Journal of the Korean Data & Information Science Society
2011, **22**(6), 1113–1121

한국데이터정보과학회지

동시 비 발생 빈도를 고려한 유사성 측도의 연관성 규칙 평가 기준 활용 방안

박희창[1]

[1]창원대학교 통계학과

접수 2011년 10월 14일, 수정 2011년 11월 2일, 게재확정 2011년 11월 8일

요 약

최근 여러 분야에서 다양한 데이터 마이닝 방법들을 현업에 적용하고 있는 추세이다. 가장 많이 활용되고 있는 데이터 마이닝 기법 중의 하나인 연관성 규칙은 대용량 데이터베이스에 내재되어 있는 항목들 간의 관련성을 수치화하여 그들 간의 연관 정도를 나타내는 기법이다. 의미 있는 연관성 규칙을 생성하기 위해 지지도, 신뢰도, 향상도 등의 측도가 가장 기본적으로 활용되고 있다. 본 논문에서는 군집 분석이나 다차원 분석법에서 많이 활용되고 있는 유사성 측도들 중에서 동시 비 발생 빈도를 고려한 유사성 측도를 연관성 평가 기준으로 제안한 후, 예제를 통하여 기존의 신뢰도 및 지지도와 비교함으로써 그 유용성을 알아보았다. 모의실험 결과를 종합해볼 때, 동시 발생 빈도 또는 동시 비 발생 빈도가 증가하면 본 논문에서 고려한 모든 유사성 측도들은 지지도 및 신뢰도와 마찬가지로 증가하며, 불일치 계수의 값이 증가하면 이 측도들은 감소하게 된다는 사실을 알 수 있었다. 또한 이들 유사성 측도들은 지지도 및 신뢰도와 매우 유의한 상관관계가 있는 것으로 나타났으며, 전항과 후항이 바뀌더라도 값의 변화가 없기 때문에 신뢰도 보다 더 바람직한 연관성 규칙 평가 기준이라고 할 수 있다.

주요용어: 동시 비 발생 빈도, 신뢰도, 연관성 규칙, 유사성 측도, 지지도.

1. 서론

데이터 마이닝 기법 중의 하나인 연관규칙 탐사 (association rule discovery) 방법은 대용량 데이터베이스에 내재되어 있는 항목들 간의 관련성을 찾아내는 데 활용되고 있으며, 항목들 간의 관계를 수치화하여 이들 간의 관련성을 표시함으로써 제조업, 유통업, 그리고 의료분야 등 여러 분야에서 다양하게 적용되고 있다. Agrawal 등 (1993)에 의해 처음 소개된 연관성 규칙은 이후 국내외적으로 많은 학자들에 의해 연구되어 왔으며, 지금도 활발한 연구가 진행되고 있다 (Agrawal과 Srikant, 1994; Park 등, 1995; Toivonen, 1996; Cai 등, 1998; Han과 Fu, 1999; Liu 등, 1999; Pasquier 등, 1999; Han 등, 2000; Pei 등, 2000; Choi와 Park, 2008; Cho와 Park, 2008; Park, 2010a, 2010b, 2010c; Park, 2011a).

연관성 규칙을 평가하기 위해 가장 기본적으로 활용하는 연관성 측도에는 지지도 (support), 신뢰도 (confidence), 향상도 (lift) 등이 있으며, 연관성 규칙이 갖는 의의는 정성적인 의미와 정량적인 의미를 동시에 해석할 수 있다는 데 있다 (Srikant와 Agrawal, 1995). 연관성 규칙 생성 시, 먼저 분석가가 지정한 최소 지지도를 만족시키는 빈발항목집합을 생성한 다음, 이들에 대해 최저신뢰도 기준을 만족하고 향상도가 1이상인 것을 연관성 규칙으로 채택하게 된다. 이 때 사용되는 지지도와 향상도는 전항과 후항이 바뀌더라도 동일한 값을 가지므로 이들 측도들은 대칭적이라고 할 수 있으나, 신뢰도는 전

