Journal of the Korean Data & Information Science Society
2013, **24**(3), 523–532

http://dx.doi.org/10.7465/jkdi.2013.24.3.523
한국데이터정보과학회지

연관 규칙 마이닝에서 비교 기여 순수 신뢰도의 제안

박희창[1]

[1]창원대학교 통계학과

접수 2013년 4월 15일, 수정 2013년 5월 3일, 게재확정 2013년 5월 11일

요 약

데이터 마이닝은 빅 데이터에 잠재되어 있는 지식이나 패턴을 찾아내는 기술이며, 대표적인 기법 중의 하나가 연관성 규칙 마이닝이다. 이 기법은 지지도, 신뢰도, 향상도 등의 연관성 평가 기준을 기반으로 하여 각 항목들 간의 관련성을 찾아내는 데 활용되고 있다. 연관성을 평가하기 위한 기준으로 여러 가지 흥미도 측도가 개발되어 있는데, 그 중에서도 신뢰도가 가장 많이 활용되고 있으나 연관성의 방향을 알 수가 없다는 단점을 가지고 있다. 이를 보완하기 위한 측도로 순수 신뢰도가 개발되었으나. 양의 신뢰도과 음의 신뢰도의 값이 동일한 경우에는 이 측도의 값이 같아지므로 정확한 연관성 규칙을 발견할 수 없게 된다. 이러한 단점을 보완하기 위해 기여 순수 신뢰도와 비교 신뢰도가 개발되었는데 이들은 이들 측도들이 취할 수 있는 값의 범위에 대한 문제를 제외하고는 흥미도 측도로서는 매우 바람직하다고 할 수 있으나 값의 범위에 대한 문제점이 존재한다. 이 문제를 해결하기 위해 본 논문에서는 기여 순수 신뢰도와 비교 신뢰도의 크기를 동시에 고려한 비교 기여 순수 신뢰도를 제안하였다. 또한 예제를 통하여 그 유용성을 알아본 결과, 비교 기여 순수 신뢰도는 그 부호에 의해 연관성 규칙의 방향을 파악할 수 있는 동시에 그 값의 범위가 [-1, +1]의 값을 가지므로 행태적 해석이 가능한 것으로 확인되었다.

주요용어: 기여 순수 신뢰도, 비교 기여 순수 신뢰도, 비교 신뢰도, 신뢰도, 연관성 평가 기준.

1. 서론

오늘날 기업이나 조직 간의 경쟁이 심화되고 정보의 중요성에 대한 인식이 확산됨에 따라 빅 데이터(big data)에서 유용한 정보를 캐내는 데이터 마이닝 (data mining) 기법이 주목 받고 있다. 데이터 마이닝은 빅 데이터에 함축적으로 들어 있는 지식이나 패턴을 찾아내는 기술이며, 대표적인 기법이 연관성 규칙 마이닝 (association rule mining)이다. 이는 빅 데이터로부터 항목들 간에 특정한 연관성을 발견하는 것으로 다양한 분야에서 많이 활용되고 있으며, Agrawal 등 (1993)에 의해 처음 소개되었다. 이후로 많은 학자들이 연관성 측정에 관한 연구를 수행하였는데, 이들은 크게 제약조건을 가지는 항목으로 구성된 트랜잭션 데이터베이스에서 빈발항목을 찾는 연구와 연관성 규칙 생성에 대한 수행속도를 향상시키기 위한 연구로 분류할 수 있다. 전자에 관한 대표적인 연구로는 Han과 Fu (1995), Srikant 등 (1997), Cai 등 (1998), Liu 등 (1999) 등이 있으며, 후자에 관한 연구로는 Bayardo (1998), Pasquier 등 (1999), Han 등 (2000), Pei 등 (2000), Toivonen (1996) 등이 있다. 특히 연관성 규칙에 대한 최근의 국내 연구로는 Cho와 Park (2011a, 2011b), Jin 등 (2011), Park (2010a, 2010b, 2011a, 2011b, 2011c) 등이 있다.

