Journal of the Korean Data & Information Science Society
2011, **22**(2), 235–243

한국데이터정보과학회지

연관 규칙 마이닝에서 기여 순수 신뢰도의 제안

박희창[1]

[1]창원대학교 통계학과

접수 2011년 2월 2일, 수정 2011년 3월 12일, 게재확정 2011년 3월 17일

요 약

데이터 마이닝 기법 중에서 가장 많이 이용되고 있는 기법은 연관성 규칙을 탐색하는 것으로, 이 기법은 지지도, 신뢰도, 향상도 등의 연관성 평가 기준을 기반으로 하여 각 항목집합들 간의 관련성을 찾아내는 데 활용되고 있다. 연관성을 평가하기 위한 기준으로 많은 흥미도 측도가 개발되어 있다. 그 중에서도 신뢰도가 가장 많이 활용되고 있으나 신뢰도는 연관성의 방향을 알 수가 없다는 단점을 가지고 있다. 이를 보완하기 위한 측도로 순수 신뢰도가 개발되었으나. 이 또한 양의 신뢰도의 값과 음의 신뢰도의 값이 동일한 경우에는 순수 신뢰도의 값이 같아지므로 이러한 경우에는 순수 신뢰도로는 차이를 알 수 없다.이에 본 논문에서는 기존의 신뢰도와 순수 신뢰도의 단점을 보완한 연관성 평가 기준인 기여 순수 신뢰도를 제안하였다. 또한 예제를 통하여 그 유용성을 알아본 결과, 기여 순수 신뢰도는 그 부호에 의해 연관성 규칙의 방향을 파악할 수 있는 동시에 순수 신뢰도에 의해서는 구분할 수 없는 상황도 해석 가능하게 할 수 있다는 사실을 확인하였다.

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

1. 서론

오늘날 기업이나 조직에서는 대용량 데이터로부터 알려지지 않은 흥미 있고 가치 있는 정보를 얻기를 원한다. 이러한 연유로 방대한 양의 데이터 속에서 쉽게 드러나지 않는 유용한 정보를 찾아내는 과정인 데이터마이닝 (data mining) 기법이 등장하게 되었다. 데이터마이닝 기법 중에서 가장 많이 활용되고 있는 연관성 규칙 (association rule)은 방대한 데이터로부터 항목집합들 간에 특정한 연관성을 발견하는 것으로 교차판매, 매장 진열, 카탈로그 디자인, 장바구니 분석 등 다양한 분야에서 많이 활용되고 있다. 연관성 규칙은 Agrawal 등 (1993)에 의해 처음 소개된 이후, 많은 학자들에 의해 연구되고 있다.(Agrawal과 Srikant, 1994; Park 등, 1995; Cai 등, 1998; Han과 Fu, 1999; Liu 등, 1999;, Pasquier 등, 1999; Han 등, 2000; Pei 등, 2000; Park과 Cho, 2005; Cho와 Park, 2008; Choi와 Park, 2008; Park, 2008, 2009 등).

이러한 연관성 규칙은 항목집합들 간의 지지도 (support), 신뢰도 (confidence), 향상도 (lift) 등의 흥미도 측도 (interestingness measure)를 바탕으로 관련성 여부를 측정한다. 의미 있는 연관성 규칙을 탐색하기 위한 흥미도 측도는 크게 객관적 흥미도 측도와 주관적 흥미도 측도로 나눌 수 있다 (Silberschatz와 Tuzhilin, 1996; Freitas, 1999). 객관적 흥미도 측도는 논리적인 또는 통계적인 방법에 의해 제안된 것으로 사용자에게 규칙을 정제할 수 있는 근거를 제시해주며, 주관적 흥미도 측도는 사용자 관점에서 해석 가능하도록 제안된 것이다. 흥미도 측도에 관해서는 많은 학자들에 의해 연구가 수행되었으며, 대표적인 연구로는 Hilderman 등 (2000)이 객관적 흥미도 측도들을 데이터마이닝에 응용하였으

