Journal of the Korean Data & Information Science Society
2013, **24**(1), 117–124

http://dx.doi.org/10.7465/jkdi.2013.24.1.117
한국데이터정보과학회지

모든 주변 비율을 고려한 확률적 흥미도 측도 기반 유사성 측도의 연관성 평가 기준 활용 방안

박희창[1]

[1]창원대학교 통계학과

접수 2012년 12월 10일, 수정 2012년 12월 28일, 게재확정 2013년 1월 7일

요 약

연관성 규칙 탐사는 상당한 양의 데이터베이스에 내재되어 있는 항목들 간의 관련성을 파악하는 것으로 쇼핑몰, 보건 및 의료, 교육분야 등의 현장에서 많이 적용되고 있다. 이러한 연관성 규칙을 생성하기 위해 연관성 규칙 평가 기준인 지지도, 신뢰도, 향상도 등이 활용되고 있다. 이들 중에서 신뢰도가 연관성 평가 기준으로 가장 많이 활용되고는 있으나 항상 양의 값을 취하는 비대칭적 측도이기 때문에 항목 간에 연관성 규칙을 생성하는 데 어려움이 존재하게 된다. 이러한 문제를 해결하기 위해 본 논문에서는 주변 비율 전부를 포함한 확률적 흥미도 기반 유사성 측도를 연관성 평가 기준으로 활용하는 방안을 고려하였다. 이 측도들은 주변비율 전부와 교차표의 모든 항을 고려하여 연관성의 강도를 측정하는 측도이므로 나타나는 모든 정보를 충실히 반영해주는 측도라고 할 수 있다. 모의실험을 통해 확인한 결과, 모든 주변 비율을 고려한 확률적 흥미도 기반 유사성 측도 대부분이 기존의 연관성 평가 기준과 마찬가지로 연관성의 정도를 파악할 수 있는 동시에 부호를 포함하고 있어서 연관성의 방향도 알 수 있었다.

주요용어: 신뢰도, 연관성 규칙, 유사성 측도, 주변 비율, 지지도, 확률적 흥미도 측도.

1. 서론

데이터 마이닝 기법들 중에서 가장 많이 연구되고 있는 연관성 규칙 마이닝 (association rule mining)은 매우 방대한 양의 데이터베이스에 내재되어 있는 항목들 간의 관련성을 탐색하는 데 활용되고 있으며, 유통업, 제조업, 보험업, 의료 및 교육 분야 등 많은 분야에 적용되고 있다 (Park, 2012b). 이러한 연관성 규칙 마이닝은 동시에 발생하는 여러 항목들을 생성된 규칙의 집합으로 나타냄으로써 항목들 간의 상호 연관성들을 쉽게 파악할 수 있는 정성적인 의미뿐만 아니라 생성된 규칙에 대하여 지지도 (support)와 신뢰도 (confidence), 그리고 향상도 (lift) 측도를 이용함으로써 정량적인 의미로도 해석이 가능하다 (Srikant와 Agrawal, 1995). Agrawal 등(1993)에 의해 처음으로 소개된 연관성 규칙 기법은 이후에도 많은 학자들이 연관성 규칙과 관련된 연구를 수행하였다. 특히 연관성 규칙과 관련한 최근의 국내 연구로는 Jin 등 (2011), Park (2010a, 2010b, 2011a, 2011b, 2011c, 2012a, 2012b, 2012c) 등이 있다.

일반적으로 연관성 규칙을 생성할 때 우선적으로 사용자가 지정한 최소 지지도의 조건을 만족하는 빈발항목집합을 생성한다. 그런 후, 향상도가 1이상인 것 중에서 최저 신뢰도의 조건을 만족하는 규칙을 연관성 규칙으로 채택하게 된다 (Park, 2011a). 이 때 규칙 생성 여부를 결정하기 위해 가장 많이 활

