Journal of the Korean Data & Information Science Society
2012, **23**(6), 1127–1135

http://dx.doi.org/10.7465/jkdi.2012.23.6.1127
한국데이터정보과학회지

확률적 흥미도를 이용한 유사성 측도의 연관성 평가 기준

박희창[1]

[1]창원대학교 통계학과

접수 2012년 10월 22일, 수정 2012년 11월 6일, 게재확정 2012년 11월 12일

요약

연관성 규칙 기법은 대용량데이터베이스에 있는 항목들 간의 관련성을 수치화 하는 것으로 데이터 마이닝 기법 중에서는 가장 많이 활용되고 있다. 연관성 규칙을 탐사하기 위한 연관성 규칙 평가 기준에는 지지도, 신뢰도, 향상도 등이 있다. 이들 중에서 가장 중심이 되는 신뢰도는 비대칭적 측도일 뿐만 아니라 항상 양의 값만을 취하고 있어서 항목 간에 연관성 규칙을 생성하는 데 여러 가지 문제가 존재한다. 이러한 문제를 해결하기 위해 본 논문에서는 확률적 흥미도 측도 기반, 특히 주변 비율을 고려하지 않은 유사성 측도를 연관성 평가 기준으로 적용하는 방안에 대해 연구하였다. 예제에 의한 비교를 통하여 Yule과 Michael의 유사성 계수와 Pearson의 파이 계수는 신뢰도와 동일하게 연관성의 정도를 파악할 수 있는 동시에 부호를 포함하고 있어서 연관성의 방향도 알 수 있었으나, 카이 제곱 통계량 기반 측도들은 항상 양의 값만 나타날 뿐만 아니라 신뢰도와는 변화하는 양상이 다르다는 것을 확인할 수 있었다.

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

1. 서론

데이터 마이닝이란 대용량 데이터베이스로부터 쉽게 드러나지 않는 유용한 정보들을 추출하는 기법을 의미한다. 지금까지 연관성 규칙, 의사결정나무, 신경망, 클러스터링, 유전자 알고리즘, 베이지안 네트워크 등 다양한 데이터 마이닝 기법들이 개발되었다. 이러한 데이터 마이닝 기법은 마케팅, 제조업 및 보험업, 그리고 의료 및 교육 등 많은 분야에서 적용되고 있다. 특히 데이터 마이닝 기법들 중에서 연관성 규칙은 하나의 거래나 사건에 포함되어 있는 둘이상의 항목들의 경향을 파악해서 상호 관련성을 발견하는 것으로 대용량 데이터베이스에 존재하는 항목간의 관련성을 찾아내는 기법이다. 연관성 규칙은 각 항목간의 연관성을 반영하는 규칙으로 둘 또는 그 이상의 항목들 사이의 지지도, 신뢰도, 향상도를 바탕으로 관련성 여부를 측정한다. 연관성 규칙은 탐색적이며, 비목적성 분석이며, 기존의 데이터를 특별한 변형 없이 계산에 용이하게 사용 가능하다는 장점을 가지고 있으며, 두 품목간의 관계를 명확히 수치화함으로써 두 개 이상의 품목간의 관련성을 나타내기 때문에 현업에서 많이 활용되고 있다 (Cho와 Park, 2011).

연관성 규칙에 관한 연구는 Agrawal 등 (1993)에 의해 처음 소개된 이후, 많은 학자들에 의해 수행된 바 있다. 이들 중에는 Han과 Fu (1995, 1999), Srikant 등 (1997), Cai 등 (1998), Liu 등 (1999)의 제약조건을 가지는 항목으로 구성된 트랜잭션 데이터베이스에서 빈발항목을 찾는 연구가 있으며, Agrawal과 Srikant (1994), Park 등 (1995), Bayardo (1998), Pasquier 등 (1999), Han 등 (2000), Pei

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

등 (2000), Toivonen (1996) 등의 연관성규칙 생성에 대한 수행속도를 향상시키기 위한 연구가 있다. 또한 연관성 규칙에 대한 국내 연구로는 Cho와 Park (2011), Lim 등 (2010), Park (2010a, 2010b, 2011a, 2011b, 2011c) 등이 있다 (Park, 2010a).