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

며, Bing 등 (2000)은 주관적 흥미도 측도를 연관성 규칙에 적용한 바 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 신뢰도와 순수 신뢰도의 단점을 보완한 기여 순수 신뢰도 (attributably pure confidence)를 제안하고자 한다. 기여 순수 신뢰도는 신뢰도와 순수 신뢰도의 크기를 동시에 고려한 것으로 양의 신뢰도와 음의 신뢰도의 크기를 상대적으로 비교해서 나타낸 흥미도 측도라고 할 수 있다. 본 논문의 2절에서는 제안하는 흥미도 측도인 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 신뢰도와의 비교를 통해 기여 순수 신뢰도의 유용성에 대해 알아본 후, 4절에서 결론을 내리고자 한다.

2. 기여 순수 신뢰도

연관성 규칙을 평가하는 기준에는 지지도, 신뢰도, 향상도 등이 있으며, 이를 식으로 기술하기 위해 다음과 같은 분할표를 고려하기로 한다.

표 2.1 2×2 분할표

		Y		합계
		1	0	
X	1	n_{11}	n_{10}	$n_{1.}$
	0	n_{01}	n_{00}	$n_{0.}$
합계		$n_{.1}$	$n_{.0}$	n

지지도 $S(A \Rightarrow B)$는 항목 집합 A와 항목 집합 B가 동시에 발생하는 거래의 비율을 의미하며, 다음과 같이 정의된다.

$$S(A \Rightarrow B) = P(A \cap B) = \frac{n_{11}}{n} \tag{2.1}$$

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.

$$C(A \Rightarrow B) = P(B|A) = \frac{n_{11}}{n_{1.}} \tag{2.2}$$

신뢰도는 항목 집합 A를 포함하는 거래 중에서 항목 집합 B가 포함될 확률이 어느 정도인지를 확인하는 기준이 될 수 있으므로 관련성 규칙의 예측 지표라고 볼 수 있다.

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.

$$L(A \Rightarrow B) = \frac{P(B|A)}{P(B)} = \frac{P(A \cap B)}{P(A) \cdot P(B)} = \frac{n_{11} \cdot n}{n_{1.} \cdot n_{.1}} \tag{2.3}$$

향상도는 항목 집합 A가 주어지지 않았을 때의 항목 집합 B의 확률과 항목 집합 A가 주어졌을 때의 항목 집합 B의 확률의 비율을 의미한다.

한편, 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. 이러한 문제를 해결하기 위해 안광일과 김성집 (2003)은

의학분야에서 널리 이용되고 있는 기여위험률 (attributable risk)을 순수 신뢰도 (net confidence ; $Nconf$)라는 이름으로 데이터 마이닝 분야에 적용한 바 있다.

$$Nconf(A \Rightarrow B) = P(Y|X) - P(Y|\bar{X}) = \frac{n_{11} \cdot n_{00} - n_{10} \cdot n_{01}}{n_{1.} \cdot n_{0.}} \tag{2.4}$$

여기서 $\bar{X}$의 의미는 X 가 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 $P(Y|\bar{X})$의 값이 어떤 값을 가지더라도 두 값의 차이가 동일하면 순수 신뢰도의 값도 동일하게 되는 단점을 가지고 있다. 이를 좀 더 구체적으로 알아보기 위해 다음과 같은 가상의 예제를 이용하여 설명하면 다음과 같다. 먼저 표 2.2와 표 2.3으로부터 신뢰도인 $P(Y|X)$을 계산하면 각각 0.5와 0.8이며, $P(Y|\bar{X})$의 값은 각각 0.3과 0.6으로 나타났다. 순수 신뢰도는 두 표 모두 0.2로 동일하므로 순수 신뢰도만을 가지고는 이 두 경우의 차이를 설명할 수 없다. 또한 표 2.2는 $P(Y|X)$이 $P(Y|\bar{X})$에 비해 3배의 크기이고, 표 2.3은 $P(Y|X)$이 $P(Y|\bar{X})$에 비해 약 1.3배의 크기이다. 그러나 순수 신뢰도를 연관성 측도를 사용하게 되면 이러한 크기를 고려할 수 없다.