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

용되고 있는 신뢰도의 전항과 후항이 바뀌게 되면 신뢰도의 값이 달라지는 비대칭적 측도가 되는 동시에 항상 양의 값을 가진다. 따라서 신뢰도의 크기로는 연관성의 방향을 파악하기 어렵다. 이러한 문제를 해결하기 위해 본 논문에서는 Warrens (2008)에 나타나는 모든 주변비율 (all marginal proportion; AMP)을 고려한 확률적 흥미도 측도 (probabilistic interestingness measures; PIM) 기반 유사성 측도에 대해 연관성 평가 기준으로서의 적용 가능 여부를 탐색하고자 한다. Park (2012c)에서 기술한 바와 같이 PIM과 관련한 연구로는 Orchard (1975)이 수행한 바 있는 PIM을 이용한 Boolean Analyzer 알고리즘 개발과 Imberman 등 (2001)의 PIM에 의한 연관성 규칙 생성에 관한 연구, Park (2012b, c)의 주변비율이 없는 PIM 기반 유사성 측도와 부분주변비율을 고려한 PIM 기반 유사성 측도에 관한 연구 등이 있다. 논문의 2절에서는 AMP를 고려한 PIM 기반 유사성 측도들을 소개한다. 3절에서는 예제를 통하여 기존의 연관성 규칙 평가 기준과 본 논문에서 고려한 유사성 측도와의 비교를 통해 유사성 측도의 유용성을 살펴본 후, 4절에서 결론을 내리고자 한다.

2. AMP를 고려한 PIM 기반 유사성 측도

Park (2012b)에서와 마찬가지로 기존의 연관성 규칙의 평가기준인 지지도, 신뢰도, 향상도 등을 수식으로 나타내기 위해 Table 2.1과 같은 분할표를 고려하고자 한다.

Table 2.1 2×2 contingency table

		B		Total
		1	0	
A	1	n_{11}	n_{10}	n_1+
	0	n_{01}	n_{00}	n_0+
Total		n_+1	n_+0	n

지지도 $S(A \Rightarrow B)$는 항목 집합 A와 항목 집합 B가 동시에 발생하는 거래의 비율로 $S(A \Rightarrow B) = n_{11}/n$으로 계산된다. 신뢰도 $C(A \Rightarrow B)$는 항목 집합 A가 포함된 거래 비율 중 항목 집합 A와 항목 집합 B가 동시에 포함된 거래의 비율을 의미하며, n_{11}/n_{1+}이 된다. 신뢰도 $C(A \Rightarrow B)$에서 전항과 후항을 바꾸어 계산되는 신뢰도 $C(B \Rightarrow A)$는 항목 집합 B가 포함된 거래 비율 중 항목 집합 A와 항목 집합 B가 동시에 포함된 거래의 비율을 의미하며, n_{11}/n_{+1}이 된다. 마지막으로 향상도 $L(A \Rightarrow B)$는 항목 집합 A를 구매한 경우 그 거래가 항목 집합 A를 포함하는 경우와 항목 집합 B가 임의로 구매되는 경우의 비를 의미하며, $n_{11} \cdot n/(n_{1+} \cdot n_{+1})$로 계산된다. 확률적 흥미도 측도인 PIM은 $n_{11}n_{00} - n_{10}n_{01}$를 의미하며, 항목 A와 B의 관련성의 정도를 나타내는 측정치이다. 이 측도의 값이 양수이면 두 항목은 정(+)의 관련성을 나타내고, 음수이면 부(-)의 관련성을 나타낸다. 만약 그 값이 0이거나 0에 가까운 값이 되면 두 항목은 서로 독립이므로 아무런 연관관계가 없다고 할 수 있다.

본 논문에서는 주변 비율 전부를 고려한 PIM 기반 유사성 측도들에 대해 연관성 평가기준으로서의 적용가능성을 평가하고자 한다. 이들을 수식으로 나타내기 위해 Table 2.1의 각 항을 총도수로 나누어서 비율의 형태로 나타내면 Table 2.2와 같다.

Table 2.2 2×2 contingency table by proportions

		B		Total
		1	0	
A	1	a	b	p_1
	0	c	d	q_1
Total		p_2	q_2	1

일반적으로 가장 많이 알려져 있는 카이 제곱 통계량은 다음과 같다.

$$\chi^2 = \frac{n(ad-bc)^2}{p_1p_2q_1q_2}. \tag{2.1}$$

본 논문에서는 PIM 기반 유사성 측도 중에서 모든 주변비율을 포함하는 유사성 측도를 AMP를 고려한 PIM 기반 유사성 측도라고 한다. 특히 이들 통계량은 기존의 연관성 규칙 평가 기준과는 달리 주변비율 전부와 교차표의 모든 항을 고려하여 $ad-bc$의 값에 대한 크기를 이용하여 연관성의 강도를 측정하는 측도이다. Warrens (2008)이 기술한 바와 같이 AMP를 고려한 PIM 기반 유사성 측도에는 Kuder와 Richardson (1937), 그리고 Cronbach (1951)가 제안한 S_{KR}, Cohen (1960)이 제안한 S_{Cohen}, Stiles (1961)가 제안한 S_{Sti}, Maxwell과 Pilliner (1968)가 제안한 S_{MP}, Fleiss (1975)가 제안한 S_{Fleiss}등이 있다. Table 2.2를 이용하여 이들을 수식으로 나타낸 후, PIM을 이용해 변환하면 다음과 같다.