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

연관성 규칙 마이닝에서는 항목들 간의 여러 가지 흥미도 측도 (interestingness measure)의 값을 근거로 하여 관련성 여부를 측정한다. 의미 있는 연관성 규칙을 탐색하기 위한 흥미도 측도에는 사용자 관점에서 해석 가능하도록 제안된 주관적 흥미도 측도와 논리적인 또는 통계적인 방법에 의해 제안된 것으로 사용자에게 규칙을 정제할 수 있는 근거를 제시해주는 객관적 흥미도 측도가 있다 (Silberschatz와 Tuzhilin, 1996; Freitas, 1999). 많은 학자들에 의해 흥미도 측도에 관한 연구가 수행되었는데, 대표적인 연구로는 객관적 흥미도 측도들을 데이터 마이닝에 응용한 Hilderman과 Hamilton (2000)의 연구와 주관적 흥미도 측도를 연관성 규칙에 적용한 바 있는 Liu 등 (2000)의 연구가 있다. 또한 Tan 등 (2002)은 여러 가지 흥미도 측도들 가운데서 올바른 선택방안에 대해 제안한 바 있다.

본 논문에서는 기존에 많이 활용되고 있는 흥미도 측도인 신뢰도 (confidence)와 순수 신뢰도 (net confidence), 그리고 기여 순수 신뢰도 (attributably pure confidence)의 단점을 보완한 비교 기여 순수 신뢰도 (compared and attributably pure confidence)를 제안하고자 한다. 비교 기여 순수 신뢰도는 기여 순수 신뢰도와 비교 신뢰도의 크기를 동시에 고려한 것으로 양의 신뢰도와 음의 신뢰도의 크기를 상대적으로 비교해서 나타낸 흥미도 측도인 동시에 값의 범위가 [-1, +1]이 되는 측도이다. 본 논문의 2절에서는 제안하는 흥미도 측도인 비교 기여 순수 신뢰도를 정의한 후 여러 가지 특성을 살펴보는 동시에 Piatetsky-Shapiro (1991)가 제안한 흥미도 측도의 기준에 대한 충족여부를 점검한다. 3절에서는 예제를 통하여 기존의 신뢰도와 순수 신뢰도, 그리고 기여 순수 신뢰도와의 비교를 통해 비교 기여 순수 신뢰도의 유용성에 대해 알아본 후, 4절에서 결론을 내리고자 한다.

2. 비교 기여 순수 신뢰도

연관성 규칙을 평가하는 기준에는 지지도, 신뢰도, 향상도 등이 있으며, Ahn과 Kim (2003), Kuo (2009), 그리고 Park (2012)의 연구 결과와 비교하기 위해 Table 2.1과 같은 분할표를 고려한다. 지지도 $S(X \Rightarrow Y)$는 항목 X와 항목 Y가 동시에 발생하는 거래의 비율을 의미하며, 위의 표로부터 n_{11}/n으로 계산된다. 신뢰도 $C(X \Rightarrow Y)$는 항목 X가 포함된 거래 비율 중 항목 X와 항목 Y가 동시에 포함된 거래의 비율을 의미하며, $n_{11}/n_{1.}$으로 얻어진다. 또한 향상도 $L(X \Rightarrow Y)$는 항목 X를 구매한 경우 그 거래가 항목 Y를 포함하는 경우와 항목 Y가 임의로 구매되는 경우의 비를 의미하며, $(n_{11} \cdot n)/(n_{1.} \cdot n_{.1})$으로 계산된다. 여기서 신뢰도는 항목 X를 포함하는 거래 중에서 항목 Y가 포함될 확률이 어느 정도인지를 확인하는 기준이 될 수 있으므로 관련성 규칙의 예측 지표라고 볼 수 있다. 그러나 신뢰도는 계산된 값만을 가지고는 양의 연관성을 가지는지 음의 연관성을 가지는지를 알 수 없을 뿐만 아니라 신뢰도만으로는 음의 연관성을 가지는 연관성 규칙을 의미 있는 양의 관계를 가지는 규칙으로 선택하게 되는 오류를 범할 수 있다.

Table 2.1 2×2 contingency table

		Y		Total
		1	0	
X	1	n_{11}	n_{10}	n_{1+}
	0	n_{01}	n_{00}	n_{0+}
Total		n_{+1}	n_{+0}	n

이러한 문제를 해결하기 위해 Ahn과 Kim (2003)은 의학 분야에서 널리 이용되고 있는 기여위험률 (attributable risk)을 순수 신뢰도 (net confidence ; $Nconf$)라는 이름으로 데이터 마이닝 분야에 적용

한 바 있다.

$$Nconf(X \Rightarrow Y) = P(Y|X) - P(Y|\overline{X}) = \frac{n_{11} \cdot n_{00} - n_{10} \cdot n_{01}}{n_{1.} \cdot n_{0.}} \tag{2.1}$$

