	0.200	0.733	0.600	0.533
39	11	31	19	0.390	0.780	0.633	0.220	0.707	0.560	0.487
38	12	32	18	0.380	0.760	0.600	0.240	0.680	0.520	0.440
37	13	33	17	0.370	0.740	0.567	0.260	0.653	0.480	0.393
36	14	34	16	0.360	0.720	0.533	0.280	0.627	0.440	0.347
35	15	35	15	0.350	0.700	0.500	0.300	0.600	0.400	0.300
34	16	36	14	0.340	0.680	0.467	0.320	0.573	0.360	0.253
33	17	37	13	0.330	0.660	0.433	0.340	0.547	0.320	0.207
32	18	38	12	0.320	0.640	0.400	0.360	0.520	0.280	0.160
31	19	39	11	0.310	0.620	0.367	0.380	0.493	0.240	0.113
30	20	40	10	0.300	0.600	0.333	0.400	0.467	0.200	0.067
29	21	41	9	0.290	0.580	0.300	0.420	0.440	0.160	0.020
28	22	42	8	0.280	0.560	0.267	0.440	0.413	0.120	-0.027
27	23	43	7	0.270	0.540	0.233	0.460	0.387	0.080	-0.073
26	24	44	6	0.260	0.520	0.200	0.480	0.360	0.040	-0.120
25	25	45	5	0.250	0.500	0.167	0.500	0.333	0.000	-0.167
24	26	46	4	0.240	0.480	0.133	0.520	0.307	-0.040	-0.213
23	27	47	3	0.230	0.460	0.100	0.540	0.280	-0.080	-0.260
22	28	48	2	0.220	0.440	0.067	0.560	0.253	-0.120	-0.307
21	29	49	1	0.210	0.420	0.033	0.580	0.227	-0.160	-0.353
20	30	50	0	0.200	0.400	0.000	0.600	0.200	-0.200	-0.400

4. 결론

빅 데이터 분석 기술 중의 하나인 연관성 규칙 기법은 탐색적이고 비목적성 분석이며 기존의 데이터를 특별히 변환하지 않고도 계산에 용이하게 사용 가능하다는 장점을 가지고 있다. 일반적인 연관성 규칙은 사용자가 지정한 최소 지지도를 만족하는 빈발항목집합을 생성한 후 최저 신뢰도를 만족하는 규칙을 연관성 규칙으로 생성한다. 이 때 사용되는 신뢰도는 그 값의 크기로는 양의 연관성이 있는지, 아니면 음의 연관성이 있는지 연관성의 방향을 파악할 수 없다. 본 논문에서는 신뢰도의 이러한 단점을 보완하기 위해 연관성 평가 기준의 관점에서 분류 모형 구축에 유용한 여러 가지 신뢰도를 비교 고찰하였다. 동시 발생빈도 및 동시 비발생빈도가 증가함에 따라 $P(Y^c|X)$을 제외하고는 모든 평가 기준의 값들이 증가하고 있으며, 인과적 신뢰도 $conf_{CA}$는 양의 값을 갖는 반면에 확인적 신뢰도 $conf_{CO}$와 인과적 확인 신뢰도 $conf_{CC}$는 양과 음의 값을 갖는다. 또한 두 가지 형태의 불일치빈도가 증가함에 따라 $P(Y^c|X)$을 제외하고는 모든 평가 기준의 값들이 감소하였으며, $conf_{CA}$는 양의 값을 갖는 반면에 $conf_{CO}$와 $conf_{CC}$는 양과 음의 값을 갖는다. 따라서 연관성 규칙의 평가 기준 관점에서는 $conf_{CA}$보

다는 연관성의 방향을 알 수 있는 $conf_{CO}$와 $conf_{CC}$가 더 바람직한 측도라고 할 수 있다. $conf_{CO}$와 $conf_{CC}$중에서는 후자인 인과적 확인 신뢰도 $conf_{CC}$가 양의 신뢰도 크기와 음의 신뢰도 크기의 영향을 좀 더 합리적으로 나타내주므로 더 바람직한 측도라고 할 수 있다. 이 측도는 백화점이나 온라인 쇼핑몰 등에서의 시장바구니 분석이나 개인화 추천 서비스, 그리고 교차 판매 분석 등에서 연관성 규칙의 방향을 고려하는 경우에 활용 가능하다.

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.

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 Datos y Aprendizaje, TAMIDA2005*, 135-144.

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.

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.

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. (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. (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.

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.

Journal of the Korean Data & Information Science Society
2014, **25**(2), 365–371

http://dx.doi.org/10.7465/jkdi.2014.25.2.365
한국데이터정보과학회지

Comparison of confidence measures useful for classification model building

Hee Chang Park[1]

[1]Department of Statistics, Changwon National University

Received 4 February 2014, revised 24 February 2014, accepted 3 March 2014

Abstract

Association rule of the well-studied techniques in data mining is the exploratory data analysis for understanding the relevance among the items in a huge database. This method has been used to find the relationship between each set of items based on the interestingness measures such as support, confidence, lift, similarity measures, etc. By typical association rule technique, we generate association rule that satisfy minimum support and confidence values. Support and confidence are the most frequently used, but they have the drawback that they can not determine the direction of the association because they have always positive values. In this paper, we compared support, basic confidence, and three kinds of confidence measures useful for classification model building to overcome this problem. The result confirmed that the causal confirmed confidence was the best confidence in view of the association mining because it showed more precisely the direction of association.

Keywords: Association rule, causal confidence, causal confirmed confidence, confirmed confidence, data mining.

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