Journal of the Korean Data & Information Science Society
2010, **21**(5), 891–899

한국데이터정보과학회지

연관 규칙 마이닝에서의 평가기준 표준화 방안

박희창[1]

[1]창원대학교 통계학과

접수 2010년 7월 19일, 수정 2010년 8월 30일, 게재확정 2010년 9월 6일

요 약

연관성 규칙은 방대한 양의 데이터베이스 속에 있는 각 항목들 간의 관련성을 수치화함으로써 두 개 이상의 항목간의 관련성을 나타내는 기법으로 데이터 마이닝 분야에서 가장 많이 활용되고 있다. 의미 있는 연관성 규칙을 탐색하기 위한 가장 기본적인 평가기준에는 지지도, 신뢰도, 향상도 등이 있으며, 이들을 이용하여 연관성 규칙을 생성하게 된다. 이 때 사용되는 향상도는 그 값의 범위가 지지도나 신뢰도와는 다르므로 지지도나 신뢰도의 범위를 동일하도록 하기 위해 표준화할 필요가 있으며, 지지도와 신뢰도도 하나의 후항변수에 대해 여러 개의 전항변수들이 있는 경우 이들 중 어느 것이 후항변수와 가장 연관성이 있는지를 객관적으로 비교하기 위해서도 표준화가 필요하다.본 논문에서는 각 항목집합의 주변 발생확률을 고려하여 객관적이고도 정확한 연관성 정도를 파악하기 위해 연관성 평가기준을 표준화하는 방안에 대해 연구하고자 한다. 또한 흥미도 측도의 세 가지 조건의 충족 여부를 점검해 본 후, 구체적인 예제를 통하여 기존의 연관성 평가기준과 표준화된 연관성 평가기준을 비교 분석하고자 한다.

주요용어: 신뢰도, 연관성 규칙, 지지도, 표준화 연관규칙 평가기준, 향상도.

1. 서론

연관성 규칙 (association rule)은 데이터마이닝 분야에서장바구니 분석 개념으로 처음 소개되었으며, 대용량 데이터베이스에 내재되어 있는 각 항목들 간의 관련성을 찾아내는 데 활용되며, 특히 유통업이나 제조업 등에서 많이 활용되고 있다. 이러한 연관성 규칙은 Agrawal 등 (1993)이 최초로 제안하였으며, 이후 많은 학자들이 연관성 규칙과 관련된 연구를 수행하였다 (Agrawal과 Srikant, 1994; Park 등, 1995; Srikant와 Agrawal, 1995; Toivonen, 1996; Bayardo, 1998; Cai 등, 1998; Han과 Fu, 1999; Liu 등, 1999; Pasquier 등, 1999; Han 등, 2000; Pei 등, 2000; Park과 Cho, 2005; Cho와 Park, 2007; Cho와 Park, 2008; Choi와 Park, 2008; Park, 2008).

연관성 규칙은 시간의 순서를 고려하지 않는 비목적성 분석기법이며, 이에 적용되는 데이터의 형태는 발생시점에서 기록되어진 항목에 관한 정보만으로 구성되어 있다. 의미 있는 연관성 규칙을 탐색하기 위한 가장 기본적인 흥미도 측도에는 지지도 (support), 신뢰도 (confidence), 향상도 (lift) 등이 있으며, 이들을 이용하여 연관성 규칙을 생성하게 된다. 일반적인 연관성 규칙 생성과정은 먼저 사용자가 지정한 최소 지지도를 만족시키는 빈발항목집합을 생성한 후, 이들에 대해 최저신뢰도 기준을 만족하고 향상도가 1이상인 것을 규칙으로 채택하게 된다. 이 때 각 항목 발생 확률의 크기를 고려하여 연관성 측도들을 표준화 하게 되면 모든 연관성 측도의 값이 0과 1 사이의 값을 갖게 되어 보다 객관적으로 연관성 규칙의 강도를 측정할 수 있다. 또한 하나의 후항변수에 대해 여러 개의 전항변수들이 있는 경우 이

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

들 중 어느 것이 후항변수와 가장 연관성이 있는지를 객관적으로 비교하기 위해서도 표준화가 필요하다. 특히 향상도의 경우에는 그 범위가 $[0, \infty]$이므로 그 크기만으로는 연관성 정도가 어느 정도인지를 판단하기가 사실상 어렵다.