일반적인 연관성 규칙 생성과정은 먼저 사용자가 지정한 최소 지지도의 조건을 만족하는 빈발항목집합을 생성한 후 이들에 대해 최저 신뢰도의 조건을 만족하는 규칙을 연관성 규칙으로 채택하게 된다 (Park, 2011a). 이 때 규칙 생성 여부를 결정하기 위해 사용되는 신뢰도는 전항과 후항이 바뀌게 되면 그 값이 달라지는 비대칭적 측도가 되는 동시에 항상 양의 값을 가진다. 따라서 신뢰도의 크기로는 양의 연관성 있는지, 아니면 음의 연관성이 있는지를 알 수 없다. 이러한 문제를 해결하기 위해 본 논문에서는 Warrens (2008)에 의해 정리된 주변 비율을 고려하지 않은 확률적 흥미도 측도 (probabilistic interestingness measures; PIM)를 이용한 유사성 측도에 대해 연관성 평가 기준으로서의 적용 가능 여부를 탐색하고자 한다. Orchard (1975)는 이러한 PIM을 이용한 Boolean Analyzer 알고리즘을 제시하였고, Imberman 등 (2001)이 머리 외상 데이터에 이를 적용하여 연관성 규칙을 생성한 바 있다. 논문의 2절에서는 주변 비율을 고려하지 않은 PIM 기반 유사성 측도들을 소개한다. 3절에서는 구체적인 예제를 통하여 기존의 연관성 규칙 평가 기준과 본 논문에서 고려한 유사성 측도와의 비교를 통해 유사성 측도의 유용성을 살펴본 후, 4절에서 결론을 내리고자 한다.

2. PIM 기반 유사성 측도

기존의 연관성 규칙의 평가기준인 지지도, 신뢰도, 향상도 등을 수식으로 나타내기 위해 Table 2.1과 같은 분할표를 고려한다.

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 비율 ($conf$)을 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 비율 ($conf_2$)을 의미하며, n_{11}/n_+1이 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.

본 논문에서는 PIM 기반 유사성 측도 중에서 원래의 공식에서 주변비율이 존재하지 않거나 존재한다고 해도 수식을 카이제곱으로 나타낸 후에는 주변비율 (marginal proportion; mp)이 존재하지 않는 유사성 측도를 주변비율이 없는 PIM 기반 유사성 측도라고 명명하기로 한다. Warrens (2008)에 의해 정리된 PIM 기반 유사성 측도들을 수식으로 나타내기 위해 Table 2.1의 각 항을 총도수로 나누어서 다시 작성하면 Table 2.2와 같다. 특히 주변비율이 없는 PIM 기반 유사성 측도는 기존의 연관규칙 평가 기준과는 달리 교차표의 모든 항을 고려하여 ad-bc의 값에 대한 크기를 이용하여 연관성의 강도를 측정하는 측도이다.

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

이 표에서 각 항은 $a = P(A \cap B)$, $b = P(A \cap B^c)$, $c = P(A^c \cap B)$, $d = P(A^c \cap B^c)$, $p_1 = a + b$, $q_1 = c + d$, $p_2 = a + c$, 그리고 $q_2 = b + d$을 의미한다. 이로부터 PIM은 다음과 같이 정의된다.