$$S_{Cohen} = \frac{2(ad-bc)}{p_1q_2+p_2q_1} = \frac{2PIM}{n^2(p_1q_2+p_2q_1)}. \tag{2.2}$$

$$S_{Sti} = \log_{10}\frac{n(|ad-bc|-n/2)^2}{p_1p_2q_1q_2} = \log_{10}\frac{(|PIM|-n^3/2)^2}{n^3p_1p_2q_1q_2}. \tag{2.3}$$

$$S_{Fleiss} = \frac{(ad-bc)(p_1q_2+p_2q_1)}{2p_1p_2q_1q_2} = \frac{PIM(p_1q_2+p_2q_1)}{2n^2p_1p_2q_1q_2}. \tag{2.4}$$

$$S_{MP} = \frac{2(ad-bc)}{p_1q_1+p_2q_2} = \frac{2PIM}{n^2(p_1q_1+p_2q_2)}. \tag{2.5}$$

$$S_{KR} = \frac{4(ad-bc)}{p_1q_1+p_2q_2+2(ad-bc)} = \frac{4PIM}{n^2(p_1q_1+p_2q_2)+2PIM}. \tag{2.6}$$

위의 식에서 보는 바와 같이 본 논문에서 고려하는 모든 유사성 측도는 p_1, q_1, p_2, q_2, 그리고 $ad-bc$의 값에 의해 그 값이 결정되며, 이들 모두 주변 비율 전부를 고려하는 측도인 동시에 PIM으로 표현이 가능하다. 이들 중에서 S_{Cohen}, S_{Fleiss}, 그리고 S_{MP}는 분자의 절대값 보다 분모의 값이 더 크고, 분자의 값의 부호에 따라 양과 음의 값을 가지므로 S_{Cohen}은 -1과 1 사이의 값을 갖는다. 반면에 S_{Sti}는 항상 양의 값을 취하며, S_{KR}의 분모를 재정리하면 $2ad+(a+d)(b+c)$가 되어 ad의 값이 bc에 비해 상당히 크면 1 보다 큰 값을 갖게 되고, 그 반대이면 -1보다 작은 음의 값을 갖는다. 한편, 수식 상으로는 측도 S_{Sti}가 음의 값을 취할 수는 있으나 현실적으로는 양의 값만을 가진다는 것을 수식의 분모 및 분자의 비교를 통해 확인할 수 있다.

본 논문에서 제시하고 있는 AMP를 고려한 PIM 기반 유사성 측도들에 대해 Piatetsky-Shapiro (1991)가 제안한 흥미도 측도의 세 가지 조건만족 여부를 조사해보면 먼저 식 (2.2)에서 식 (2.6)으로부터 $P(A\cap B)=P(A)P(B)$이면 각 식들의 분자 $ad-bc$가 0이 되므로 측도 S_{Sti}를 제외한 유사성 측도들 S_{KR}, S_{Cohen}, S_{MP}, 그리고 S_{Fleiss}는 0이 된다. 두 번째로 p_1이 증가한다는 것은 a또는 b가 증가한다는 의미이고, p_2가 증가한다는 것은 a또는 c가 증가한다는 의미이므로 p_1또는 p_2의 값이 증가함에 따라 측도 S_{KR}, S_{Cohen}, S_{MP}, 그리고 S_{Fleiss}가 모두 감소한다. 마지막으로 $P(A\cap B)$의 값이 증가한다는 의미는 a가 증가한다는 것이므로 식 (2.2)에서 식 (2.6)로부터 a가 증가하면 측도 S_{Sti}와 S_{KR}을 제외한 측도 S_{Cohen}, S_{MP}, 그리고 S_{Fleiss}가 증가함을 알 수 있다. 따라서 본 논문에서 고려하는 측도들 중에서 흥미도 측도의 조건을 모두 충족하는 측도는 S_{Cohen}, S_{MP}, 그리고 S_{Fleiss}임을 알 수 있다.