본 논문에서는 각 항목집합의 주변 발생확률을 고려하여 객관적이고도 정확한 연관성 정도를 파악하기 위해 연관성 평가기준을 표준화하는 방안에 대해 연구하고자 한다. 특히 McNicholas 등 (2008)의 연구와 연관성 측도들 간의 관계식을 고려하여 또 다른 표준화 향상도를 제시한 후, Piatetsky-Shapiro (1991)가 제안한 흥미도 측도의 세 가지 조건의 충족 여부를 점검해보고자 한다. 또한 구체적인 예제를 통하여 기존의 연관성 평가기준과 표준화된 연관성 평가기준을 비교 분석하고자 한다.

2. 표준화된 연관성 평가 기준

연관성 규칙을 평가하는 가장 기본적인 기준에는 지지도, 신뢰도, 향상도 등이 있다. 일반적으로 연관성 규칙은 공집합이 아닌 항목집합 I가 주어졌을 때 $(A,\ B \subset I)$, $A \neq \phi$, $B \neq \phi$, 그리고 $A \cap B = \phi$인 I의 부분집합 A와 B에 대해 $A \Rightarrow B$의 형태로 기술된다. 이러한 연관성 규칙을 평가하는 가장 기본적인 기준에는 지지도, 신뢰도, 향상도 등이 있다. 지지도 $S(A \Rightarrow B)$는 항목 집합 A와 항목 집합 B가 동시에 발생하는 거래량 (transaction)의 비율을 의미하며, 다음과 같이 정의된다.

$$S(A \Rightarrow B) = \text{A와 B를 동시에 구매하는 거래수/전체 거래수} = P(A \text{ and } B) \tag{2.1}$$

신뢰도 $C(A \Rightarrow B)$는 항목 집합 A가 포함된 거래 비율 중 항목 집합 A와 항목 집합 B가 동시에 포함된 거래의 비율을 의미하며, 다음과 같이 정의된다.

$$C(A \Rightarrow B) = P(B|A) \tag{2.2}$$

향상도 $L(A \Rightarrow B)$는 항목 집합 A를 구매한 경우 그 거래가 항목 집합 B를 포함하는 경우와 항목 집합 B가 임의로 구매되는 경우의 비를 의미하며, 다음과 같이 정의된다.

$$L(A \Rightarrow B) = \frac{P(B|A)}{P(B)} = \frac{P(A \text{ and } B)}{P(A) \cdot P(B)} \tag{2.3}$$

따라서 연관 규칙 마이닝에서는 향상도가 1 이상이고, 최저지지도를 만족하는 규칙들 중에서 최저 신뢰도 기준을 초과하는 경우에 일반적으로 연관성 규칙이 생성되는 것으로 간주한다.일반적으로 지지도와 신뢰도는 0과 1사이의 값을 취하는 반면에 향상도는 0 이상의 값을 갖게 된다. 지지도와 신뢰도의 범위와 향상도의 범위를 동일하게 만들기 위해 McNicholas 등 (2008)은 향상도를 표준화하는 방안에 대해 연구하였으며, 다음과 같이 표준화된 향상도를 제안하였다.

$$L_{ST}(A \Rightarrow B) = \frac{L(A \Rightarrow B) - \lambda}{\nu - \lambda} \tag{2.4}$$

여기서 $\lambda = \max[P(A) + P(B) - 1, 1/n]/(P(A) \cdot P(B))$, $\nu = 1/\max[P(A),\ P(B)]$이다.

McNicholas 등 (2008)이 연구한 내용의 논리적 전개를 고려하면 지지도와 신뢰도에 대해서도 표준화가 가능한 것으로 판단된다. 이들은 만약 $P(A)+P(B) \leq 1$이면 연관성 규칙의 정의에서 $A \cap B = \phi$이므로 $P(A \text{ and } B) \geq 1/n$이고, $P(A) + P(B) > 1$이면 $P(A \text{ and } B) \geq P(A) + P(B) - 1$이라는 두 가지 경우만을 고려하여 $P(A \text{ and } B)$의 하한을 $\max[P(A) + P(B) - 1,\ 1/n]$으로 제안하였다. 또한 $P(A \text{ and } B)$는 $P(A)$와 $P(B)$ 중에서 작은 것보다 작으므로 $P(A \text{ and } B)$의 상한은 $\min[P(A),\ P(B)]$이므로 지지도는 다음과 같은 구간을 취한다고 주장하였다.

$$\max[P(A) + P(B) - 1,\ 1/n] \leq S(A \Rightarrow B) \leq \min[P(A),\ P(B)] \tag{2.5}$$

그러나 McNicholas 등 (2008)은 지지도와 신뢰도와의 관계를 고려하지 않았으므로 이 절에서는 이들에 대한 명확한 관계를 제시하기 위해 좀 더 정확한 지지도에 대한 상하한의 값을 찾고자 한다. 다음과 같은 분할표를 고려해보자.