여기서 $\overline{X}$의 의미는 X가 일어나지 않음을 의미한다. 이러한 순수 신뢰도는 순수하게 특정 요인에 의해서만 결과가 얼마인가를 나타내주는 측도이며, 부호에 의해 양의 관련성과 음의 관련성을 판단할 수 있기는 하나, $P(Y|X)$와 $P(Y|\overline{X})$의 값이 어떤 값을 가지더라도 두 값의 차이가 동일하면 순수 신뢰도의 값도 동일하게 되는 단점을 가지고 있다. 이러한 문제를 보완하기 위해 Park (2012)은 다음과 같은 기여 순수 신뢰도 ($APconf$)를 제안하였다.

$$APconf(X \Rightarrow Y) = \frac{P(Y|X) - P(Y|\overline{X})}{P(Y|X)} = \frac{n_{11} \cdot n_{00} - n_{10} \cdot n_{01}}{n_{11} \cdot n_{0.}} \tag{2.2}$$

이 측도는 의학 분야에서 적용되고 있는 기여 분율 (attributable fraction)을 연관성 규칙의 평가기준에 적합하도록 변형한 것이다. 위의 표에서 기여 순수 신뢰도를 계산하면 각각 0.4와 0.25로 나타났다. 따라서 기여 순수 신뢰도는 $P(Y|X)$와 $P(Y|\overline{X})$의 차이 크기를 $P(Y|X)$에 대해 상대적으로 나타낸 것으로 이를 이용하면 $P(Y|X)$와 $P(Y|\overline{X})$의 크기를 반영할 수 있게 된다. 그러나 Park (2012)이 밝힌 바와 같이 이 측도의 범위는 $(-\infty, 1]$이므로 흥미도 측도로서는 바람직하지 않다. 또한 Kuo (2009)는 다음과 같은 비교 신뢰도 (compared confidence)를 제안한 바 있다.

$$Cconf(X \Rightarrow Y) = \frac{P(Y|X) - P(Y|\overline{X})}{P(Y|\overline{X})} \tag{2.3}$$

이 측도는 식 (2.2)의 기여 순수 신뢰도의 분모인 $P(Y|X)$를 $P(Y|\overline{X})$으로 한 측도이며, 이 측도의 범위는 $[-1, \infty)$으로 나타나서 이 측도 역시 흥미도 측도로서는 바람직하지 않다고 할 수 있다.

한편, 기여 순수 신뢰도 ($APconf$)는 순수 신뢰도를 양의 신뢰도인 $P(Y|X)$과 비교한 것이고, 비교 신뢰도는 순수 신뢰도를 음의 신뢰도인 $P(Y|\bar{X})$과 비교한 것이다. 이들 측도들이 취할 수 있는 값의 범위에 대한 문제를 제외하고는 흥미도 측도로서는 매우 바람직하다고 할 수 있다. 따라서 본 논문에서는 기여 순수 신뢰도와 비교 신뢰도를 동시에 고려하면서 행태적 해석 (operational interpretation)이 가능할 수 있도록 [-1, +1]의 범위를 가지는 측도를 다음과 같이 개발하였으며, 이를 비교 기여 순수 신뢰도라고 명명하였다.

$$CAPconf(X \Rightarrow Y) = \begin{cases} \dfrac{P(Y|X) - P(Y|\overline{X})}{P(Y|X)}, \text{ if } P(Y|X) > P(Y|\overline{X}), P(Y|X) \neq 0 \\ \dfrac{P(Y|X) - P(Y|\overline{X})}{P(Y|\overline{X})}, \text{ if } P(Y|X) < P(Y|\overline{X}), P(Y|\overline{X}) \neq 0 \end{cases} \tag{2.4}$$

Park (2011b)은 기여 순수 신뢰도 $APconf$이 Piatetsky-Shapiro (1991)가 제안한 흥미도 측도의 조건을 충족하는지의 여부를 증명한 바 있으며, 그 성질들을 다음과 같이 규명하였다.

성질 2.1 $APconf > 0$이면 X와 Y가 양의 연관성을 가지고, $APconf < 0$이면 X와 Y가 음의 연관성을 가지며, $APconf = 0$이면 X와 Y가 서로 독립관계임을 의미한다.

성질 2.2 일반적으로 $APconf(X \Rightarrow Y)$와 $APconf(Y \Rightarrow X)$의 값은 동일하지 않다.