표 2.2 가상의 분할표(1)

		Y		합계
		1	0	
X	1	50	50	100
	0	30	70	100
합계		80	120	200

표 2.3 가상의 분할표(2)

		Y		합계
		1	0	
X	1	80	20	100
	0	60	40	100
합계		140	60	200

이러한 문제를 보완하기 위해 본 논문에서는 다음과 같은 기여 순수 신뢰도 ($APconf$)를 제안하고자 한다.

$$APconf(X \Rightarrow Y) = \frac{P(Y|X) - P(Y|\bar{X})}{P(Y|X)} = \frac{n_{11} \cdot n_{00} - n_{10} \cdot n_{01}}{n_{11} \cdot n_{0.}} \tag{2.5}$$

예를 들어 기여 순수 신뢰도의 값이 2/3라는 의미는 항목 X를 구매한 자로서 항목 Y를 구매한 자 중에서 2/3만이 항목 X 구매자에 의해 항목 Y가 구매된다는 의미이고, 나머지 1/3은 항목 X를 구매하지 않았어도 항목 Y를 구매할 사람들의 비율을 의미한다. 이 측도는 의학분야에서 노출군과 비노출군을 합한 전체 집단에서 발생한 환자 중에서 요인에 의해서 발생한 환자가 차지하는 분율을 나타내는 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 $P(Y|\bar{X})$의 차이 크기를 $P(Y|X)$에 대해 상대적으로 나타낸 것으로 이를 이용하면 $P(Y|X)$와 $P(Y|\bar{X})$의 크기를 반영할 수 있게 된다.

먼저 기여 순수 신뢰도 $APconf$가 가지고 있는 성질을 기술하면 다음과 같다.

<성질 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.

(설명) : $APconf > 0$이면 순수 신뢰도와 마찬가지로 식 (2.5)의 분자의 값이 양의 값이 되어야 하므로, 즉 $P(Y|X) > P(Y|\bar{X})$이므로 항목집합 X가 포함된 트랜잭션에서 항목집합 Y가 발견되는 확률이 항목집합 X가 포함되지 않은 트랜잭션에서 항목집합 Y가 발견되는 확률보다 크므로 이들 두 항목 간에는 양의 연관성이 존재한다고 볼 수 있다. 만약 $APconf < 0$이면 이 또한 순수 신뢰도와 마찬가지로 항목집합 X가 포함된 트랜잭션에서 항목집합 Y가 발견되는 확률이 항목집합 X가 포함되지 않은 트랜잭션에서 항목집합 Y가 발견되는 확률보다 작으므로 이들 두 항목 간에는 음의 연관성이 존재한다고 볼 수 있다. 그리고 $APconf = 0$이면 항목집합 Y는 항목집합 X가 포함된 트랜잭션에서뿐만 아니라 X가 발견되지 않는 트랜잭션에서도 동일한 확률로 발견되므로 X와 Y가 서로 독립관계라고 볼 수 있다.

<성질 2> 일반적으로 $APconf(X \Rightarrow Y)$와 $APconf(Y \Rightarrow X)$의 값은 동일하지 않다.

(설명) : 연관성 규칙의 흥미도 측도는 방향성을 갖는 것이 바람직한데 (Berzal 등, 2001), 식 (2.5)를 통해서 알 수 있는 바와 같이 $APconf$는 $Y \Rightarrow X$와 $X \Rightarrow Y$의 값이 다르므로 방향성이 고려된다.

<성질 3> $APconf(X \Rightarrow Y) = APconf(Y \Rightarrow X)$이면 항목집합 X와 Y는 서로 독립이다.