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

항과 후항이 바뀌게 되면 그 값이 달라지므로 대칭적 측도라고 할 수 없다. 또한 신뢰도는 전항과 후항을 다르게 하여 계산한 두 값이 모두 최저 기준을 만족하게 되면 연관성 규칙이 생성된 것으로 볼 수 있다. 반면에 두 값 중에서 어느 한 값이 신뢰도의 최저 기준을 만족하지 않는 경우에는 연관성 규칙이 생성된 것으로 보기가 곤란하다. 이러한 문제를 해결하기 위해 Park (2011b)은 유사성 측도들을 연관성 평가 기준으로 제안한 바 있다. 그러나 이 연구에서는 동시 비 발생 빈도 (negative co-occurrence frequencies)를 고려하지 않은 유사성 측도들에 대해 고찰하였다. 동시 비 발생 빈도는 동시발생빈도와 더불어 두 항목간의 관련성에 대한 순방향성을 의미하고 있으므로 이도 함께 고려하는 것이 바람직할 것으로 생각되어 본 논문에서는 다차원 분석이나 군집분석에서 이용되고 있는 동시 비 발생 빈도를 고려한 유사성 측도들을 연관성 평가 기준으로 제안하고자 한다. 먼저 제 2절에서는 본 논문에서 고려하는 동시 비 발생 빈도를 고려한 유사성 측도들을 소개한 후, 이들과 기본적인 연관성 평가 기준인 지지도와 신뢰도와의 관계를 탐색하고자 한다. 제 3절에서는 구체적인 예제를 통하여 기존의 지지도 및 신뢰도와 본 논문에서 제안한 유사성 측도들을 비교해봄으로써, 그 유용성을 살펴본 후, 결론을 맺고자 한다.

2. 동시 비 발생 빈도를 고려한 유사성 측도

본 논문에서는 동시 비 발생 확률을 고려한 유사성 측도들에 대해 연관성 평가기준으로서의 적용가능성을 평가하고자 한다. 먼저 동시 비 발생 빈도를 고려한 유사성 측도를 수식으로 나타내기 위해 다음과 같은 분할표를 고려한다.

표 2.1 2×2분할표

		B		합계
		1	0	
A	1	a	b	$a+b$
	0	c	d	$c+d$
합계		$a+c$	$b+d$	n

Meyer (2002)에 의하면 동시 비 발생 빈도를 고려한 유사성 측도에는 Sokal과 Michener의 단순매칭측도 (simple matching measure), Russel과 Rao 측도, Rogers와 Tanimoto 측도, Hamann 측도, Sokal과 Sneath 측도, 그리고 Ochiai II 측도 등이 있다. 이들 중에서 Russel과 Rao가 제안한 유사성 측도는 지지도와 동일하므로 고찰 대상에서 제외하고, 그 이외의 유사성 측도들을 표 2.1의 기호를 사용하여 수식으로 나타내면 다음과 같다.

$$\text{Simple matching} : Sokm(A,B) = \frac{a+d}{a+b+c+d} \tag{2.1}$$

$$\text{Rogers과 Tanimoto} : Rog(A,B) = \frac{a+d}{a+d+2(b+c)} \tag{2.2}$$

$$\text{Hamann} : Ham(A,B) = \frac{(a+d)-(b+c)}{a+b+c+d} \tag{2.3}$$

$$\text{Sokal과 Sneath} : Soks(A,B) = \frac{2(a+d)}{2(a+d)+b+c} \tag{2.4}$$

$$\text{Ochiai II} : Och_2(A,B) = \frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} \tag{2.5}$$

본 논문에서 제시한 동시 비 발생 빈도를 고려한 유사성 측도들에 대해 지지도 및 신뢰도와의 관계를 파악할 수 있는데 신뢰도의 분자가 지지도를 의미하므로 신뢰도에 대한 이들 유사성 측도와의 관계식을 구하면 다음과 같다. 여기서 $C(A,B)$는 신뢰도 $P(B|A)$를 의미한다.