3. 모의실험

본 절에서는 기존의 연관성 규칙 평가기준들과의 비교를 통해 AMP를 고려한 PIM 기반 유사성 측도들의 유용성에 대해 알아보고자 한다. 이를 위해 Park (2012b)에서 사용된 예제와 동일하게 항목 집합 A, B를 가정하였다. 먼저 데이터베이스에 있는 총 트랜잭션의 수 (t)를 100명으로 하고, 항목 집합 A는 구매한 물품의 금액을 기준으로 특정금액 이상 (1) 구매한 사람 수와 특정금액 미만 (0)을 구매한 사람 수를 각각 50명으로 하였다. 또한 항목 집합 B를 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 A와 B가 동시에 발생한 빈도 수, 즉 특정금액 이상의 물품을 구매하면서 특정방법으로 결제한 빈도수는 h명으로 하였다. 이를 정리하면 Table 3.1과 같고, 여기서 h가 취할 수 있는 정수 값의 범위는 $0 \le h \le 30$이다.

Table 3.1 Simulation data(1)

		B		Total
		1	0	
A	1	h	$50-h$	50
	0	$30-h$	$h+20$	50
Total		30	70	100

이 모의실험 데이터를 동시발생비율의 변화에 따라 주변비율 전부를 고려한 유사성 측도들과 지지도 및 신뢰도를 미니탭 16을 이용하여 계산한 결과를 Table 3.2에 나타내었다.

Table 3.2 Variation of PIM based similarity measures with AMP by Table 3.1

a	b	c	d	$supp$	$conf$	$conf_2$	$lift$	ν	ν^2	χ^2	S_{KR}	S_{Cohen}	S_{Sti}	S_{MP}	S_{FLeiss}
0.01	0.49	0.29	0.21	0.010	0.020	0.033	0.04	-0.1400	0.0196	37.3333	-3.1111	-0.5600	6.6753	-0.6087	-0.6667
0.02	0.48	0.28	0.22	0.020	0.040	0.067	0.08	-0.1300	0.0169	32.1905	-2.6000	-0.5200	6.6755	-0.5652	-0.6190
0.03	0.47	0.27	0.23	0.030	0.060	0.100	0.12	-0.1200	0.0144	27.4286	-2.1818	-0.4800	6.6757	-0.5217	-0.5714
0.04	0.46	0.26	0.24	0.040	0.080	0.133	0.16	-0.1100	0.0121	23.0476	-1.8333	-0.4400	6.6759	-0.4783	-0.5238
0.05	0.45	0.25	0.25	0.050	0.100	0.167	0.20	-0.1000	0.0100	19.0476	-1.5385	-0.4000	6.6760	-0.4348	-0.4762
0.06	0.44	0.24	0.26	0.060	0.120	0.200	0.24	-0.0900	0.0081	15.4286	-1.2857	-0.3600	6.6762	-0.3913	-0.4286
0.07	0.43	0.23	0.27	0.070	0.140	0.233	0.28	-0.0800	0.0064	12.1905	-1.0667	-0.3200	6.6764	-0.3478	-0.3810
0.08	0.42	0.22	0.28	0.080	0.160	0.267	0.32	-0.0700	0.0049	9.3333	-0.8750	-0.2800	6.6766	-0.3043	-0.3333
0.09	0.41	0.21	0.29	0.090	0.180	0.300	0.36	-0.0600	0.0036	6.8571	-0.7059	-0.2400	6.6767	-0.2609	-0.2857
0.10	0.40	0.20	0.30	0.100	0.200	0.333	0.40	-0.0500	0.0025	4.7619	-0.5556	-0.2000	6.6769	-0.2174	-0.2381
0.11	0.39	0.19	0.31	0.110	0.220	0.367	0.44	-0.0400	0.0016	3.0476	-0.4211	-0.1600	6.6771	-0.1739	-0.1905
0.12	0.38	0.18	0.32	0.120	0.240	0.400	0.48	-0.0300	0.0009	1.7143	-0.3000	-0.1200	6.6773	-0.1304	-0.1429
0.13	0.37	0.17	0.33	0.130	0.260	0.433	0.52	-0.0200	0.0004	0.7619	-0.1905	-0.0800	6.6774	-0.0870	-0.0952
0.14	0.36	0.16	0.34	0.140	0.280	0.467	0.56	-0.0100	0.0001	0.1905	-0.0909	-0.0400	6.6776	-0.0435	-0.0476
0.15	0.35	0.15	0.35	0.150	0.300	0.500	0.60	0.0000	0.0000	0.0000	0.0000	0.0000	6.6778	0.0000	0.0000
0.16	0.34	0.14	0.36	0.160	0.320	0.533	0.64	0.0100	0.0001	0.1905	0.0833	0.0400	6.6776	0.0435	0.0476
0.17	0.33	0.13	0.37	0.170	0.340	0.567	0.68	0.0200	0.0004	0.7619	0.1600	0.0800	6.6774	0.0870	0.0952
0.18	0.32	0.12	0.38	0.180	0.360	0.600	0.72	0.0300	0.0009	1.7143	0.2308	0.1200	6.6773	0.1304	0.1429
0.19	0.31	0.11	0.39	0.190	0.380	0.633	0.76	0.0400	0.0016	3.0476	0.2963	0.1600	6.6771	0.1739	0.1905
0.20	0.30	0.10	0.40	0.200	0.400	0.667	0.80	0.0500	0.0025	4.7619	0.3571	0.2000	6.6769	0.2174	0.2381
0.21	0.29	0.09	0.41	0.210	0.420	0.700	0.84	0.0600	0.0036	6.8571	0.4138	0.2400	6.6767	0.2609	0.2857
0.22	0.28	0.08	0.42	0.220	0.440	0.733	0.88	0.0700	0.0049	9.3333	0.4667	0.2800	6.6766	0.3043	0.3333
0.23	0.27	0.07	0.43	0.230	0.460	0.767	0.92	0.0800	0.0064	12.1905	0.5161	0.3200	6.6764	0.3478	0.3810
0.24	0.26	0.06	0.44	0.240	0.480	0.800	0.96	0.0900	0.0081	15.4286	0.5625	0.3600	6.6762	0.3913	0.4286
0.25	0.25	0.05	0.45	0.250	0.500	0.833	1.00	0.1000	0.0100	19.0476	0.6061	0.4000	6.6760	0.4348	0.4762
0.26	0.24	0.04	0.46	0.260	0.520	0.867	1.04	0.1100	0.0121	23.0476	0.6471	0.4400	6.6759	0.4783	0.5238
0.27	0.23	0.03	0.47	0.270	0.540	0.900	1.08	0.1200	0.0144	27.4286	0.6857	0.4800	6.6757	0.5217	0.5714
0.28	0.22	0.02	0.48	0.280	0.560	0.933	1.12	0.1300	0.0169	32.1905	0.7222	0.5200	6.6755	0.5652	0.6190
0.29	0.21	0.01	0.49	0.290	0.580	0.967	1.16	0.1400	0.0196	37.3333	0.7568	0.5600	6.6753	0.6087	0.6667