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

$$\frac{P(Y|X) - P(Y|\bar{X})}{P(Y|X)} = \frac{P(X|Y) - P(X|\bar{Y})}{P(X|Y)} \tag{2.6}$$

식 (2.6)을 정리하면 $P(X \cap Y) = P(X)P(Y)$가 되어 두 항목집합 X와 Y는 서로 독립관계에 놓이게 된다. 따라서 $APconf(X \Rightarrow Y) = APconf(Y \Rightarrow X)$이면 항목집합 X와 Y는 서로 독립이 된다.

<성질 4> $APconf(X \Rightarrow Y)$ 값의 범위는 $[-\infty, 1]$이다.

(설명) : $APconf(X \Rightarrow Y)$ 값이 1이라는 의미는 $APconf(X \Rightarrow Y)$의 분자와 분모 값이 1이므로 항목집합 X가 발견되는 모든 트랜잭션에서 항목집합 Y가 발견되고 X가 없는 트랜잭션에서는 Y가 전혀 발생하지 않는다는 의미이다. 그리고 $APconf(X \Rightarrow Y)$의 값이 $-\infty$라는 의미는 $APconf(X \Rightarrow Y)$의 분자 값이 -1이므로 항목집합 X가 발견되지 않는 트랜잭션에서만 거의 모든 항목집합 Y가 발견된다는 것이다.

다음으로는 본 논문에서 제안한 기여 순수 신뢰도 $APconf$이 Piatetsky-Shapiro (1991)가 제안한 흥미도 측도의 조건을 충족하는지의 여부를 증명하면 다음과 같다.

[조건 1] $P(X \cap Y) = P(X)P(Y)$이면 $APconf$의 값은 0이 된다.

(증명) : 먼저 식 (2.5)의 $APconf(X \Rightarrow Y)$를 정리하면 다음과 같이 표현된다.

$$APconf(X \Rightarrow Y) = 1 - \frac{[P(Y) - P(X \cap Y)]P(X)}{P(X \cap Y)[1 - P(X)]} \tag{2.7}$$

$P(X \cap Y) = P(X)P(Y)$이면 $APconf$의 분자와 분모의 값이 같으므로 $APconf$의 값은 0이 된다.

[조건 2] $APconf$는 $P(Y)$의 값에 따라 단조 감소한다.

(증명) : 위의 식 (2.7)로부터 $P(Y)$의 값이 증가하면 $APconf$는 감소한다는 사실을 알 수 있다.

[조건 3] $APconf$ 는 $P(X \cap Y)$의 값에 따라 단조 증가한다.

(증명) : 식 (2.7)의 $APconf$를 다시 한 번 정리하면 다음과 같이 나타낼 수 있다.

$$APconf(X \Rightarrow Y) = 1 - \frac{P(X)P(Y)}{P(X \cap Y)[1 - P(X)]} + \frac{P(X)}{[1 - P(X)]} \tag{2.8}$$

이로부터 $P(X \cap Y)$의 값이 증가함에 따라 $APconf$는 단조 증가하는 것을 알 수 있다.

3. 예제 데이터의 적용

본 절에서는 신뢰도와 순수 신뢰도의 문제점과 기여 순수 신뢰도의 유용성을 예제를 통해 고찰하고자 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 냉장고의 금액을 기준으로 100만원 이상(1) 구매한 사람 수를 50명으로 하고 100만원 미만(0)을 구매한 사람 수를 50명으로 하였다. 또한 항목 집합 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 집합 X와 Y가 동시에 발생한 빈도 수, 즉 100만원 이상의 냉장고를 구매하면서 신용카드로 결제한 빈도수는 a명으로 하였다. 이를 정리하면 표 3.1과 같다. 이 표에서 a가 취할 수 있는 범위는 $0 \le a \le 30$이다.