$$Sokm(A,B) = 1 - P(B) - P(A)[1 - 2C(A,B)] \tag{2.6}$$

$$Rog(A,B) = \frac{1 - P(B) - P(A)[1 - C(A,B)]}{1 + P(B) + P(A)[1 - 2C(A,B)]} \tag{2.7}$$

$$Ham(A,B) = 1 - 2P(B) - 2P(A)[1 - 2C(A,B)] \tag{2.8}$$

$$Soks(A,B) = \frac{2[1 - P(B) - P(A)\{1 - 2C(A,B)\}]}{2 - P(B) - P(A)[1 - 2C(A,B)]} \tag{2.9}$$

$$Och_2(A,B) = \frac{P(A)C(A,B)[1 - P(B) - P(A)\{1 - C(A,B)\}]}{\sqrt{P(A)[1 - P(A)]P(B)[1 - P(B)]}} \tag{2.10}$$

위의 식들로부터 알 수 있는 바와 같이 신뢰도가 증가함에 따라 본 논문에서 고려하는 동시 비 발생 빈도를 고려한 유사성 측도들 모두가 증가하게 된다. 또한 수식을 관찰해 본 결과, 이들은 전항과 후항이 바뀌더라도 값의 변화가 없기 때문에 신뢰도 보다 더 바람직한 연관성 규칙 평가 기준이라고 할 수 있다. 특히 항목 A와 B가 독립 관계이면 Hamann이 제안한 유사성 측도 $Ham(A,B)$는 다음의 값을 취하게 된다.

$$Ham(A,B) = [1 - 2P(A)][1 - 2P(B)]$$

이로부터 알 수 있는 사실은 $P(A)$또는 $P(B)$가 0.5이면 $Ham(A,B)$은 0이 되며, 두 값 모두 0.5보다 작거나 크면 $Ham(A,B)$은 양의 값을 취하나 이들 중 하나만 0.5보다 작으면 음의 값을 취하게 된다는 것이다.

3. 예제를 통한 고찰

본 절에서는 동시 비 발생 빈도를 고려한 유사성 측도들의 유용성에 대해 예제를 통하여 탐색하고자 한다. 이를 위해 항목 집합 X, Y에 대해 다음과 같이 가정하였다. 먼저 데이터베이스에 있는 총 트랜잭션의 수 (t)를 100명으로 하고, 항목 집합 X는 구매한 물품의 금액을 기준으로 특정금액 이상 (1) 구매한 사람 수와 특정금액 미만 (0)을 구매한 사람 수를 각각 50명으로 하였다. 또한 항목 집합 Y를 결제 방식을 기준으로 특정 방법 (예 : 신용카드)으로 결제 (1)한 사람 수를 30명으로 하고 그 외의 방법으로 결제 (0)한 사람의 수를 70명으로 하였다. 항목 집합 X와 Y가 동시에 발생한 빈도 수, 즉 특정금액 이상의 물품을 구매하면서 특정방법으로 결제한 빈도수는 a명으로 하였다. 이를 정리하면 표 3.1과 같다.

표 3.1 모의실험 데이터(1)

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

이 표에서 a가 취할 수 있는 정수 값의 범위는 $0 \le a \le 30$이다. 표 3.1을 이용하여 동시발생빈도의 변화에 따라 계산된 유사성 측도들과 지지도 및 신뢰도를 계산하여 그 일부를 나타내면 표 3.2와 같다.