$$PIM = n^2(ad - bc) \tag{2.1}$$

여기서 a와 d는 A, B 두 항목이 모두 발생하거나 두 항목 모두 발생하지 않는 경우의 비율을 나타내므로 두 항목의 연관성의 방향이 동일하다고 할 수 있다. 반면에 b와 c는 A, B 두 항목 중에서 하나는 발생하고 다른 하나는 발생하지 않는 경우의 비율을 의미하므로 두 항목의 연관성의 방향이 동일하지 않다고 할 수 있다. 따라서 $ad - bc$의 값이 양이면 정(+)의 연관성이 있다고 할 수 있고, 음의 값을 취하면 부(-)의 연관성이 있다고 할 수 있으며, 그 값이 클수록 연관성 강도가 더 강하다고 볼 수 있다. 따라서 PIM은 연관 정도를 수치적으로 표현 가능하게 하는 척도로서 연관 정도의 순위까지도 알 수 있다. Table 2.2에서 항목 A와 항목 B가 서로 독립이면 PIM이 0이 되어 아무런 연관 관계가 없는 것으로 판단한다. 만약 항목 A와 B가 독립이 아니면 PIM이 0이 아닌 값으로 나타나므로 이를 통해 연관 정도를 확인할 수 있다.

Table 2.2를 이용하여 일반적으로 가장 많이 알려져 있는 카이 제곱 통계량을 나타내면 식 (2.1)과 같다.

$$\chi^2 = \frac{n(ad - bc)^2}{p_1 p_2 q_1 q_2} \tag{2.2}$$

Warrens (2008)가 정리한 바와 같이 주변비율을 고려하지 않는 PIM 기반 유사성 측도에는 Doolittle (1885)와 Pearson (1926)의 s_{Doo}, Yule (1900)과 Montgomery와 Crittenden (1977)의 s_{Yule1}, Yule(1912)과 Pearson과 Heron(1913)의 파이 계수 s_{Phi}, Michael (1920)의 s_{Mich}등이 있으며, 이들을 Table 2.2를 이용하여 수식으로 나타내면 다음과 같다. 이들 측도들은 기존의 연관규칙 평가 기준과는 달리 교차표의 모든 항을 고려한 PIM의 값을 이용하여 연관성의 강도를 측정하는 측도이며, 특히 s_{Yule1}는 Q 계수로 많이 알려져 있다.

$$s_{Doo} = \frac{(ad - bc)^2}{p_1 p_2 q_1 q_2} = \frac{\chi^2}{n} \tag{2.3}$$

$$s_{Yule1} = \frac{ad - bc}{ad + bc} \tag{2.4}$$

$$s_{Phi} = \frac{ad - bc}{\sqrt{p_1 p_2 q_1 q_2}} = \pm\sqrt{\frac{\chi^2}{n}} \tag{2.5}$$

$$s_{Mich} = \frac{4(ad - bc)}{(a + d)^2 + (b + c)^2} \tag{2.6}$$

카이 제곱 통계량과 s_{Doo}는 ad와 bc의 값의 차이가 크게 날수록, 즉 연관성 강도가 강할수록 더 큰 값을 가지게 된다, 그러나 이들 두 측도는 PIM을 제곱한 값을 측도의 계산에 고려하므로 연관성의 방향을 알 수가 없다. 따라서 이들 두 측도를 신뢰도값 $conf$와 $conf_2$와 비교하여 설명하기가

곤란하므로 이들 측도를 이용해서는 신뢰도의 의미를 해석할 수 없는 동시에 신뢰도를 보완한 측도로 보기는 어렵다. 따라서 측도 S_{Doo}를 제외한 나머지 측도들에 대해 Piatetsky-Shapiro (1991)가 제안한 흥미도 측도의 세 가지 조건만족 여부를 조사해보면 먼저 식 (2.3), (2.4), 그리고 (2.5)로부터 $P(A\cap B) = P(A)P(B)$이면 각 식들의 분자가 0이 되므로 측도 s_{Yule1}, s_{Phi}, s_{Mich}의 값은 모두 0이 된다. 그리고 p_1이 증가한다는 것은 a 또는 b가 증가한다는 의미이므로 이들 측도들의 분모는 증가하는 반면에 분자는 감소하게 되므로 s_{Yule1}, s_{Phi}, s_{Mich}는 p_1의 값에 따라 단조 감소한다. 또한 먼저 p_2가 증가한다는 것은 a 또는 c가 증가한다는 의미이므로 이들 측도들의 분모는 증가하는 반면에 분자는 감소하게 되므로 s_{Yule1}, s_{Phi}, s_{Mich}는 p_2의 값에 따라 단조 감소한다. 마지막으로 $P(X\cap Y)$의 값이 증가한다는 의미는 a가 증가한다는 것이므로 식 (2.3), (2.4), 그리고 (2.5)로부터 a가 증가하면 측도 s_{Yule1}, s_{Phi}, s_{Mich}의 값은 모두 증가한다는 것을 알 수 있다.