표 2.1 예제 분할표

		Y		합계
		1	0	
X	1	30	50	80
	0	50	70	120
합계		80	120	200

이 표에서 지지도, 지지도의 상한, 그리고 신뢰도는 각각 $S(A \Rightarrow B) = 0.15$, $\min[P(A),\ P(B)] = 0.4$, $C(A \Rightarrow B) = 0.375$으로 나타났다. 따라서 McNicholas 등 (2008)가 제시한 식 (2.5)의 구간과 $S(A \Rightarrow B) \leq C(A \Rightarrow B)$의 신뢰도와의 관계식을 동시에 고려하는 것이 바람직하며, 이를 토대로 새로운 지지도의 구간을 구하면 다음과 같다.

$$\max[P(A) + P(B) - 1,\ 1/n] \leq S(A \Rightarrow B) \leq \min[P(A),\ P(B),\ P(B|A)] \tag{2.6}$$

이를 표준화하여 0과 1사이의 값을 갖게 되는 표준화된 지지도는 다음과 같이 정의할 수 있다.

$$S_{ST}(A \Rightarrow B) = \frac{S(A \Rightarrow B) - \max[P(A) + P(B) - 1, 1/n]}{\min[P(A),\ P(B),\ P(B|A)] - \max[P(A) + P(B) - 1, 1/n]} \tag{2.7}$$

지지도와 신뢰도와의 관계를 고려하여 식 (2.6)를 변형하면 신뢰도의 범위는 다음과 같이 표현된다.

$$\begin{aligned} \max[1 + P(B)/P(A) - 1/P(A),\ 1/[nP(A)]] &\leq C(A \Rightarrow B) \\ &\leq \min[1,\ P(B)/P(A),\ P(B|A)/P(A)] \end{aligned} \tag{2.8}$$

따라서 표준화된 신뢰도는 다음과 같이 얻어진다.

$$C_{ST}(A \Rightarrow B) = \frac{C(A \Rightarrow B) - L.B._C}{U.B._C - L.B._C} \tag{2.9}$$

여기서

$$\begin{aligned} U.B._C &= \min[1,\ P(B)/P(A),\ P(B|A)/P(A)], \\ L.B._C &= \max[1 + P(B)/P(A) - 1/P(A), 1/[nP(A)]] \end{aligned}$$

이다. 또한 향상도와 신뢰도와의 관계를 고려하여 식 (2.8)을 변형하면 다음과 같은 향상도의 범위를 구할 수 있다.

$$\begin{aligned} \max[1/P(B) + 1/P(A) - 1/[P(A)P(B)], 1/[nP(A)P(B)]] &\leq L(A \Rightarrow B) \\ &\leq \min[1/P(B),\ 1/P(A),\ P(B|A)/[P(A)P(B)]] \end{aligned} \tag{2.10}$$

이로부터 표준화된 향상도는 다음과 같이 얻어진다.

$$L_{ST}(A \Rightarrow B) = \frac{L(A \Rightarrow B) - L.B._L}{U.B._L - L.B._L} \tag{2.11}$$

여기서

$$U.B._L = \min[1/P(B),\ 1/P(A),\ P(B|A)/[P(A)P(B)],$$
$$L.B._L = \max[1/P(B) + 1/P(A) - 1/[P(A)P(B)], 1/[nP(A)P(B)]]$$

이다.

이러한 표준화 측도에 대해 Piatetsky-Shapiro (1991)가 제안한 흥미도 측도의 세 가지 조건 중에서 기본적인 연관성 평가 기준에는 적용되지 않는 [조건 1]을 제외한 나머지 조건들의 충족 여부를 점검해 보면 다음과 같다.

[조건 2] 표준화 연관성 평가기준들은 $P(A)$ 또는 $P(B)$의 값에 따라 단조 감소한다.

(점검): 먼저 $P(A) + P(B) \leq 1$인 경우 식 (2.7)의 $S_{ST}(A \Rightarrow B)$를 정리하면 다음의 식을 얻을 수 있다.

$$S_{ST}(A \Rightarrow B) = \frac{S(A \Rightarrow B) - 1/n}{\min[P(A),\ P(B),\ P(B|A)] - 1/n} \tag{2.12}$$

이때, $P(A)$ 또는 $P(B)$가 $P(B|A)$보다 작으면 $P(A)$와 $P(B)$의 값이 증가함에 따라 $S_{ST}(A \Rightarrow B)$는 단조 감소하는 것을 알 수 있다. 그러나 $P(B|A)$가 $P(A)$ 또는 $P(B)$보다 작게 나타나는 특수한 경우에는 그렇지 않다. 식 (2.9)의 $C_{ST}(A \Rightarrow B)$를 정리하면 다음의 식을 얻을 수 있다.