표 3.2 모의실험 데이터(1)에 의한 연관성 규칙 기준의 변화

a	b	c	d	supp	conf	conf2	Sokm	Rog	Ham	Och2	Soks
1	49	29	21	0.010	0.020	0.033	0.220	0.124	-0.560	0.010	0.361
2	48	28	22	0.020	0.040	0.067	0.240	0.136	-0.520	0.010	0.387
3	47	27	23	0.030	0.060	0.100	0.260	0.149	-0.480	0.011	0.413
4	46	26	24	0.040	0.080	0.133	0.280	0.163	-0.440	0.012	0.438
5	45	25	25	0.050	0.100	0.167	0.300	0.176	-0.400	0.013	0.462
6	44	24	26	0.060	0.120	0.200	0.320	0.190	-0.360	0.014	0.485
7	43	23	27	0.070	0.140	0.233	0.340	0.205	-0.320	0.015	0.507
8	42	22	28	0.080	0.160	0.267	0.360	0.220	-0.280	0.016	0.529
9	41	21	29	0.090	0.180	0.300	0.380	0.235	-0.240	0.017	0.551
10	40	20	30	0.100	0.200	0.333	0.400	0.250	-0.200	0.017	0.571
11	39	19	31	0.110	0.220	0.367	0.420	0.266	-0.160	0.018	0.592
12	38	18	32	0.120	0.240	0.400	0.440	0.282	-0.120	0.019	0.611
13	37	17	33	0.130	0.260	0.433	0.460	0.299	-0.080	0.020	0.630
14	36	16	34	0.140	0.280	0.467	0.480	0.316	-0.040	0.021	0.649
15	35	15	35	0.150	0.300	0.500	0.500	0.333	0.000	0.022	0.667
16	34	14	36	0.160	0.320	0.533	0.520	0.351	0.040	0.023	0.684
17	33	13	37	0.170	0.340	0.567	0.540	0.370	0.080	0.024	0.701

여기서 $b = n(X \text{ and } Y^c)$, $c = n(X^c \text{ and } Y)$, $d = n(X^c \text{ and } Y^c)$를 의미하며, $conf_2$는 $P(X|Y)$를 의미한다. 이 표에서 보는 바와 같이 동시발생빈도 a의 값이 증가함에 따라, 또한 동시 비 발생 빈도 d의 값이 증가함에 따라 지지도와 신뢰도가 증가하고 있으며, 본 논문에서 고려하는 유사성 측도 모두가 증가하고 있다. 또한 유사성 측도 중에서는 Ham을 제외한 모든 유사성 측도들은 모두 0과 1 사이의 값을 가지며, $Soks$가 가장 큰 값으로 나타나고 있고, 측도 Och_2가 가장 작은 값으로 나타나고 있다. Ham측도는 -1과 1 사이의 값을 갖는다.

지지도 및 신뢰도와 동시 비 발생 빈도를 고려한 유사성 측도들 간의 상관계수를 구하면 표 3.3과 같다. 여기서 **는 0.01 수준에서 상관계수의 값이 유의함을 의미한다.

표 3.3 모의실험 데이터(1)에 의한 연관성 평가 기준들 간의 상관 분석

	supp	conf	conf2	Sokm	Rog	Ham	Och2
conf	1.000**						
conf2	1.000**	1.000**					
Sokm	1.000**	1.000**	1.000**				
Rog	.995**	.995**	.995**	.995**			
Ham	1.000**	1.000**	1.000**	1.000**	.995**		
Och2	1.000**	1.000**	1.000**	1.000**	.995**	1.000**	
Soks	.995**	.995**	.995**	.995**	.980**	.995**	.995**

지지도 및 신뢰도와 유사성 측도들을 비교해보면 Rog와 $Soks$를 제외하고는 지지도 및 신뢰도와 완전 상관을 나타내고 있으나, 이들 측도마저도 상관관계가 매우 유의한 것으로 나타났다. 또한 이들은 전항과 후항이 바뀌더라도 값의 변화가 없기 때문에 신뢰도 보다 더 바람직한 연관성 규칙 평가 기준이라고 할 수 있다.

이번에는 불일치 빈도 b가 변화함에 따라 동시 비 발생 빈도를 고려한 유사성 측도들의 변화하는 양상을 살펴보기 위해 표 3.4를 이용한 결과를 표 3.5에 나타내었다.