3. 예제를 통한 유용성 고찰

본 절에서는 예제를 통하여 조건부 확률에 의한 대칭 유사성 측도들의 유용성에 대해 탐색하고자 한다. 이를 위해 항목 집합 A, B에 대해 Park (2011a)에서와 같이 가정하였다. 먼저 데이터베이스에 있는 총 트랜잭션의 수 (t)를 100명으로 하고, 항목 집합 A는 구매한 물품의 금액을 기준으로 특정 금액 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 : 현금)으로 결제 (1)한 사람 수를 30명으로 하고 그 외의 방법으로 결제 (0)한 사람의 수를 70명으로 하였다. 항목 집합 A와 B가 동시에 발생한 빈도수, 즉 특정금액 이상의 물품을 구매하면서 특정방법으로 결제한 빈도수는 h명으로 하였다. 이를 정리하면 Table 3.1과 같다. 여기서 동시발생빈도 h의 정수 값 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

Table 3.1을 이용하여 동시발생비율 a의 변화에 따라 계산된 유사성 측도들과 지지도 및 신뢰도를 미니탭 16에 의하여 계산한 후, 그 일부를 나타내면 Table 3.2와 같다. 여기서 a, b, c, d는 Table 2.2에서의 각 셀을 의미하고, nu는 $ad-bc$, $supp$는 지지도, $lift$는 향상도, 그리고 신뢰도 $conf$ 및 $conf_2$는 전항 변수가 각각 A와 B인 신뢰도를 의미한다. 이 표를 살펴보면 동시발생비율 a의 값이 증가함에 따라 $supp$와 $lift$, 그리고 $conf$ 및 $conf_2$가 증가하고 있으며, 본 논문에서 고려하는 PIM 기반 유사성 측도들인 s_{Yule1}, s_{Phi}, s_{Mich}역시 증가하고 있다. 반면에 측도 chi 및 s_{Doo}는 PIM의 제곱의 형태로 식이 나타나므로 감소하다가 증가하는 경향을 나타내고 있다. 그리고 측도 chi 및 s_{Doo}는 각각 0과 1 사이의 값을 취하는 반면에, 측도 s_{Yule1}, s_{Phi}그리고 s_{Mich}는 각각 -1과 1사이의 값을 갖는다. 따라서 카이 제곱 통계량 기반 측도인 chi 및 s_{Doo}는 항상 양의 값만 나타날 뿐만 아니라 동시발생비율 a가 증가함에 따라 기존의 연관성 평가 기준들이 모두 증가함에도 불구하고 감소하다가 증가하는 형태를 나타내고 있으므로 연관성 평가 기준으로는 바람직하지 못하다고 할 수 있다. 그러나 PIM 기반 유사성 측도인 s_{Yule1}, s_{Phi}, s_{Mich}는 모두 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. 뿐만 아니라 기존의 연관성 평

가기준에서 나타내지 못하는 연관성의 방향을 나타내주는 매우 바람직한 연관성 규칙 평가기준으로 고려해볼만 하다. 한편, PIM 기반 유사성 측도인 s_{Yule1}, s_{Phi}, s_{Mich}을 좀 더 구체적으로 비교해보면 이 표에서는 s_{Yule1}이 가장 바람직한 것으로 판단된다. 그 이유는 다른 측도들에 비해 a가 증가함에 따라 측도 s_{Yule1}이 변화하는 폭이 가장 크다는 사실을 확인할 수 있다. 또한 a의 값이 증가함에 따라 nu의 값이 증가하고 있으며, 변화의 폭은 s_{Yule1}, s_{Mich}, s_{Phi}의 순으로 s_{Yule1}이 가장 크며, 이 측도는 최소값과 최대값이 각각 -1과 1에 가까운 값이 되어서 연관성의 강도를 쉽게 파악할 수 있다.