Table 3.2에서 $supp$는 지지도 $S(A \Rightarrow B)$, $conf$는 신뢰도 $C(A \Rightarrow B)$, $conf_2$는 $C(B \Rightarrow A)$, $lift$는 $L(A \Rightarrow B)$, 그리고 ν는 $ad-bc$를 의미한다. 이 표에서 보는 바와 같이 동시발생비율 a의 값이 증가함에 따라 지지도와 향상도, 그리고 신뢰도 $conf$및 $conf_2$가 증가하고 있으며, 본 논문에서 고려하는 유사성 측도들 중에서 S_{Sti}를 제외한 유사성 측도들 S_{KR}, S_{Cohen}, S_{MP}, 그리고 S_{Fleiss}는 증가하고 있다. 반면에 카이제곱 통계량 기반 유사성 측도인 χ^2와 S_{Sti}는 각각 ν의 제곱 및 절대값 형태로 식이 나

타나므로 감소하다가 증가하는 경향과 증가하다가 감소하는 경향을 나타내고 있다. 그리고 측도 χ^2와 S_{Sti}는 0과 1 사이의 값을 취하는 반면에, 유사성 측도 S_{KR}을 제외한 S_{Cohen}, S_{MP}, 그리고 S_{Fleiss}는 모두 -1과 1 사이의 값을 갖는다. 따라서 카이 제곱 통계량 기반 측도는 항상 양의 값만 나타날 뿐만 아니라 a가 증가함에 따라 기존의 연관성 평가 기준들이 모두 증가함에도 불구하고 감소하다가 증가하거나 증가하다가 감소하는 형태를 나타내고 있으므로 연관성 평가 기준으로는 바람직하지 못하다고 할 수 있다. 그러나 본 논문에서 고려하고 있는 유사성 측도 중에서 S_{KR}, S_{Cohen}, S_{MP}, 그리고 S_{Fleiss}는 a가 증가함에 따라 기존의 연관성 평가 기준들과 마찬가지로 모두 증가하는 형태를 나타내고 있을 뿐만 아니라 양, 음의 부호를 가지고 있으므로 연관성의 방향도 파악할 수 있어서 연관성 평가 기준으로 사용가능하다고 할 수 있다. 더구나 이들 측도는 기존의 연관성 평가기준에서 나타내지 못하는 연관성의 방향을 나타내고 있으므로 매우 바람직한 연관성 규칙 평가기준으로 고려할 수 있다. 한편, 본 논문에서 고려하는 유사성 측도는 주변 비율의 조합을 사용하기 때문에 어떤 조합을 고려하느냐에 따라 그 값이 달라진다. 이 표에서 주변 비율 p_1, p_2, q_1, q_2의 값이 각각 0.5, 0.3, 0.5, 0.7이고, 각 유사성 측도에서 고려하는 주변비율의 조합이 각각 $p_2q_1 = 0.15$, $p_2q_2 = 0.21$, $p_1q_1 = 0.25$, $p_1q_2 = 0.35$의 순으로 나타나고 있으며, 본 논문에서 고려한 의미 있는 유사성 측도는 주변비율의 조합의 값에 영향을 받아서 S_{Cohen}, S_{MP}, $S_{Fleiss}S_{KR}$의 순으로 절대값의 크기가 커지고 있다.