표 3.4 모의실험 데이터(2)

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

표 3.5 모의실험 데이터(2)에 의한 연관성 규칙 기준의 변화

a	b	c	d	supp	conf	conf2	Sokm	Rog	Ham	Och2	Soks
17	13	33	37	0.170	0.567	0.340	0.540	0.370	0.080	0.275	0.701
16	14	34	36	0.160	0.533	0.320	0.520	0.351	0.040	0.251	0.684
15	15	35	35	0.150	0.500	0.300	0.500	0.333	0.000	0.229	0.667
14	16	36	34	0.140	0.467	0.280	0.480	0.316	-0.040	0.208	0.649
13	17	37	33	0.130	0.433	0.260	0.460	0.299	-0.080	0.187	0.630
12	18	38	32	0.120	0.400	0.240	0.440	0.282	-0.120	0.168	0.611
11	19	39	31	0.110	0.367	0.220	0.420	0.266	-0.160	0.149	0.592
10	20	40	30	0.100	0.333	0.200	0.400	0.250	-0.200	0.131	0.571
9	21	41	29	0.090	0.300	0.180	0.380	0.235	-0.240	0.114	0.551
8	22	42	28	0.080	0.267	0.160	0.360	0.220	-0.280	0.098	0.529
7	23	43	27	0.070	0.233	0.140	0.340	0.205	-0.320	0.082	0.507
6	24	44	26	0.060	0.200	0.120	0.320	0.190	-0.360	0.068	0.485
5	25	45	25	0.050	0.167	0.100	0.300	0.176	-0.400	0.055	0.462
4	26	46	24	0.040	0.133	0.080	0.280	0.163	-0.440	0.042	0.438
3	27	47	23	0.030	0.100	0.060	0.260	0.149	-0.480	0.030	0.413
2	28	48	22	0.020	0.067	0.040	0.240	0.136	-0.520	0.019	0.387
1	29	49	21	0.010	0.033	0.020	0.220	0.124	-0.560	0.009	0.361
0	30	50	20	0.000	0.000	0.000	0.200	0.111	-0.600	0.000	0.333

이 표에서 보는 바와 같이 불일치 빈도 b가 증가함에 따라 지지도 및 신뢰도가 감소하고 있으며, 본 논문에서 고려하는 동시 비 발생 빈도를 고려한 유사성 측도들도 모두 감소하는 것으로 나타났다. 여기서도 Ham측도는 -1과 1 사이의 값을 갖는다.

동시 비 발생 빈도를 고려한 유사성 측도들과 지지도 및 신뢰도 간의 상관계수를 구하면 표 3.6과 같다. 여기서 **는 0.01 수준에서 상관계수의 값이 유의함을 의미한다. 이 표에서 보는 바와 같이 지지도 및 신뢰도와 유사성 측도들 간에는 매우 유의한 상관관계가 있는 것으로 나타났다.

표 3.6 모의실험 데이터(2)에 의한 연관성 평가 기준들 간의 상관 분석

	supp	conf	conf2	Sokm	Rog	Ham	Och2
conf	1.000**						
conf2	1.000**	1.000**					
Sokm	1.000**	1.000**	1.000**				
Rog	.994**	.994**	.994**	.994**			
Ham	1.000**	1.000**	1.000**	1.000**	.994**		
Och2	.987**	.987**	.987**	.987**	.999**	.987**	
Soks	.994**	.994**	.994**	.994**	.977**	.994**	.965**

이번에는 불일치 빈도 c가 변화함에 따라 동시 비 발생 빈도를 고려한 유사성 측도들의 변화하는 양상을 살펴보기 위해 표 3.7을 이용하여 계산된 결과를 표 3.8에 나타내었다.