표 3.1 예제 데이터(1)

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

표 3.1로부터 동시발생빈도(a)에 따른 지지도, 신뢰도, 음의 신뢰도, 순수 신뢰도, 그리고 기여 순수 신뢰도를 계산하면 다음의 표 3.2와 같은 결과를 얻을 수 있다. 여기서 $b = P(X = 1, Y = 0)$, $c = P(X = 0, Y = 1)$, $d = P(X = 0, Y = 0)$을 의미한다. 이 표로부터 알 수 있는 바와 같이 a의 값이 커질수록 신뢰도와 순수 신뢰도, 그리고 기여 순수 신뢰도는 증가하고 있는 반면에 참고하기 위해 계산된 음의 신뢰도는 감소하고 있다. 또한 신뢰도는 모두 양의 값을 가지므로 방향이 없어서 그 값만으로는 양의 연관성이 있는지 아니면 음의 연관성이 있는지를 알 수 없으나, 순수 신뢰도와 기여 순수 신뢰도는 그 부호에 의해 연관성 규칙의 방향을 알 수 있다. 이에 대해 좀 더 구체적으로 알아보기 위해 a=12, b=38, c=18, d=32인 경우와 a=18, b=32, c=12, d=38인 경우를 비교해보면, 신뢰도와 음의 신뢰도, 순수 신뢰도, 그리고 기여 순수 신뢰도는 각각 0.240, 0.360, -0.120, -0.500과 0.360, 0.240, 0.120, 0.333으로 나타나서 a가 증가하면 신뢰도와 순수 신뢰도, 그리고 기여 순수 신뢰도는 증가하며, 음의 신뢰도는 감소하고 있다. 또한 신뢰도는 0.240과 0.360으로 두 경우 모두 양의 값으로 나타나나 순수 신뢰도와 기여 순수 신뢰도는 양의 신뢰도와 음의 신뢰도를 함께 고려함으로써 각각 -0.120 및 -0.500과 0.120 및 0.333으로 나타나게 되어 연관성의 방향을 가늠할 수 있는 측도가 된다. 반면에 순수 신뢰도는 단지 양의 신뢰도와 음의 신뢰도의 차이만을 고려하고 있으며, 기여 순수 신뢰도는 순수 신뢰도와 양의 신뢰도의 비율을 나타내고 있다. 위의 예에서 기여 순수 신뢰도가 0.333이라는 의미는 냉장고를 100만원 이상 구매한 자 중에서 신용카드로 결제한 자 중에서 33.3%만 냉장고를 100만원 이상 구매한 자에 의해 신용카드로 결제한다는 의미이고, 나머지 66.7%는 냉장고를 100만원 이상 구매하지 않았어도 신용카드로 결제한다는 의미이다.

이번에는 b의 값의 변화에 따른 순수 신뢰도와 조건부 순수 신뢰도의 값을 비교하기 위해 다음과 같이 각 셀의 값을 바꾸어 실험하였다.
표 3.3에서 b가 취할 수 있는 정수 값의 범위는 $0 \le b \le 30$이다. 이 표로부터 각 셀 값의 변화에 따른 신뢰도, 순수 신뢰도, 그리고 조건부 순수 신뢰도를 계산하면 다음의 표 3.4와 같은 결과를 얻을 수 있다. 이 표로부터 알 수 있는 바와 같이 b의 값이 커질수록 신뢰도와 순수 신뢰도, 그리고 기여 순수 신뢰도

표 3.2 예제 데이터(1)에 의해 계산된 각종 흥미도 측도의 결과

a	b	c	d	지지도	신뢰도	음의 신뢰도	순수신뢰도	기여 순수 신뢰도
6	44	24	26	0.060	0.120	0.480	-0.360	-3.000
7	43	23	27	0.070	0.140	0.460	-0.320	-2.286
8	42	22	28	0.080	0.160	0.440	-0.280	-1.750
9	41	21	29	0.090	0.180	0.420	-0.240	-1.333
10	40	20	30	0.100	0.200	0.400	-0.200	-1.000
11	39	19	31	0.110	0.220	0.380	-0.160	-0.727
12	38	18	32	0.120	0.240	0.360	-0.120	-0.500
13	37	17	33	0.130	0.260	0.340	-0.080	-0.308
14	36	16	34	0.140	0.280	0.320	-0.040	-0.143
15	35	15	35	0.150	0.300	0.300	0.000	0.000
16	34	14	36	0.160	0.320	0.280	0.040	0.125
17	33	13	37	0.170	0.340	0.260	0.080	0.235
18	32	12	38	0.180	0.360	0.240	0.120	0.333
19	31	11	39	0.190	0.380	0.220	0.160	0.421
20	30	10	40	0.200	0.400	0.200	0.200	0.500
21	29	9	41	0.210	0.420	0.180	0.240	0.571
22	28	8	42	0.220	0.440	0.160	0.280	0.636
23	27	7	43	0.230	0.460	0.140	0.320	0.696
24	26	6	44	0.240	0.480	0.120	0.360	0.750
25	25	5	45	0.250	0.500	0.100	0.400	0.800