성질 2.3 $APconf(X \Rightarrow Y) = APconf(Y \Rightarrow X)$이면 항목 X와 Y는 서로 독립이다.

성질 2.4 $APconf(X \Rightarrow Y)$값의 범위는 $[-\infty, 1]$이다.

Kuo (2009)는 비교 신뢰도를 제안한 후 예제를 통하여 그 의미는 기술하였으나 흥미도 측도의 타당성을 의미하는 조건 충족 여부에 대해서는 증명하지 않았으므로 여기서 Piatetsky- Shapiro (1991)가 제안한 흥미도 측도의 조건을 충족하는지의 여부를 증명하면 다음과 같다.

조건 2.1 $P(X \text{ and } Y) = P(X)P(Y)$이면 $Cconf$의 값은 0이 된다.

(증명) : 먼저 식 (2.3)의 $Cconf(X \Rightarrow Y)$를 정리하면 다음과 같이 표현된다.

$$Cconf(X \Rightarrow Y) = \frac{P(X \text{ and } Y) - P(X)P(Y)}{P(X)P(Y) - P(X)P(X \text{ and } Y)} \tag{2.5}$$

$P(X \text{ and } Y) = P(X)P(Y)$이면 $Cconf$의 분자의 값이 0이므로 $Cconf$의 값은 0이 된다.

조건 2.2 $Cconf$는 $P(Y)$의 값에 따라 단조 감소한다.

(증명) : 식 (2.6)의 $Cconf$를 다시 한 번 정리하면 다음과 같이 나타낼 수 있다.

$$Cconf(X \Rightarrow Y) = \frac{P(X \text{ and } Y)}{P(Y) - P(X \text{ and } Y)} \cdot \frac{1 - P(X)}{P(X)} \tag{2.6}$$

이로부터 $P(Y)$의 값이 증가함에 따라 $Cconf$는 단조 감소하는 것을 알 수 있다.

조건 2.3 $Cconf$는 $P(X \text{ and } Y)$의 값에 따라 단조 증가한다.

(증명) : 식 (2.6)으로부터 $P(X \text{ and } Y)$의 값이 증가함에 따라 $Cconf$는 단조 증가하는 것을 알 수 있다.

또한 비교 신뢰도가 가지는 성질을 규명하면 다음과 같다.

성질 2.5 $Cconf > 0$이면 X와 Y가 양의 연관성을 가지고, $Cconf < 0$이면 X와 Y가 음의 연관성을 가지며, $Cconf = 0$이면 X와 Y가 서로 독립관계임을 의미한다.

(설명) : $Cconf > 0$가 되기 위해서는 순수 신뢰도와 마찬가지로 식 (2.3)의 분자의 값이 양의 값이 되어야 하므로, 즉 $P(Y|X) > P(Y|\bar{X})$이므로 항목 X가 포함된 트랜잭션에서 항목 Y가 발견되는 확률이 항목 X가 포함되지 않은 트랜잭션에서 항목 Y가 발견되는 확률보다 커야 한다. 따라서 $Cconf > 0$이면 이들 두 항목 간에는 양의 연관성이 존재한다고 볼 수 있다. 만약 $Cconf < 0$이면 이 또한 순수 신뢰도와 마찬가지로 항목 X가 포함된 트랜잭션에서 항목 Y가 발견되는 확률이 항목 X가 포함되지 않은 트랜잭션에서 항목 Y가 발견되는 확률보다 작으므로 이들 두 항목 간에는 음의 연관성이 존재한다고 볼 수 있다. 그리고 $Cconf = 0$이면 항목 Y는 항목 X가 포함된 트랜잭션에서뿐만 아니라 X가 발견되지 않는 트랜잭션에서도 동일한 확률로 발견되므로 X와 Y가 서로 독립관계라고 볼 수 있다.

성질 2.6 일반적으로 $Cconf(X \Rightarrow Y)$와 $Cconf(Y \Rightarrow X)$의 값은 동일하지 않다.

(설명) : Park (2012)에서 기술한 바와 같이 연관성 규칙의 흥미도 측도는 방향성을 갖는 것이 바람직한데 (Berzal 등, 2004), 식 (2.3)을 통해서 알 수 있는 바와 같이 $Cconf$는 $Y \Rightarrow X$와 $X \Rightarrow Y$의 값이 다르므로 방향성이 고려된다.

성질 2.7 $Cconf(X \Rightarrow Y) = Cconf(Y \Rightarrow X)$이면 항목 X와 Y는 서로 독립이다.