이번에는 주변비율 전부를 고려한 PIM 기반 유사성 측도들의 유용성을 좀 더 살펴보기 위해 Table 3.3과 같이 불일치 빈도 j의 값의 변화함에 따라 각각의 측도들을 계산한 결과를 Table 3.4에 나타내었다. 여기서 j가 취할 수 있는 값의 범위는 $0 \le j \le 20$이다.

Table 3.3 Simulation data(2)

		B		Total
		1	0	
A	1	$30-j$	$50+j$	80
	0	j	$20-j$	20
Total		30	70	100

Table 3.4에서 보는 바와 같이 불일치비율 c의 값이 증가함에 따라 지지도와 신뢰도가 감소하고 있으며, 바람직한 측도 중에서는 S_{Cohen}, S_{MP}, S_{Fleiss}은 감소하고 있으며, 모두 -1과 1 사이의 값을 갖고 있으나 S_{KR}은 증가하는 동시에 모든 값이 0보다 크다.

Table 3.4 Variation of PIM based similarity measures without mp by Table 3.3

a	b	c	d	$supp$	$conf$	$conf_2$	$lift$	ν	ν^2	χ^2	S_{KR}	S_{Cohen}	S_{Sti}	S_{MP}	S_{FLeiss}
0.29	0.51	0.01	0.19	0.290	0.363	0.967	1.208	0.050	0.0025	7.4405	8.5106	0.1613	6.87073	0.2703	0.4613
0.28	0.52	0.02	0.18	0.280	0.350	0.933	1.167	0.040	0.0016	4.7619	8.8889	0.1290	6.87091	0.2162	0.3690
0.27	0.53	0.03	0.17	0.270	0.338	0.900	1.125	0.030	0.0009	2.6786	9.3023	0.0968	6.87108	0.1622	0.2768
0.26	0.54	0.04	0.16	0.260	0.325	0.867	1.083	0.020	0.0004	1.1905	9.7561	0.0645	6.87125	0.1081	0.1845
0.25	0.55	0.05	0.15	0.250	0.313	0.833	1.042	0.010	0.0001	0.2976	10.2564	0.0323	6.87143	0.0541	0.0923
0.24	0.56	0.06	0.14	0.240	0.300	0.800	1.000	0.000	0.0000	0.0000	10.8108	0.0000	6.87160	0.0000	0.0000
0.23	0.57	0.07	0.13	0.230	0.288	0.767	0.958	-0.010	0.0001	0.2976	11.4286	-0.0323	6.87143	-0.0541	-0.0923
0.22	0.58	0.08	0.12	0.220	0.275	0.733	0.917	-0.020	0.0004	1.1905	12.1212	-0.0645	6.87125	-0.1081	-0.1845
0.21	0.59	0.09	0.11	0.210	0.263	0.700	0.875	-0.030	0.0009	2.6786	12.9032	-0.0968	6.87108	-0.1622	-0.2768
0.20	0.60	0.10	0.10	0.200	0.250	0.667	0.833	-0.040	0.0016	4.7619	13.7931	-0.1290	6.87091	-0.2162	-0.3690
0.19	0.61	0.11	0.09	0.190	0.238	0.633	0.792	-0.050	0.0025	7.4405	14.8148	-0.1613	6.87073	-0.2703	-0.4613
0.18	0.62	0.12	0.08	0.180	0.225	0.600	0.750	-0.060	0.0036	10.7143	16.0000	-0.1935	6.87056	-0.3243	-0.5536
0.17	0.63	0.13	0.07	0.170	0.213	0.567	0.708	-0.070	0.0049	14.5833	17.3913	-0.2258	6.87038	-0.3784	-0.6458
0.16	0.64	0.14	0.06	0.160	0.200	0.533	0.667	-0.080	0.0064	19.0476	19.0476	-0.2581	6.87021	-0.4324	-0.7381
0.15	0.65	0.15	0.05	0.150	0.188	0.500	0.625	-0.090	0.0081	24.1071	21.0526	-0.2903	6.87004	-0.4865	-0.8304
0.14	0.66	0.16	0.04	0.140	0.175	0.467	0.583	-0.100	0.0100	29.7619	23.5294	-0.3226	6.86986	-0.5405	-0.9226
0.13	0.67	0.17	0.03	0.130	0.163	0.433	0.542	-0.110	0.0121	36.0119	26.6667	-0.3548	6.86969	-0.5946	-1.0149
0.12	0.68	0.18	0.02	0.120	0.150	0.400	0.500	-0.120	0.0144	42.8571	30.7692	-0.3871	6.86951	-0.6486	-1.1071
0.11	0.69	0.19	0.01	0.110	0.138	0.367	0.458	-0.130	0.0169	50.2976	36.3636	-0.4194	6.86934	-0.7027	-1.1994
0.10	0.70	0.20	0.00	0.100	0.125	0.333	0.417	-0.140	0.0196	58.3333	44.4444	-0.4516	6.86917	-0.7568	-1.2917