표 3.3 예제 데이터(2)

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

는 감소하고 있는 반면에 음의 신뢰도는 증가하고 있다. 이 표에서도 신뢰도는 모두 양의 값을 가지므로 방향이 없어서 그 값만으로는 양의 연관성이 있는지 아니면 음의 연관성이 있는지를 알 수 없으나, 순수 신뢰도와 기여 순수 신뢰도는 그 부호에 의해 연관성 규칙의 방향을 알 수 있다. 이에 대해 좀 더 구체적으로 알아보기 위해 a=38, b=12, c=32, d=18인 경우와 a=32, b=18, c=38, d=12인 경우를 비교해보면, 신뢰도와 음의 신뢰도, 순수 신뢰도, 그리고 기여 순수 신뢰도는 각각 0.760, 0.640, 0.120, 0.158과 0.640, 0.760, -0.120, -0.188로 나타나서 b가 증가하면 신뢰도와 순수 신뢰도, 그리고 기여 순수 신뢰도는 감소하며, 음의 신뢰도는 증가하고 있다. 또한 신뢰도는 0.760과 0.640으로 두 경우 모두 양의 값으로 나타나나 순수 신뢰도와 기여 순수 신뢰도는 양의 신뢰도와 음의 신뢰도를 함께 고려함으로써 각각 0.120 및 0.158과 -0.120 및 -0.188로 나타나게 되어 연관성의 방향을 가늠할 수 있는 측도가 된다.

표 3.2와 표 3.4를 동시에 고려해보면 순수 신뢰도와 기여 순수 신뢰도에 대한 차이를 확연하게 알 수 있다. 위에서 고려한 예에서 순수 신뢰도가 0.120인 경우를 고려해보면 각각 a=18, b=32, c=12, d=38인 경우와 a=38, b=12, c=32, d=18인 경우이다. 이 두 경우에 기여 순수 신뢰도는 각각 0.333과 0.158이다. 따라서 양의 신뢰도와 음의 신뢰도의 값의 차이가 동일한 경우에는 순수 신뢰도의 값만 가지고는 그 의미를 해석할 수 없는 반면에 기여 순수 신뢰도는 양의 신뢰도와 음의 신뢰도의 값의 차이가 같지만 두 값이 다른 경우에도 그 의미를 해석할 수 있으므로 순수 신뢰도에 비해 기여 순수 신뢰도가 더 바람직한 측도라고 할 수 있다.