$$C_{ST}(A \Rightarrow B) = \frac{C(A \Rightarrow B) - 1/[nP(A)]}{\min[1,\ P(B)/P(A),\ P(B|A)/P(A)] - 1/[nP(A)]} \tag{2.13}$$

이를 정리하면 식 (2.12)과 같은 결과가 나오므로 $P(A)$ 또는 $P(B)$가 $P(B|A)$보다 작으면 $P(A)$와 $P(B)$의 값이 증가함에 따라 $C_{ST}(A \Rightarrow B)$는 단조 감소하는 것을 알 수 있다. 식 (2.11)의 $L_{ST}(A \Rightarrow B)$도 동일한 결과를 얻을 수 있으며, $P(A) + P(B) > 1$인 경우에도 동일한 방법으로 점검이 가능하다.

[조건 3] 표준화 연관성 평가기준들은 $P(A \text{ and} B)$ 의 값에 따라 단조 증가한다.

(점검): $P(A) + P(B) \leq 1$인 경우 $S_{ST}(A \Rightarrow B)$, $C_{ST}(A \Rightarrow B)$, $L_{ST}(A \Rightarrow B)$는 위에서 정리한 수식으로부터 $P(A \text{ and} B)$의 값이 증가함에 따라 단조 증가하는 것을 알 수 있으며, $P(A) + P(B) > 1$인 경우에도 동일한 방법으로 증명할 수 있다.

3. 적용 예제

본 절에서는 연관성 규칙을 위한 3가지 평가기준값과 표준화된 평가기준값들에 대해 예제를 통하여 비교하고자 한다. 이를 위해 항목 집합 A, B에 대해 다음과 같이 가정하였다. 먼저 데이터베이스에 있는 총 트랜잭션의 수 (t)를 100명으로 하고, 항목 집합 A는 구매한 물품의 금액을 기준으로 특정금액 이상 (1) 구매한 사람 수를 $(70+u+w)$명으로 하고 특정금액 미만 (0)을 구매한 사람 수를 $(30-u-w)$명으로 하였다. 또한 항목 집합 B를 결제 방식을 기준으로 특정 방법 (예: 신용카드)으로 결제 (1)한 사람 수를 $(55+u-w)$명으로 하고 그 외의 방법으로 결제 (0)한 사람의 수를 $(45-u+w)$명으로 하였다. 항목 집합 A와 B가 동시에 발생한 빈도 수, 즉 특정금액 이상의 물품을 구매하면서 특정방법으로 결제한 빈도수는 $(50+u)$명으로 하였다. 이를 정리하면 표 3.1과 같다. 표에서 u 및 w가 취할 수 있는 정수값의 범위는 각각 $0 \leq u \leq 25$, $0 \leq w \leq 5$이다.

표 3.1 모의실험 데이터

		B		합계
		1	0	
A	1	$50+u$	$20+w$	$70+u+w$
	0	$5-w$	$25-u$	$30-u-w$
합계		$55+u-w$	$45-u+w$	100

이로부터 u와 w의 변화에 따른 지지도와 표준화된 지지도를 계산하여 그 일부를 표 3.2에 제시하였다. 여기서 $L.B._S$와 $U.B._S$는 각각 지지도의 하한 및 상한을 의미하고, $Range_S$는 지지도의 범위를 의미한다.