(설명) : $Cconf(X \Rightarrow Y) = Cconf(Y \Rightarrow X)$이기 위해서는 다음의 식을 만족하여야 한다.

$$\frac{P(Y|X) - P(Y|\overline{X})}{P(Y|\overline{X})} = \frac{P(X|Y) - P(X|\overline{Y})}{P(X|\overline{Y})}. \tag{2.7}$$

식 (2.7)을 정리하면 $P(X \text{ and } Y) = P(X)P(Y)$가 되어 두 항목 X와 Y는 서로 독립관계에 놓이게 된다. 따라서 $Cconf(X \Rightarrow Y) = Cconf(Y \Rightarrow X)$이면 항목 X와 Y는 서로 독립이 된다.

성질 2.8 $Cconf(X \Rightarrow Y)$값의 범위는 $[-1, \infty)$이다.
(설명) : 식 (2.3)을 정리하면 다음과 같이 나타낼 수 있다.

$$Cconf(X \Rightarrow Y) = \frac{P(Y|X)}{P(Y|\overline{X})} - 1$$

이 식으로부터 $Cconf(X \Rightarrow Y)$값이 -1이라는 의미는 $P(Y|X) = 0$이므로 항목 X가 발견되는 모든 트랜잭션에서 Y가 전혀 발생하지 않는다는 의미이다. 그리고 $Cconf(X \Rightarrow Y)$의 값이 $+\infty$라는 의미는 항목 X가 발견되는 모든 트랜잭션에서 항목 Y가 발견될 확률이 X가 없는 트랜잭션에서 Y가 발견될 확률보다 훨씬 크다는 의미이다.

3. 예제 데이터의 적용

본 절에서는 신뢰도와 기여 순수 신뢰도, 비교 신뢰도의 문제점을 탐색하고 비교 기여 순수 신뢰도의 유용성을 예제를 통해 고찰하고자 한다. 이를 위해 항목 X, Y에 대해 다음과 같이 가정하였다 (Park, 2012).

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

먼저 데이터베이스에 있는 총 트랜잭션의 수 (t)를 100명으로 하고, 항목 X는 구매한 냉장고의 금액을 기준으로 100만원 이상 (1) 구매한 사람 수를 50명으로 하고 100만원 미만 (0)을 구매한 사람 수를 50명으로 하였다. 또한 항목 Y를 결제 방식을 기준으로 신용 카드로 결제 (1)한 사람 수를 30명으로 하고 신용 카드 이외의 방법으로 결제 (0)한 사람의 수를 70명으로 하였다. 항목 X와 Y가 동시에 발생한 빈도 수, 즉 100만원 이상의 냉장고를 구매하면서 신용카드로 결제한 빈도수는 a명으로 하였다. 이를 정리하면 Table 3.1과 같다. 이 표에서 a가 취할 수 있는 범위는 $0 \le a \le 30$이다. Table 3.1로부터 동시발생빈도 (a)에 따른 신뢰도 ($conf_1$, $conf_2$), 기여 순수 신뢰도 ($APconf$), 그리고 비교신뢰도 ($Cconf$), 그리고 비교 기여 순수 신뢰도 ($CAPconf$)를 계산하면 다음의 Table 3.2와 같은 결과를 얻을 수 있다. 여기서 $b = P(X = 1, Y = 0)$, $c = P(X = 0, Y = 1)$, $d = P(X = 0, Y = 0)$이며, $conf_1 = P(Y|X)$이고, $conf_2 = P(X|Y)$을 의미한다. 이에 대한 계산은 미니탭 16의 계산기 기능을 이용하였다. 이 표로부터 알 수 있는 바와 같이 a의 값이 커질수록 신뢰도 $conf_1$과 비교 신뢰도, 기여 순수 신뢰도, 그리고 비교 기여 순수 신뢰도가 증가하고 있다. 그러나 $conf_1$은 항상 양의 값을 취하고 있고, 기여 순수 신뢰도는 양과 음의 값을 취하고는 있으나 음의 값인 경우에는 $-\infty$의 값을 취할 수 있기 때문에 흥미도 측도로서는 바람직하다고 할 수 없다. 비교 신뢰도 역시 양과 음의 값을 취하고는 있으나 이 측도는 양의 값인 경우에 $+\infty$의 값을 취할 수 있어서 흥미도 측도로서는 바람직하다고 할 수 없다. 반면에 비교 기여 순수 신뢰도는 양과 음의 값을 취할 수 있는 동시에 그 값의 범위가 $[-1, +1]$이므로 바람직한 흥미도 측도라고 할 수 있다.