Table 3.2 Variation of PIM based similarity measures without mp by Table 3.1

a	b	c	d	$supp$	$conf$	$conf2$	$lift$	nu	nu^2	chi	s_{Doo}	s_{Yule1}	s_{Phi}	s_{Mich}
0.11	0.39	0.19	0.31	0.110	0.220	0.367	0.440	-0.040	0.0016	3.0476	0.0305	-0.3697	-0.1746	-0.3120
0.12	0.38	0.18	0.32	0.120	0.240	0.400	0.480	-0.030	0.0009	1.7143	0.0171	-0.2809	-0.1309	-0.2366
0.13	0.37	0.17	0.33	0.130	0.260	0.433	0.520	-0.020	0.0004	0.7619	0.0076	-0.1890	-0.0873	-0.1590
0.14	0.36	0.16	0.34	0.140	0.280	0.467	0.560	-0.010	0.0001	0.1905	0.0019	-0.0951	-0.0436	-0.0799
0.15	0.35	0.15	0.35	0.150	0.300	0.500	0.600	0.000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.16	0.34	0.14	0.36	0.160	0.320	0.533	0.640	0.010	0.0001	0.1905	0.0019	0.0951	0.0436	0.0799
0.17	0.33	0.13	0.37	0.170	0.340	0.567	0.680	0.020	0.0004	0.7619	0.0076	0.1890	0.0873	0.1590
0.18	0.32	0.12	0.38	0.180	0.360	0.600	0.720	0.030	0.0009	1.7143	0.0171	0.2809	0.1309	0.2366
0.19	0.31	0.11	0.39	0.190	0.380	0.633	0.760	0.040	0.0016	3.0476	0.0305	0.3697	0.1746	0.3120
0.20	0.30	0.10	0.40	0.200	0.400	0.667	0.800	0.050	0.0025	4.7619	0.0476	0.4545	0.2182	0.3846
0.21	0.29	0.09	0.41	0.210	0.420	0.700	0.840	0.060	0.0036	6.8571	0.0686	0.5348	0.2619	0.4539
0.22	0.28	0.08	0.42	0.220	0.440	0.733	0.880	0.070	0.0049	9.3333	0.0933	0.6098	0.3055	0.5193
0.23	0.27	0.07	0.43	0.230	0.460	0.767	0.920	0.080	0.0064	12.1905	0.1219	0.6791	0.3491	0.5806
0.24	0.26	0.06	0.44	0.240	0.480	0.800	0.960	0.090	0.0081	15.4286	0.1543	0.7426	0.3928	0.6374
0.25	0.25	0.05	0.45	0.250	0.500	0.833	1.000	0.100	0.0100	19.0476	0.1905	0.8000	0.4364	0.6897
0.26	0.24	0.04	0.46	0.260	0.520	0.867	1.040	0.110	0.0121	23.0476	0.2305	0.8514	0.4801	0.7373
0.27	0.23	0.03	0.47	0.270	0.540	0.900	1.080	0.120	0.0144	27.4286	0.2743	0.8969	0.5237	0.7802
0.28	0.22	0.02	0.48	0.280	0.560	0.933	1.120	0.130	0.0169	32.1905	0.3219	0.9366	0.5674	0.8186
0.29	0.21	0.01	0.49	0.290	0.580	0.967	1.160	0.140	0.0196	37.3333	0.3733	0.9709	0.6110	0.8526

본 논문에서 제안한 유사성 측도들의 유용성을 좀 더 살펴보기 위해 이번에는 Table 3.3과 같이 불일치 빈도 i의 값의 변화함에 따라 각각의 측도들을 계산하여 그 결과를 Table 3.4에 나타내었다.