표 3.2 셀 빈도수에 따른 지지도와 표준화 지지도의 변화

n_{11}	n_{10}	n_{01}	n_{00}	$P(A)$	$P(B)$	$L.B._S$	$U.B._S$	$Range_S$	$S(A\Rightarrow B)$	$S_{ST}(A\Rightarrow B)$
50	24	1	25	0.7400	0.5100	0.2500	0.5100	0.2600	0.5000	0.9615
51	24	1	24	0.7500	0.5200	0.2700	0.5200	0.2500	0.5100	0.9600
52	24	1	23	0.7600	0.5300	0.2900	0.5300	0.2400	0.5200	0.9583
53	24	1	22	0.7700	0.5400	0.3100	0.5400	0.2300	0.5300	0.9565
54	24	1	21	0.7800	0.5500	0.3300	0.5500	0.2200	0.5400	0.9545
55	24	1	20	0.7900	0.5600	0.3500	0.5600	0.2100	0.5500	0.9524
56	24	1	19	0.8000	0.5700	0.3700	0.5700	0.2000	0.5600	0.9500
57	24	1	18	0.8100	0.5800	0.3900	0.5800	0.1900	0.5700	0.9474
58	24	1	17	0.8200	0.5900	0.4100	0.5900	0.1800	0.5800	0.9444
59	24	1	16	0.8300	0.6000	0.4300	0.6000	0.1700	0.5900	0.9412
60	24	1	15	0.8400	0.6100	0.4500	0.6100	0.1600	0.6000	0.9375
61	24	1	14	0.8500	0.6200	0.4700	0.6200	0.1500	0.6100	0.9333
62	24	1	13	0.8600	0.6300	0.4900	0.6300	0.1400	0.6200	0.9286
50	23	2	25	0.7300	0.5200	0.2500	0.5200	0.2700	0.5000	0.9259
51	23	2	24	0.7400	0.5300	0.2700	0.5300	0.2600	0.5100	0.9231
63	24	1	12	0.8700	0.6400	0.5100	0.6400	0.1300	0.6300	0.9231
52	23	2	23	0.7500	0.5400	0.2900	0.5400	0.2500	0.5200	0.9200
53	23	2	22	0.7600	0.5500	0.3100	0.5500	0.2400	0.5300	0.9167
64	24	1	11	0.8800	0.6500	0.5300	0.6500	0.1200	0.6400	0.9167
54	23	2	21	0.7700	0.5600	0.3300	0.5600	0.2300	0.5400	0.9130
55	23	2	20	0.7800	0.5700	0.3500	0.5700	0.2200	0.5500	0.9091
65	24	1	10	0.8900	0.6600	0.5500	0.6600	0.1100	0.6500	0.9091
56	23	2	19	0.7900	0.5800	0.3700	0.5800	0.2100	0.5600	0.9048

이 표에서 보는 바와 같이 두 항목의 발생확률 $P(A)$와 $P(B)$가 증가할수록 $S_{ST}(A\Rightarrow B)$는 감소하고 있는 반면에 $S(A\Rightarrow B)$는 그렇지 않다는 것을 알 수 있다. 이를 좀 더 구체적으로 살펴보면 $n_{11}=54$, $n_{10}=24$, $n_{01}=1$, $n_{00}=21$인 경우와 $n_{11}=56$, $n_{10}=24$, $n_{01}=1$, $n_{00}=19$인 경우에 $S(A\Rightarrow B)$와 $S_{ST}(A\Rightarrow B)$의 값은 각각 0.54와 0.9545 및 0.56과 0.9500으로 나타났다. 따라서 $P(A)$와 $P(B)$가 큰 경우의 $S(A\Rightarrow B)$는 $P(A)$와 $P(B)$가 작은 경우의 $S(A\Rightarrow B)$값은 크게 되는 반면에 $S_{ST}(A\Rightarrow B)$는 정반대로 나타나고 있다. 그 이유는 $S_{ST}(A\Rightarrow B)$는 $Range_S$의 값에 영향을 받기 때문인 것으로 생각된다. 또한 위의 두 경우와 $n_{11}=51$, $n_{10}=23$, $n_{01}=2$, $n_{00}=24$인 경우를 살펴보아도 위와 같은 결과를 얻게 된다. 그리고 $S_{ST}(A\Rightarrow B)$는 그 값이 취할 수 있는 범위를 계산하는 데 각 항목의 발생확률을 고려하였으므로 보다 객관적이고 정확한 측도라고 할 수 있다. 따라서 $S(A\Rightarrow B)$보다는 $S_{ST}(A\Rightarrow B)$을 연관성 평가기준으로 사용하는 것이 더 바람직하다.

표 3.1로부터 u와 w의 변화에 따른 신뢰도와 표준화된 신뢰도를 계산하여 그 일부를 표 3.3에 제시하였다. 여기서 $L.B._C$와 $U.B._C$는 각각 신뢰도의 하한 및 상한을 의미하고, $Range_C$는 신뢰도의 범위를

의미한다.