Table 3.2 Output of some association thresholds by simulation data(1)

a	b	c	d	$supp$	$conf_1$	$conf_2$	$APconf$	$Cconf$	$CAPconf$
5	45	25	25	0.050	0.100	0.500	-4.000	-0.800	-0.800
6	44	24	26	0.060	0.120	0.480	-3.000	-0.750	-0.750
7	43	23	27	0.070	0.140	0.460	-2.286	-0.696	-0.696
8	42	22	28	0.080	0.160	0.440	-1.750	-0.636	-0.636
9	41	21	29	0.090	0.180	0.420	-1.333	-0.571	-0.571
10	40	20	30	0.100	0.200	0.400	-1.000	-0.500	-0.500
11	39	19	31	0.110	0.220	0.380	-0.727	-0.421	-0.421
12	38	18	32	0.120	0.240	0.360	-0.500	-0.333	-0.333
13	37	17	33	0.130	0.260	0.340	-0.308	-0.235	-0.235
14	36	16	34	0.140	0.280	0.320	-0.143	-0.125	-0.125
15	35	15	35	0.150	0.300	0.300	0.000	0.000	0.000
16	34	14	36	0.160	0.320	0.280	0.125	0.143	0.125
17	33	13	37	0.170	0.340	0.260	0.235	0.308	0.235
18	32	12	38	0.180	0.360	0.240	0.333	0.500	0.333
19	31	11	39	0.190	0.380	0.220	0.421	0.727	0.421
20	30	10	40	0.200	0.400	0.200	0.500	1.000	0.500
21	29	9	41	0.210	0.420	0.180	0.571	1.333	0.571
22	28	8	42	0.220	0.440	0.160	0.636	1.750	0.636
23	27	7	43	0.230	0.460	0.140	0.696	2.286	0.696

비교 신뢰도, 기여 순수 신뢰도, 그리고 비교 기여 순수 신뢰도의 변화하는 양상을 좀 더 구체적으로 알아보기 위해 이들 값들이 음의 값인 경우를 먼저 살펴보면 $a = 7$, $b = 43$, $c = 23$, $d = 27$인 경우에는 $APconf = -2.286$, $Cconf = -0.696$, $CAPconf = -0.696$가 되어 비교 신뢰도와 비교 기여 순수 신뢰도는 $[-1, +1]$의 범위에 속하는 값을 취하는 반면에 기여 순수 신뢰도는 그 범위를 넘어선 값을 취하고 있다. 이번에는 이들이 양의 값을 가지는 경우를 살펴보면 $a = 21$, $b = 29$, $c = 9$, $d = 41$인 경우 $APconf = 0.571$, $Cconf = 1.333$, $CAPconf = 0.571$이 되어 기여 순수 신뢰도와 비교 기여 순수 신뢰도는 $[-1, +1]$의 범위에 속하는 값을 취하는 반면에 비교 신뢰도는 그 범위를 넘어선 값을 취하고 있다. 따라서 위에서 기술한 바와 같이 비교 기여 순수 신뢰도는 항상 $[-1, +1]$의 값을 가지므로 본 논문에서 비교하는 신뢰도 중에서는 가장 바람직한 흥미도 측도라고 할 수 있다.

이번에는 b의 값의 변화에 따라 신뢰도와 기여 순수 신뢰도, 비교 신뢰도의 문제점을 탐색하고 비교 기여 순수 신뢰도의 유용성을 고찰하고자 한다. 이를 위해 Table 3.3과 같이 각 셀의 값을 바꾸어 실험하였다.

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

Table 3.3에서 b가 취할 수 있는 정수 값의 범위는 $0 \leq b \leq 30$이다. 이 표로부터 각 셀 값의 변화에 따른 신뢰도, 순수 신뢰도, 그리고 조건부 순수 신뢰도를 계산하여 그 일부를 나타내면 다음 Table 3.4와 같은 결과를 얻을 수 있다.