카이 제곱 통계량 측도는 항상 양의 값만 나타날 뿐만 아니라 c가 증가함에 따라 기존의 연관성 평가 기준들이 모두 증가함에도 불구하고 감소하다가 증가하고 있으며 기존의 연관성 평가 기준인 신뢰도는 항상 0 보다 큰 값만을 취하고 있으므로 이들을 이용한 연관성 측정에는 무리가 따른다. 하지만 본 논문에서 고려하고 있는 유사성 측도들 중에서 S_{Cohen}, S_{MP}, S_{Fleiss}는 c가 증가함에 따라 모두 감소하는 형태를 나타내고 있을 뿐만 아니라 양, 음의 부호를 가지고 있으므로 연관성의 방향도 파악할 수 있어서 연관성 평가 기준으로 매우 바람직한 연관성 규칙 평가기준으로 고려할 수 있다. 이 표에서도 어떤 주변비율의 조합을 고려하느냐에 따라 유사성 측도들의 값이 달라진다. 이 표에서 주변 비율 p_1, p_2, q_1, q_2의 값이 각각 0.8, 0.3, 0.2, 0.7이고, 각 유사성 측도에서 고려하는 값들이 각각 $p_2q_1 = 0.06$, $p_1q_1 = 0.16$, $p_2q_2 = 0.21$, $p_1q_2 = 0.56$의 순으로 나타나고 있으므로 본 논문에서 고려하는 의미 있는 유사성 측도는 S_{Cohen}, S_{MP}, S_{Fleiss}의 순서대로 절대값의 크기가 나타났다.

위의 실험 결과를 종합해볼 때, 주변비율 전부를 고려한 PIM 기반 유사성 측도들 중에서 S_{Cohen}, S_{MP}, S_{Fleiss}은 교차표의 각 항의 값이 어떤 형태로 변하더라도 항상 -1과 1 사이의 값을 갖는 동시에 기존의 연관성 변화 양상과도 일치하고 있다. 이들 중에서는 S_{Fleiss}가 변동 폭이 가장 커서 구분하기가 쉬우므로 가장 바람직한 연관성 규칙 평가 기준으로 생각된다.