표 3.3 셀 빈도수에 따른 신뢰도와 표준화 신뢰도의 변화

n_{11}	n_{10}	n_{01}	n_{00}	$P(A)$	$P(B)$	$S(A \Rightarrow B)$	$L.B._C$	$U.B._C$	$Range_C$	$C(A \Rightarrow B)$	$C_{ST}(A \Rightarrow B)$
61	24	1	14	0.8500	0.6200	0.6100	0.5529	0.7294	0.1765	0.7176	0.9333
62	24	1	13	0.8600	0.6300	0.6200	0.5698	0.7326	0.1628	0.7209	0.9286
50	23	2	25	0.7300	0.5200	0.5000	0.3425	0.7123	0.3699	0.6849	0.9259
51	23	2	24	0.7400	0.5300	0.5100	0.3649	0.7162	0.3514	0.6892	0.9231
63	24	1	12	0.8700	0.6400	0.6300	0.5862	0.7356	0.1494	0.7241	0.9231
52	23	2	23	0.7500	0.5400	0.5200	0.3867	0.7200	0.3333	0.6933	0.9200
53	23	2	22	0.7600	0.5500	0.5300	0.4079	0.7237	0.3158	0.6974	0.9167
64	24	1	11	0.8800	0.6500	0.6400	0.6023	0.7386	0.1364	0.7273	0.9167
54	23	2	21	0.7700	0.5600	0.5400	0.4286	0.7273	0.2987	0.7013	0.9130
55	23	2	20	0.7800	0.5700	0.5500	0.4487	0.7308	0.2821	0.7051	0.9091
65	24	1	10	0.8900	0.6600	0.6500	0.6180	0.7416	0.1236	0.7303	0.9091
56	23	2	19	0.7900	0.5800	0.5600	0.4684	0.7342	0.2658	0.7089	0.9048
66	24	1	9	0.9000	0.6700	0.6600	0.6333	0.7444	0.1111	0.7333	0.9000
57	23	2	18	0.8000	0.5900	0.5700	0.4875	0.7375	0.2500	0.7125	0.9000
58	23	2	17	0.8100	0.6000	0.5800	0.5062	0.7407	0.2346	0.7160	0.8947
50	22	3	25	0.7200	0.5300	0.5000	0.3472	0.7361	0.3889	0.6944	0.8929
67	24	1	8	0.9100	0.6800	0.6700	0.6484	0.7473	0.0989	0.7363	0.8889
59	23	2	16	0.8200	0.6100	0.5900	0.5244	0.7439	0.2195	0.7195	0.8889
51	22	3	24	0.7300	0.5400	0.5100	0.3699	0.7397	0.3699	0.6986	0.8889
52	22	3	23	0.7400	0.5500	0.5200	0.3919	0.7432	0.3514	0.7027	0.8846
60	23	2	15	0.8300	0.6200	0.6000	0.5422	0.7470	0.2048	0.7229	0.8824
53	22	3	22	0.7500	0.5600	0.5300	0.4133	0.7467	0.3333	0.7067	0.8800
68	24	1	7	0.9200	0.6900	0.6800	0.6630	0.7500	0.0870	0.7391	0.8750

이 표에서 보는 바와 같이 두 항목의 발생확률 $P(A)$ 또는 $P(B)$가 증가할수록 $C(A \Rightarrow B)$는 증가하고 $C_{ST}(A \Rightarrow B)$는 감소하고 있으며, $P(A \text{ and } B)$가 증가할수록 $C(A \Rightarrow B)$는 증가하고 있으며, $C_{ST}(A \Rightarrow B)$는 대체적으로 감소하고 있다. $P(A \text{and} B)$가 증가할수록 간혹 $C_{ST}(A \Rightarrow B)$가 매우 적게 감소하는 경우가 나타나는 데, 이는 범위의 크기가 영향을 준 것으로 생각된다. 이를 좀 더 구체적으로 알아보기 위해 $n_{11} = 50$, $n_{10} = 22$, $n_{01} = 3$, $n_{00} = 25$인 경우와 $n_{11} = 54$, $n_{10} = 23$, $n_{01} = 2$, $n_{00} = 21$인 경우를 비교해보면 $C(A \Rightarrow B)$와 $C_{ST}(A \Rightarrow B)$의 값은 각각 0.6944와 0.8929 및 0.7013과 0.9130으로 나타났다. 따라서 이 두 측도는 공히 $P(A)$ 또는 $P(B)$가 작은 경우의 두 측도의 값이 큰 경우에 비해 큰 값을 가지며, $P(A \text{ and } B)$의 값이 큰 경우의 두 측도의 값이 작은 경우에 비해 큰 값을 가진다. 따라서 두 측도 모두 연관성 평가기준을 잘 충족하고 있으나 두 항목발생확률을 고려한 표준화된 측도인 $C_{ST}(A \Rightarrow B)$을 사용함으로써 연관성 정도를 보다 객관적으로 평가할 수 있다.

또한 표 3.1로부터 u와 w의 변화에 따른 지지도와 표준화된 향상도를 계산하여 그 일부를 표 3.4에 제시하였다. 여기서 $L.B._L$과 $U.B._L$은 각각 향상도의 하한 및 상한을 의미하고, $Range_L$는 향상도의 범위를 의미한다.