Table 3.4 Output of some association thresholds by simulation data(2)

a	b	c	d	$supp$	$conf_1$	$conf_2$	$APconf$	$Cconf$	$CAPconf$
48	2	22	28	0.480	0.960	0.440	0.542	1.182	0.542
47	3	23	27	0.470	0.940	0.460	0.511	1.043	0.511
46	4	24	26	0.460	0.920	0.480	0.478	0.917	0.478
45	5	25	25	0.450	0.900	0.500	0.444	0.800	0.444
44	6	26	24	0.440	0.880	0.520	0.409	0.692	0.409
43	7	27	23	0.430	0.860	0.540	0.372	0.593	0.372
42	8	28	22	0.420	0.840	0.560	0.333	0.500	0.333
41	9	29	21	0.410	0.820	0.580	0.293	0.414	0.293
40	10	30	20	0.400	0.800	0.600	0.250	0.333	0.250
39	11	31	19	0.390	0.780	0.620	0.205	0.258	0.205
38	12	32	18	0.380	0.760	0.640	0.158	0.188	0.158
37	13	33	17	0.370	0.740	0.660	0.108	0.121	0.108
36	14	34	16	0.360	0.720	0.680	0.056	0.059	0.056
35	15	35	15	0.350	0.700	0.700	0.000	0.000	0.000
34	16	36	14	0.340	0.680	0.720	-0.059	-0.056	-0.056
33	17	37	13	0.330	0.660	0.740	-0.121	-0.108	-0.108
32	18	38	12	0.320	0.640	0.760	-0.188	-0.158	-0.158
31	19	39	11	0.310	0.620	0.780	-0.258	-0.205	-0.205
30	20	40	10	0.300	0.600	0.800	-0.333	-0.250	-0.250
29	21	41	9	0.290	0.580	0.820	-0.414	-0.293	-0.293
28	22	42	8	0.280	0.560	0.840	-0.500	-0.333	-0.333
27	23	43	7	0.270	0.540	0.860	-0.593	-0.372	-0.372
26	24	44	6	0.260	0.520	0.880	-0.692	-0.409	-0.409
25	25	45	5	0.250	0.500	0.900	-0.800	-0.444	-0.444
24	26	46	4	0.240	0.480	0.920	-0.917	-0.478	-0.478
23	27	47	3	0.230	0.460	0.940	-1.043	-0.511	-0.511

이 표에서 나타나는 바와 같이 b의 값이 커질수록 신뢰도 $conf_2$를 제외한 모든 신뢰도의 값이 감소하고 있으나 신뢰도 $conf_1$은 항상 양의 값을 취하고 있어서 흥미도 측도로서는 바람직하다고 할 수 없다. 기여 순수 신뢰도와 비교 신뢰도는 양과 음의 값을 취하고는 있으나 전자는 음의 값인 경우에 $-\infty$의 값을 취하고 있고 후자는 양의 값인 경우에 $+\infty$의 값을 취하고 있어서 흥미도 측도로서 바람직하다고 할 수 없다. 반면에 비교 기여 순수 신뢰도는 양과 음의 값을 취하고 있는 동시에 그 값의 범위가 $[-1,+1]$이므로 위에서 기술한 바와 같이 바람직한 흥미도 측도라고 할 수 있다. 이를 좀 더 구체적으로 알아보기 위해 $a=48$, $b=2$, $c=22$, $d=28$인 경우와 $a=23$, $b=27$, $c=47$, $d=3$인 경우를 살펴보기로 한다. 전자의 경우에는 $APconf=0.542$, $Cconf=1.182$, $CAPconf=0.542$가 되어 기여 순수 신뢰도와 비교 기여 순수 신뢰도는 $[-1,+1]$의 범위에 속하는 값을 취하는 반면에 비교 신뢰도는 그 범위를 넘어선 값을 취하고 있다. 후자의 경우에는 $APconf=-1.043$, $Cconf=-0.511$, $CAPconf=-0.511$이 되어 비교 신뢰도와 비교 기여 순수 신뢰도의 값은 $[-1,+1]$의 범위에 속하는 반면에 기여 순수 신뢰도의 값은 그 범위를 넘어서고 있다. 따라서 이 경우에도 비교 기여 순수 신뢰도가 본 논문에서 비교하는 신뢰도 중에서는 가장 바람직한 흥미도 측도라고 할 수 있다.

불일치빈도 c와 동시 비발생빈도 d의 값의 변화에 따른 신뢰도, 비교 신뢰도, 기여 순수 신뢰도, 그리고 비교 기여 순수 신뢰도의 값을 비교하기 위해 각 셀의 값을 바꾸어 실험해 보았는데, 이 경우에도 위의 결과와 동일하게 나타난 사실을 확인할 수 있었다.