Table 3.3 Simulation data (2)

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

이 표에서 보는 바와 같이 불일치비율 b의 값이 증가함에 따라 $supp$와 $lift$, $conf$ 및 $conf_2$모두가 감소하고 있으며, nu와 PIM 기반 유사성 측도 s_{Yule1}, s_{Phi}, s_{Mich}도 감소하고 있다. 반면에 카이제곱 통계량 기반 유사성 측도 chi 및 s_{Doo}는 PIM의 제곱의 형태로 식이 나타나므로 감소하다가 증가하는 경향을 나타내고 있다. 앞의 표의 결과에서와 마찬가지로 카이제곱 통계량 기반 유사성 측도 chi 및 s_{Doo}는 모두 0과 1 사이의 값을 취하는 반면에, PIM 기반 유사성 측도들은 모두 -1과 1사이의 값을 갖는다. 따라서 PIM 기반 유사성 측도들은 모두 b가 증가함에 따라 기존의 연관성 평가 기준들과 마찬가지로 모두 감소하는 형태를 나타내고 있을 뿐만 아니라 양, 음의 부호를 가지고 있으므로 연관성의 방향도 파악할 수 있어서 연관성 평가 기준으로 사용가능하다고 할 수 있다. 특히 b의 값이 증가함에 따라 nu의 값이 감소하고 있으며, PIM 기반 유사성 측도의 변화폭은 s_{Yule1}, s_{Mich}, s_{Phi}의 순으로 s_{Yule1}이 가장 크고, 또한 측도 s_{Yule1}은 최대값과 최소값이 각각 1과 -1에 가까운 값이 되어서 이 값만으로 연관성의 강도를 쉽게 파악할 수 있으므로 s_{Yule1}이 가장 바람직한 측도라고 생각된다.

한편, 불일치비율 $P(\bar{A} \cap B)$의 변화하는 양상에 따라 유사성 측도들을 계산해 보았는데, Table 3.4에서와 마찬가지로 불일치비율이 증가함에 따라 $supp$와 $lift$, $conf$ 및 $conf_2$모두가 감소하였으며,

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

a	b	c	d	$supp$	$conf$	$conf2$	$lift$	nu	nu^2	chi	s_{Doo}	s_{Yule1}	s_{Phi}	s_{Mich}
0.39	0.11	0.31	0.19	0.390	0.780	0.557	1.114	0.04	0.0016	3.0476	0.0305	0.3697	0.1746	0.3120
0.38	0.12	0.32	0.18	0.380	0.760	0.543	1.086	0.03	0.0009	1.7143	0.0171	0.2809	0.1309	0.2366
0.37	0.13	0.33	0.17	0.370	0.740	0.529	1.057	0.02	0.0004	0.7619	0.0076	0.1890	0.0873	0.1590
0.36	0.14	0.34	0.16	0.360	0.720	0.514	1.029	0.01	0.0001	0.1905	0.0019	0.0951	0.0436	0.0799
0.35	0.15	0.35	0.15	0.350	0.700	0.500	1.000	0.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.34	0.16	0.36	0.14	0.340	0.680	0.486	0.971	-0.01	0.0001	0.1905	0.0019	-0.0951	-0.0436	-0.0799
0.33	0.17	0.37	0.13	0.330	0.660	0.471	0.943	-0.02	0.0004	0.7619	0.0076	-0.1890	-0.0873	-0.1590
0.32	0.18	0.38	0.12	0.320	0.640	0.457	0.914	-0.03	0.0009	1.7143	0.0171	-0.2809	-0.1309	-0.2366
0.31	0.19	0.39	0.11	0.310	0.620	0.443	0.886	-0.04	0.0016	3.0476	0.0305	-0.3697	-0.1746	-0.3120
0.30	0.20	0.40	0.10	0.300	0.600	0.429	0.857	-0.05	0.0025	4.7619	0.0476	-0.4545	-0.2182	-0.3846
0.29	0.21	0.41	0.09	0.290	0.580	0.414	0.829	-0.06	0.0036	6.8571	0.0686	-0.5348	-0.2619	-0.4539
0.28	0.22	0.42	0.08	0.280	0.560	0.400	0.800	-0.07	0.0049	9.3333	0.0933	-0.6098	-0.3055	-0.5193
0.27	0.23	0.43	0.07	0.270	0.540	0.386	0.771	-0.08	0.0064	12.1905	0.1219	-0.6791	-0.3491	-0.5806
0.26	0.24	0.44	0.06	0.260	0.520	0.371	0.743	-0.09	0.0081	15.4286	0.1543	-0.7426	-0.3928	-0.6374
0.25	0.25	0.45	0.05	0.250	0.500	0.357	0.714	-0.10	0.0100	19.0476	0.1905	-0.8000	-0.4364	-0.6897
0.24	0.26	0.46	0.04	0.240	0.480	0.343	0.686	-0.11	0.0121	23.0476	0.2305	-0.8514	-0.4801	-0.7373
0.23	0.27	0.47	0.03	0.230	0.460	0.329	0.657	-0.12	0.0144	27.4286	0.2743	-0.8969	-0.5237	-0.7802
0.22	0.28	0.48	0.02	0.220	0.440	0.314	0.629	-0.13	0.0169	32.1905	0.3219	-0.9366	-0.5674	-0.8186
0.21	0.29	0.49	0.01	0.210	0.420	0.300	0.600	-0.14	0.0196	37.3333	0.3733	-0.9709	-0.6110	-0.8526