표 3.7 모의실험 데이터(3)

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

표 3.8 모의실험 데이터(3)에 의한 연관성 규칙 기준의 변화

a	b	c	d	supp	conf	conf2	Sokm	Rog	Ham	Och2	Soks
42	28	8	22	0.420	0.600	0.840	0.640	0.471	0.280	0.403	0.780
41	29	9	21	0.410	0.586	0.820	0.620	0.449	0.240	0.376	0.765
40	30	10	20	0.400	0.571	0.800	0.600	0.429	0.200	0.349	0.750
39	31	11	19	0.390	0.557	0.780	0.580	0.408	0.160	0.323	0.734
38	32	12	18	0.380	0.543	0.760	0.560	0.389	0.120	0.299	0.718
37	33	13	17	0.370	0.529	0.740	0.540	0.370	0.080	0.275	0.701
36	34	14	16	0.360	0.514	0.720	0.520	0.351	0.040	0.251	0.684
35	35	15	15	0.350	0.500	0.700	0.500	0.333	0.000	0.229	0.667
34	36	16	14	0.340	0.486	0.680	0.480	0.316	-0.040	0.208	0.649
33	37	17	13	0.330	0.471	0.660	0.460	0.299	-0.080	0.187	0.630
32	38	18	12	0.320	0.457	0.640	0.440	0.282	-0.120	0.168	0.611
31	39	19	11	0.310	0.443	0.620	0.420	0.266	-0.160	0.149	0.592
30	40	20	10	0.300	0.429	0.600	0.400	0.250	-0.200	0.131	0.571
29	41	21	9	0.290	0.414	0.580	0.380	0.235	-0.240	0.114	0.551
28	42	22	8	0.280	0.400	0.560	0.360	0.220	-0.280	0.098	0.529
27	43	23	7	0.270	0.386	0.540	0.340	0.205	-0.320	0.082	0.507

위의 불일치 빈도 b의 결과와 마찬가지로 불일치 빈도 c가 증가함에 따라 지지도 및 신뢰도가 감소하고 있으며, 본 논문에서 고려하는 동시 비 발생 빈도를 고려한 유사성 측도들도 모두 감소하는 것으로 나타났다. 동시 비 발생 빈도를 고려한 유사성 측도들과 지지도 및 신뢰도 간의 상관계수를 구하면 표 3.9와 같다. 여기서도 지지도 및 신뢰도와 유사성 측도들 간에는 매우 유의한 상관관계가 있는 것으로 나타났다.

표 3.9 모의실험 데이터(3)에 의한 연관성 평가 기준들 간의 상관 분석

	supp	conf	conf2	Sokm	Rog	Ham	Och2
conf	1.000**						
conf2	1.000**	1.000**					
Sokm	1.000**	1.000**	1.000**				
Rog	.994**	.994**	.994**	.994**			
Ham	1.000**	1.000**	1.000**	1.000**	.994**		
Och2	.987**	.987**	.987**	.987**	.999**	.987**	
Soks	.994**	.994**	.994**	.994**	.977**	.994**	.965**

마지막으로 동시 비 발생 빈도 d가 변화함에 따라 본 논문에서 고려한 유사성 측도들의 변화하는 양상을 살펴보았는데, 동시 비 발생 빈도 d가 증가함에 따라 지지도 및 신뢰도가 증가하고 있으며, 본 논문에서 고려하는 동시 비 발생 빈도를 고려한 유사성 측도들도 모두 증가하는 것으로 나타났다. 또한 동시 비 발생 빈도를 고려한 유사성 측도들과 지지도 및 신뢰도 간의 상관계수를 구해본 결과, 지지도 및 신뢰도와 유사성 측도들 간에는 매우 유의한 상관관계가 있는 것으로 나타났다.

4. 결론

연관성 규칙 기법이 적용되는 데이터는 거래발생시점에서 기록된 항목에 관한 정보만으로 구성된다. 이러한 기법을 적용하여 의미 있는 연관성 규칙을 생성하기 위해서는 신뢰도가 가장 많이 활용되고 있다. 그런데 전항과 후항이 바뀌게 되면 신뢰도의 값이 달라지므로 대칭적인 측도라고 말할 수 없어서 전항과 후항을 다르게 하여 계산한 두 신뢰도 값 중에서 어느 한 값이 신뢰도의 최저 기준을 만족하지 않는 경우에는 연관성 규칙이 생성된 것으로 보기가 곤란하다.