이 표에서 보는 바와 같이 두 항목의 발생확률 $P(A)$ 또는 $P(B)$가 증가할수록 $L(A \Rightarrow B)$와 $L_{ST}(A \Rightarrow B)$는 감소하고 있으며, $P(A \text{ and } B)$가 증가할수록 $L(A \Rightarrow B)$는 감소하고 $L_{ST}(A \Rightarrow B)$는 대체적으로 감소하고 있다. 이 경우에도 $P(A \text{ and } B)$가 증가할수록 간혹 $L_{ST}(A \Rightarrow B)$가 매우 적게 감소하는 경우가 나타나는 데, 이도 역시 범위가 영향을 준 것으로 생각된다. 이를 좀 더 구체적으로 알아보기 위해 $n_{11} = 61$, $n_{10} = 24$, $n_{01} = 1$, $n_{00} = 14$인 경우와 $n_{11} = 62$, $n_{10} = 24$, $n_{01} = 1$, $n_{00} = 13$인 경우를 비교해보면 $L(A \Rightarrow B)$와 $L_{ST}(A \Rightarrow B)$의 값은 각각 1.1575와 0.9333 및 1.1443과 0.9286으로 나타났다. 따라서 이 두 측도는 공히 $P(A)$ 또는 $P(B)$가 큰 경우의 두 측도의 값이 작은

표 3.4 셀 빈도수에 따른 향상도와 표준화 향상도의 변화

n_{11}	n_{10}	n_{01}	n_{00}	$P(A)$	$P(B)$	$S(A \Rightarrow B)$	$L.B._L$	$U.B._L$	$Range_L$	$L(A \Rightarrow B)$	$L_{ST}(A \Rightarrow B)$
60	24	1	15	0.8400	0.6100	0.6000	0.8782	1.1905	0.3123	1.1710	0.9375
61	24	1	14	0.8500	0.6200	0.6100	0.8918	1.1765	0.2846	1.1575	0.9333
62	24	1	13	0.8600	0.6300	0.6200	0.9044	1.1628	0.2584	1.1443	0.9286
50	23	2	25	0.7300	0.5200	0.5000	0.6586	1.3699	0.7113	1.3172	0.9259
51	23	2	24	0.7400	0.5300	0.5100	0.6884	1.3514	0.6629	1.3004	0.9231
63	24	1	12	0.8700	0.6400	0.6300	0.9159	1.1494	0.2335	1.1315	0.9231
52	23	2	23	0.7500	0.5400	0.5200	0.7160	1.3333	0.6173	1.2840	0.9200
64	24	1	11	0.8800	0.6500	0.6400	0.9266	1.1364	0.2098	1.1189	0.9167
53	23	2	22	0.7600	0.5500	0.5300	0.7416	1.3158	0.5742	1.2679	0.9167
54	23	2	21	0.7700	0.5600	0.5400	0.7653	1.2987	0.5334	1.2523	0.9130
55	23	2	20	0.7800	0.5700	0.5500	0.7872	1.2821	0.4948	1.2371	0.9091
65	24	1	10	0.8900	0.6600	0.6500	0.9363	1.1236	0.1873	1.1066	0.9091
56	23	2	19	0.7900	0.5800	0.5600	0.8075	1.2658	0.4583	1.2222	0.9048
66	24	1	9	0.9000	0.6700	0.6600	0.9453	1.1111	0.1658	1.0945	0.9000
57	23	2	18	0.8000	0.5900	0.5700	0.8263	1.2500	0.4237	1.2076	0.9000
58	23	2	17	0.8100	0.6000	0.5800	0.8436	1.2346	0.3909	1.1934	0.8947
50	22	3	25	0.7200	0.5300	0.5000	0.6551	1.3889	0.7338	1.3103	0.8929
67	24	1	8	0.9100	0.6800	0.6700	0.9535	1.0989	0.1454	1.0827	0.8889
59	23	2	16	0.8200	0.6100	0.5900	0.8597	1.2195	0.3599	1.1795	0.8889
51	22	3	24	0.7300	0.5400	0.5100	0.6849	1.3699	0.6849	1.2938	0.8889
52	22	3	23	0.7400	0.5500	0.5200	0.7125	1.3514	0.6388	1.2776	0.8846
60	23	2	15	0.8300	0.6200	0.6000	0.8745	1.2048	0.3304	1.1660	0.8824
53	22	3	22	0.7500	0.5600	0.5300	0.7381	1.3333	0.5952	1.2619	0.8800