4. 결론

연관성 규칙 마이닝은 사건 발생 기록 데이터로부터 항목 간의 연관성을 측정하는 기법으로, 분석대상은 대용량 데이터베이스이다. 의미 있는 연관성 규칙을 탐색하기 위해서는 기본적인 연관성 평가 기준 중에서 신뢰도가 가장 많이 활용되고 있다. 그러나 신뢰도는 전항과 후항이 바뀌게 되면 그 값이 달라지는 비대칭 측도이며, 항상 양의 값만을 가지기 때문에 연관성의 방향을 알 수 없다는 약점을 안고 있다. 이러한 문제를 해결하기 위해 본 논문에서는 AMP를 고려한 PIM 기반 유사성 측도를 연관성 평가 기준으로 고려하였다. 특히 이들 측도들은 기존의 연관성 규칙 평가 기준과는 달리 주변비율 전부와 교차표의 모든 항을 고려하여 연관성의 강도를 측정하는 측도이므로 발생하는 모든 정보를 충실히 반영해 주는 측도이다. 카이 제곱 통계량 기반 측도는 항상 양의 값만 나타날 뿐만 아니라 기존의 연관성 평가 기준들의 변화하는 양상과는 다르게 감소하다가 증가하거나 증가하다가 감소하는 형태를 나타내고 있으므로 연관성 평가 기준으로는 바람직하지 못하다고 할 수 있다. 또한 측도 S_{KR}은 기존의 연관성 평가 기준들과 변화하는 양상이 반대로 나타나는 경우도 있고, 값이 1보다 큰 값이 나타나는 경우도 있으므로 연관성 평가 기준으로는 바람직하지 못하다고 할 수 있다. 모의실험 결과를 종합해볼 때, AMP를 고려한 PIM 기반 유사성 측도들 중에서 측도 S_{Cohen}, S_{MP}, 그리고 S_{Fleiss}는 교차표의 각 항의 값이 어떤 형태로 변하더라도 항상 -1과 1 사이의 값을 갖는 동시에 기존의 연관성 변화 양상과도 일치하고 있다. 이들 중에서는 S_{Fleiss}가 변동 폭이 가장 커서 구분하기가 쉬우므로 가장 바람직한 연관성 규칙 평가 기준으로 생각된다. 또한 주변비율의 어떤 조합을 고려하느냐에 따라 유사성 측도들의 값이 달라진다는 사실도 확인할 수 있었다.

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.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 37-46.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, **16**, 297-334.

Fleiss, J. L. (1975). Measuring agreement between two judges on the presence or absence of a trait. *Biometrics*, **31**, 651-659.

Imberman S., Domanski B. and Thompson H.(2001). Boolean analyer-An algorithm that uses a probabilistic interestingness measure to find dependency /association rules in a head trauma data. *Proceedings of Americas Conference on Information Systems*, 369-375.

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.

Kuder, G. F. and Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika*, **2**, 151-160.

Maxwell, A. E. and Pilliner, A. E. G. (1968). Deriving coefficients of reliability and agreement for ratings. *British Journal of Mathematical and Statistical Psychology*, **21**, 105-116.

Orchard, R. A. (1975). On the determination of relationships between computer system state variables. *Bell Laboratories Technical Memorandum*, Bell Laboratories, New Jersey.

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. (2011a). Proposition of negatively pure association rule threshold. *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. (2011c). The application of some similarity measures to association rule thresholds. *Journal of the Korean Data Analysis Society*, **13**, 1331-1342.

Park, H. C. (2012a). Negatively attributable and pure confidence for generation of negative association rules. *Journal of the Korean Data & Information Science Society*, **14**, 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. (2012c). Exploration of PIM based similarity measures with PMP as association rule thresholds. *Journal of the Korean Data Analysis Society*, to be published.

Piatetsky-Shapiro, G (1991). Discovery, analysis and presentation of strong rules. *Proceedings of the 9th National Conference on Artificial Intelligence: Knowledge Discovery in Databases*, 229-248.

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

Stiles, H. E. (1961). The association factor in information retrieval. *Journal of the Association for Computing Machinery*, **8**, 271-279.

Warrens M. J. (2008). *Similarity coefficients for binary data, properties of coefficients, coefficient matrices, multi-way metrics and multivariate coefficients*, The Doctoral paper of Leiden University, Netherlands.

Journal of the Korean Data & Information Science Society
2013, **24**(1), 117–124

http://dx.doi.org/10.7465/jkdi.2013.24.1.117
한국데이터정보과학회지

Utilization of similarity measures by PIM with AMP as association rule thresholds

Hee Chang Park[1]

[1]Department of Statistics, Changwon National University

Received 10 December 2012, revised 28 December 2012, accepted 7 January 2013

Abstract

Association rule of data mining techniques is the method to quantify the relationship between a set of items in a huge database, andhas been applied in various fields like internet shopping mall, healthcare, insurance, and education. There are three primary interestingness measures for association rule, support and confidence and lift. Confidence is the most important measure of these measures, and we generate some association rules using confidence. But it is an asymmetric measure and has only positive value. So we can face with difficult problems in generation of association rules.In this paper we apply the similarity measures by probabilistic interestingness measure (PIM) with all marginal proportions (AMP) to solve this problem. The comparative studies with support, confidences, lift, chi-square statistics, and some similarity measures by PIM with AMPare shown by numerical example. As the result, we knew that the similarity measures by PIM with AMP could be seen the degree of association same as confidence. And we could confirm the direction of association because they had the sign of their values, and select the best similarity measure by PIM with AMP.

Keywords: All marginal proportion, association rule, confidence, probabilistic interestingness measure, similarity measure, support.

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