nu와 PIM 기반 유사성 측도 s_{Yule1}, s_{Phi}, s_{Mich}도 감소하였다. 반면에 카이제곱 통계량 기반 유사성 측도 chi 및 s_{Doo}는 PIM의 제곱의 형태로 식이 나타나므로 감소하다가 증가하는 경향을 나타내었다. 앞의 결과와 마찬가지로 카이제곱 통계량 기반 유사성 측도 chi 및 s_{Doo}는 모두 0과 1 사이의 값을 취하는 반면에, PIM 기반 유사성 측도들은 모두 -1과 1사이의 값을 갖는다. 따라서 PIM 기반 유사성 측도들은 모두 $P(\bar{A} \cap B)$이 증가함에 따라 기존의 연관성 평가 기준들과 마찬가지로 모두 감소하는 형태를 나타내고 있을 뿐만 아니라 양, 음의 부호를 가지고 있으므로 연관성의 방향도 파악할 수 있어서 연관성 평가 기준으로 사용가능하다고 할 수 있다. 특히 c의 값이 증가함에 따라 nu의 값이 감소하고 있으며, PIM 기반 유사성 측도의 변화폭은 s_{Yule1}, s_{Phi}, s_{Mich}의 순으로 s_{Yule1}이 가장 크고, 또한 측도 s_{Yule1}은 최대값과 최소값이 각각 1과 -1에 가까운 값이므로 이 측도만으로 연관성의 강도를 쉽게 파악할 수 있다. 또한 동시비발생비율이 증가함에 따라 $supp$와 $lift$, $conf$ 및 $conf_2$모두가 증가하였으며, nu와 PIM 기반 유사성 측도 s_{Yule1}, s_{Phi}, s_{Mich}도 증가한다는 사실을 확인하였다.

4. 결론

연관성 규칙은 둘이상의 항목들 간의 상호 관련성을 발견하고 분석하는 방법으로 발생시점에서 기록되어진 거래에 관한 방대한 양의 데이터베이스를 분석대상으로 한다. 일반적으로 의미 있는 연관성 규칙을 생성하기 위해서는 신뢰도가 가장 많이 활용되고 있으나 이는 전항과 후항이 바뀌면 그 값이 달라지는 비대칭 측도인 동시에 항상 양의 값을 취하기 때문에 연관성의 방향을 알 수 없다. 이러한 문제를 해결하기 위해 본 논문에서는 PIM 기반 유사성 측도 중에서 원래의 공식에서 주변비율이 존재하지 않거나 존재한다고 해도 수식을 카이제곱으로 나타낸 후에 주변비율이 존재하지 않는 유사성 측도를 연관성 평가 기준으로 고려해 보았다. 그 결과, s_{Yule1}, s_{Phi}, s_{Mich}는 기존의 연관성 평가 기준과 동일하게 연관성의 정도를 파악할 수 있는 동시에 값의 범위가 [-1, 1]로 부호를 포함하고 있어서 연관성의 방향도 알 수 있었으나, 카이 제곱 통계량 기반 측도인 chi 및 s_{Doo}는 항상 양의 값만 나타날 뿐만 아니라 증가하다가 감소하거나, 또는 감소하다가 증가하는 형태를 나타내고 있으므로 기존의 연관성 평가 기준과는 다른 양상을 보이는 것을 확인할 수 있었다. 따라서 PIM 기반 유사성 측도인 s_{Yule1}, s_{Phi}, s_{Mich}는 모두 양, 음의 부호를 가지고 있으므로 기존의 연관성 평가기준에서 나타내지 못하는 연관성의 방향을 나타내주는 매우 바람직한 연관성 규칙 평가기준으로 고려해볼만 하다. 특히 측도 s_{Yule1}은 다른 측도들에 비해 변화하는 폭이 가장 크고 최소값과 최대값이 각각 -1