4. 결론

데이터 마이닝 기법 중에서 많이 활용되고 있는 연관성 규칙은 여러 가지 흥미도 측도를 평가 기준으로 활용하여 의미 있는 규칙을 찾아낸다. 본 논문에서는 기존의 신뢰도 기반 흥미도 측도들이 가지고 있는 단점을 보완한 비교 기여 순수 신뢰도를 연관성 규칙의 새로운 평가 기준으로 제안한 후, 흥미도 측도의 조건의 충족여부를 조사하는 동시에 여러 가지 특성들을 살펴보았다. 그리고 예제 데이터를 이용하여 비교 기여 순수 신뢰도를 기존의 흥미도 측도인 신뢰도와 순수 신뢰도, 비교 신뢰도, 그리고 기여 순수 신뢰도와 비교하였다. 그 결과, 신뢰도는 모두 양의 값을 가지므로 연관성의 방향을 알 수 없어서 흥미도 측도로서는 바람직하다고 볼 수 없다. 비교 신뢰도와 기여 순수 신뢰도는 그 부호에 의해 연관성 규칙의 방향을 알 수 있으나, 기여 순수 신뢰도는 값의 범위가 $(-\infty,\ 1]$로 나타나고 있고, 비교 신뢰도는 값의 범위가 $[-1,\ \infty)$으로 나타나고 있어서 이 측도들 역시 흥미도 측도로서는 바람직하지 않다. 반면에 본 논문에서 제안한 비교 기여 순수 신뢰도는 값의 범위가 [-1, +1]로 나타나고 있어서 매우 바람직한 흥미도 측도라는 사실을 확인할 수 있었다.

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.

Ahn, K. and Kim, S. (2003). A new interstingness measure in association rules mining. *Journal of the Korean Institute of Industrial Engineers*, **29**, 41-48.

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

Berzal, F., Cubero, J. C., Marin, N. and Sanchez, D. (2004). Building multi-way decision trees with numerical attributes. *Information Sciences*, **165**, 73-90.

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

Freitas, A. (1999). On rule interestingness measures. *Knowledge-based System*, **12**, 309-315.

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., Pei, J. and Yin, Y. (2000). Mining frequent patterns without candidate generation. *Proceedings of ACM SIGMOD Conference on Management of Data*, 1-12.

Hilderman, R. J. and Hamilton, H. J. (2000). Applying objective interestingness measures in data mining systems. *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, 432-439.

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.

Kuo, Y. T. (2009) *Mining surprising patterns*, The doctoral paper of Melbourne university, Australia.

Liu, B., Hsu, W., Chen, S. and Ma, Y. (2000). Analyzing the subjective interestingness of association rules. *IEEE Intelligent Systems*, **15**, 47-55.

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

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

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. (2012). Exploration of symmetric similarity measures by conditional probabilities as association rule thresholds. *Journal of the Korean Data Analysis Society*, **14**, 707-716.

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

Silberschatz, A. and Tuzhilin, A. (1996) What makes patterns interesting in knowledge discovery systems. *IEEE transactions on Knowledge Data Engineering*, **8**, 970-974.

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.

Tan, P. N., Kumar, V. and Srivastava, J. (2002). Selecting the right interestingness measure for association patterns. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 32-41.

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
2013, **24**(3), 523–532

http://dx.doi.org/10.7465/jkdi.2013.24.3.523
한국데이터정보과학회지

The proposition of compared and attributably pure confidence in association rule mining

Hee Chang Park[1]

[1]Department of Statistics, Changwon National University

Received 15 April 2013, revised 3 May 2013, accepted 11 May 2013

Abstract

Generally, data mining is the process of analyzing big data from different perspectives and summarizing it into useful information. The most widely used data mining technique is to generate association rules, and it finds the relevance between two items in a huge database. This technique has been used to find the relationship between each set of items based on the interestingness measures such as support, confidence, lift, etc. Among many interestingness measures, confidence is the most frequently used, but it has the drawback that it can not determine the direction of the association. The attributably pure confidence and compared confidence are able to determine the direction of the association, but their ranges are not [-1, +1]. So we can not interpret the degree of association operationally by their values. This paper propose a compared and attributably pure confidence to compensate for this drawback, and then describe some properties for a proposed measure. The comparative studies with confidence, compared confidence, attributably pure confidence, and a proposed measure are shown by numerical example. The results show that the a compared and attributably pure confidence is better than any other confidences.

Keywords: Association threshold, attributably pure confidence, compared and attributably pure confidence, compared confidence, confidence.

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