이러한 문제를 해결하기 위해 본 논문에서는 동시 비 발생 빈도를 고려한 유사성 측도들을 연관성 평가 기준으로 제안하였다. 또한 이들 측도들과 기본적인 연관성 평가 기준인 지지도와 신뢰도와의 관계를 탐색하였으며, 구체적인 예제를 통하여 기존의 지지도 및 신뢰도와 본 논문에서 제안한 유사성 측도들과의 비교해보았다. 그 결과, 동시 발생 빈도 또는 동시 비 발생 빈도가 증가하면 위에서 고려한 모든 유사성 측도들은 지지도 및 신뢰도와 마찬가지로 증가하며, 불일치 계수의 값이 증가하면 본 논문에서 고려한 모든 측도들은 감소한다. 특히 Ham측도는 방향성을 갖게 되는 유사성 측도이므로 다른 유사성 측도들에 비해 더 바람직한 측도인 것으로 나타났다. 그리고 유사성 측도들은 지지도 및 신뢰도와 매우 유의한 상관관계가 있는 것으로 나타났다. 수식을 관찰해 본 결과, 이들은 전항과 후항이 바뀌더라도 값의 변화가 없기 때문에 신뢰도 보다 더 바람직한 연관성 규칙 평가 기준이라고 할 수 있다.

추후 연구로 동시 비 발생 빈도를 고려한 측도들과 이를 고려하지 않은 측도들을 비교하는 연구가 필요하다.

참고문헌

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.

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. (2008). A study of association rule application using self-organizing map for fused data. *Journal of the Korean Data & Information Science Society*, **19**, 95-104.

Choi, J. H. and Park, H. C. (2008). Comparative study of quantitative data binning methods in association rule. *Journal of the Korean Data & Information Science Society*, **19**, 903-910.

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.

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.

Meyer A. (2002). *Comparison of similarity coefficients used in cluster analysis with dominant markers data,* MSc thesis, Universidade de Sao Paulo, Piracicaba.

Park, H. C. (2010a). Weighted association rules considering item RFM scores. *Journal of the Korean Data & Information Science Society*, **21**, 1147-1154.

Park, H. C. (2010b). Standardization for basic association measures in association rule mining. *Journal of the Korean Data & Information Science Society*, **21**, 891-899.

Park, H. C. (2010c). Decision process for right association rule generation. *Journal of the Korean Data & Information Science Society*, **21**, 263-270.

Park, H. C. (2011a). The application for predictive similarity measures of binary data in association rule mining. *Journal of the Korean Data & Information Science Society*, **22**, 495-503.

Park, H. C. (2011b). The application of some similarity measures to association rule thresholds. *Journal of the Korean Data Analysis Society*, **13**, 1331-1342.

Park, J. S., Chen, M. S. and Philip, S. Y. (1995). An effective hash-based algorithms for mining association rules. *Proceedings of ACM SIGMOD Conference on Management of Data*, 175-186.

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.

Srikant, R. and Agrawal, R. (1995). Mining generalized association rules. *Proceedings of the 21st VLDB Conference*, 407-419.

Toivonen H. (1996). Sampling large database for association rules. *Proceedings of the 22nd VLDB Conference*, 134-145.

Journal of the Korean Data & Information Science Society
2011, **22**(6), 1113–1121

한국데이터정보과학회지

Association rule thresholds of similarity measures considering negative co-occurrence frequencies

Hee-Chang Park[1]

[1]Department of Statistics, Changwon National University

Received 14 October 2011, revised 2 November 2011, accepted 8 November 2011

Abstract

Recently, a variety of data mining techniques has been applied in various fields like healthcare, insurance, and internet shopping mall. Association rule mining is a popular and well researched method for discovering interesting relations among large set of data items. Association rule mining is the method to quantify the relationship between each set of items in very huge database based on the association thresholds. There are three primary quality measures for association rules; support and confidence and lift. In this paper we consider some similarity measures with negative co-occurrence frequencies which is widely used in cluster analysis or multi-dimensional analysis as association thresholds. The comparative studies with support, confidence and some similarity measures are shown by numerical example.

Keywords: Association rule, confidence, negative co-occurrence frequency, support.

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