경우에 비해 작은 값을 가진다. 또한 $n_{11} = 65$, $n_{10} = 24$, $n_{01} = 1$, $n_{00} = 10$인 경우와 $n_{11} = 56$, $n_{10} = 23$, $n_{01} = 2$, $n_{00} = 19$인 경우를 비교해보면 $L(A \Rightarrow B)$와 $L_{ST}(A \Rightarrow B)$의 값은 각각 1.1066과 0.9091 및 1.2222과 0.9048로 나타났다. 따라서 이 두 측도는 공히 $P(A \text{ and } B)$가 큰 경우의 두 측도의 값이 작은 경우에 비해 큰 값을 가진다. 이 경우에도 두 측도 모두 연관성 평가기준을 잘 충족하고 있으나 두 항목발생확률을 고려한 0과 1 사이의 값을 가지는 표준화된 측도인 $L_{ST}(A \Rightarrow B)$을 사용함으로써 연관성 정도를 보다 객관적으로 평가할 수 있다.

4. 결론

본 논문에서는 데이터 마이닝 분야에서 가장 많이 활용되고 있는 연관성 규칙에 대해 그 평가기준의 표준화 방안에 대해 연구하였다. 특히 McNicholas 등 (2008)의 연구와 연관성 측도들 간의 관계식을 고려하여 또 다른 표준화 향상도를 제시한 후, 지도와 신뢰도에 대해서도 표준화하였다. 표준화를 함으로써 연관성 평가기준값들이 모두 0과 1 사이의 값을 갖게 되는 동시에 연관성 규칙들의 중요도를 보다 객관적이고도 정확하게 평가할 수 있었다. 또한 흥미도 측도가 되기 위한 조건들을 점검한 결과, 일반적인 경우에는 대부분 조건을 충족하는 것으로 나타났으나, 조건부 확률의 값이 주변확률값보다 작은 경우에는 그렇지 않다는 사실도 발견하였다.

모의실험을 통해 알아 본 결과, 두 항목의 발생확률이 증가할수록 표준화 지지도는 감소하고 있는 반면에 지지도는 그렇지 않다는 것을 알 수 있었다. 신뢰도와 표준화 신뢰도도 두 항목의 발생확률이 증가할수록 감소하고 있으며, 동시발생확률이 증가할수록 신뢰도는 증가하였으며, 표준화 신뢰도인 경우에는 대체적으로 증가하고 있다. 동시발생확률이 증가할수록 간혹 표준화 신뢰도가 미약하게나마 감소하는 경우가 나타나는 데, 이는 신뢰도의 범위가 영향을 준 것으로 생각된다. 또한 표준화 신뢰도는 그 값이 취할 수 있는 범위를 계산하는 데 각 항목의 발생확률을 고려하였으므로 보다 객관적이고 정확한

측도라고 할 수 있다. 그리고 향상도와 표준화 향상도도 두 항목의 발생확률이 증가할수록 감소하였으며, 동시발생확률이 증가할수록 향상도는 증가하였으며, 표준화 향상도는 대체적으로 증가하는 경향을 나타내었다. 이 경우에도 두 측도 모두 연관성 평가기준을 잘 충족하고 있으나 두 항목발생확률을 고려한 0과 1 사이의 값을 가지는 표준화된 향상도를 사용함으로써 연관성 정도를 보다 객관적으로 평가할 수 있을 것으로 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. *Processing 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. (2007). Association rule mining by environmental data fusion. *Journal of the Korean Data & Information Science Society*, **18**, 279-287.

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.

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. (2008). The proposition of conditionally pure confidence in association rule mining. *Journal of the Korean Data & Information Science Society*, **19**, 1141-1151.

Park, H. C. and Cho, K. H. (2005). Waste database analysis joined with local information using association rules. *Journal of the Korean Data Analysis Society*, **7**, 763-772.

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.

Piatetsky, S. G. (1991). Discovery, analysis and presentation of strong rules. *Knowledge Discovery in Databases*, AAAI/MIT Press, 229-248.

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
2010, **21**(5), 891–899

한국데이터정보과학회지

Standardization for basic association measures in association rule mining

Hee Chang Park[1]

[1]Department of Statistics, Changwon National University

Received 19 July 2010, revised 30 August 2010, accepted 6 September 2010

Abstract

Association rule is the technique to represent the relationship between two or more items by numerical representing for the relevance of each item in vast amounts of databases, and is most being used in data mining. The basic thresholds for association rule are support, confidence, and lift. these are used to generate the association rules. We need standardization of lift because the range of lift value is different from that of support and confidence. And also we need standardization of support and confidence to compare objectively association level of antecedent variables for one descendant variable. In this paper we propose a method for standardization of association thresholds considering marginal probability for each item to grasp objectively and exactly association level, check the conditions for association criteria and then compare association thresholds with standardized association thresholds using some concrete examples.

Keywords: Association rule, confidence, lift, standardized threshold, support.

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