과 1에 가까운 값이 되어서 연관성의 강도를 쉽게 파악할 수 있어서 본 논문에서 고려한 유사성 측도 중에서는 가장 바람직한 연관성 평가 기준으로 판단된다.

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.

Bayardo, R. J. (1998). Efficiently mining long patterns from databases. *Proceedings of ACM SIGMOD Conference on Management of Data*, 85-93.

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. (2011). Discovery of insignificant association rule s using external variable. *Journal of the Korean Data Analysis Society*, **13**, 1343-1352.

Doolittle, M. H. (1885). The verification of predictions. *Bulletin of the Philosophical Society of Washington*, **7**, 122-127.

Han, J. and Fu, Y. (1995). Discovery of multiple-level association rules from large databases. *Proceeding of the 21st VLDB Conference*, 420-431.

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.

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.

Lim, J., Lee, K. and Cho, Y. (2010). A study of association rule by considering the frequency. *Journal of the Korean Data & Information Science Society*, **21**, 1061-1069.

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.

Michael, E. L. (1920). Marine ecology and the coefficient of association. *Journal of Animal Ecology*, **8**, 54-59.

Montgomery, A. C. and Crittenden, K. S. (1977). Improving coding reliability for open-ended questions. *Public Opinion Quarterly*, **41**, 235-243.

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, 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.

Pearson, K. (1926). On the coefficient of racial likeness. *Biometrika*, 9, 105-117.

Pearson, K and Heron, D. (1913). On theories of association. *Biometrika*, **9**, 159-315.

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.

Srinkant R., Vu Q. and Agrawal R. (1997). Mining association rules with item constraints. *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, 67-73.

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

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.

Yule, G. U. (1900). On the association of attributes in statistics. *Philosophical Transactions of the Royal Society*, **75**, 257-319.

Yule, G. U. (1912). On the methods of measuring the association between two attributes. *Journal of the Royal Statistical Society* , **75**, 579-652.

Journal of the Korean Data & Information Science Society
2012, **23**(6), 1127–1135

http://dx.doi.org/10.7465/jkdi.2012.23.6.1127
한국데이터정보과학회지

Exploration of PIM based similarity measures as association rule thresholds

Hee Chang Park[1]

[1]Department of Statistics, Changwon National University

Received 22 October 2012, revised 6 November 2012, accepted 12 November 2012

Abstract

Association rule mining is the method to quantify the relationship between each set of items in a large database. One of the well-studied problems in data mining is exploration for association rules. There are three primary quality measures for association rule, support and confidence and lift. We generate some association rules using confidence. Confidence is the most important measure of these measures, but it is an asymmetric measure and has only positive value. Thus we can face with difficult problems in generation of association rules.In this paper we apply the similarity measures by probabilistic interestingness measure to find a solution to this problem. The comparative studies with support, two confidences, lift, and some similarity measures by probabilistic interestingness measure are shown by numerical example. As the result, we knew that the similarity measures by probabilistic interestingness measure 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.

Keywords: Association rule, confidence, lift, 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