c 와 d 의 값의 변화에 따른 신뢰도 및 순수 신뢰도와 기여 순수 신뢰도의 값을 비교하기 위해 각 셀의

표 3.4 예제 데이터(2)에 의해 계산된 각종 흥미도 측도의 결과

a	b	c	d	지지도	신뢰도	음의 신뢰도	순수신뢰도	기여 순수 신뢰도
45	5	25	25	0.450	0.900	0.500	0.400	0.444
44	6	26	24	0.440	0.880	0.520	0.360	0.409
43	7	27	23	0.430	0.860	0.540	0.320	0.372
42	8	28	22	0.420	0.840	0.560	0.280	0.333
41	9	29	21	0.410	0.820	0.580	0.240	0.293
40	10	30	20	0.400	0.800	0.600	0.200	0.250
39	11	31	19	0.390	0.780	0.620	0.160	0.205
38	12	32	18	0.380	0.760	0.640	0.120	0.158
37	13	33	17	0.370	0.740	0.660	0.080	0.108
36	14	34	16	0.360	0.720	0.680	0.040	0.056
35	15	35	15	0.350	0.700	0.700	0.000	0.000
34	16	36	14	0.340	0.680	0.720	-0.040	-0.059
33	17	37	13	0.330	0.660	0.740	-0.080	-0.121
32	18	38	12	0.320	0.640	0.760	-0.120	-0.188
31	19	39	11	0.310	0.620	0.780	-0.160	-0.258
30	20	40	10	0.300	0.600	0.800	-0.200	-0.333
29	21	41	9	0.290	0.580	0.820	-0.240	-0.414
28	22	42	8	0.280	0.560	0.840	-0.280	-0.500
27	23	43	7	0.270	0.540	0.860	-0.320	-0.593
26	24	44	6	0.260	0.520	0.880	-0.360	-0.692

값을 바꾸어 실험해 보았는데, 이 경우에도 위의 결과와 동일하게 나타난 사실을 확인할 수 있었다.

4. 결론

데이터 마이닝 기법 중에서 가장 많이 활용되고 있는 연관성 규칙은 여러 가지 흥미도 측도를 평가 기준으로 활용하여 의미 있는 규칙을 찾아낸다. 본 논문에서는 기존의 신뢰도와 순수 신뢰도가 가지고 있는 약점을 보완한 기여 순수 신뢰도를 연관성 규칙의 새로운 평가 기준으로 제안한 후, 이에 대한 여러 가지 특성에 대해 살펴보았다. 기여 순수 신뢰도의 값이 양이면 두 항목집합은 양의 연관성을 가지고, 음이면 음의 연관성을 가지며, 값이 0이면 두 항목집합은 서로 독립관계임을 알 수 있었으며, 그 값의 범위는 $[-\infty, 1]$이다. 또한 두 항목 집합에 대한 연관성의 방향이 다른 경우에는 기여 순수 신뢰도의 값이 달라지며, 만약 방향이 다른 경우에도 그 값이 동일하면 두 항목 집합은 독립이 된다는 특성을 가지고 있는 것으로 파악되었다. 또한 Piatetsky-Shapiro가 제안한 흥미도 측도의 조건의 충족여부에 대해서도 알아보았다. 그리고 예제 데이터를 이용하여 기여 순수 신뢰도를 기존의 흥미도 측도인 신뢰도와 순수 신뢰도와 비교하였다. 그 결과, 신뢰도는 모두 양의 값을 가지므로 방향이 없으며, 순수 신뢰도는 그 부호에 의해 연관성 규칙의 방향을 알 수는 있으나 양의 신뢰도와 음의 신뢰도의 값의 차이가 동일한 경우에는 그 의미를 해석할 수 없다는 사실을 확인할 수 있었다. 반면에 기여 순수 신뢰도는 양의 신뢰도와 음의 신뢰도의 값의 차이가 같은 경우에도 그 의미를 해석할 수 있으므로 순수 신뢰도에 비해 기여 순수 신뢰도가 더 바람직한 측도라는 사실을 확인할 수 있었다.

향후 연구과제로는 본 논문에서 제안한 결과를 바탕으로 기여 순수 지지도와 기여 순수 향상도를 고안하여 기여 순수 연관성 규칙이 제안되어야 할 것이다.

참고문헌

안광일, 김성집 (2003). 연관규칙 탐색에서의 새로운 흥미도 척도의 제안. <대한산업공학회지>, **29**, 41-48.

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.

Berzal, F., Blanco, I., Sanchez, D. and Vila, M. (2001). A new framework to assess association rules. *Proceedings of the 4th International Conference on Intelligent Data Analysis*, 95-104.

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

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.

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

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.

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.

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. (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. (2009). Proposition of pure association rule for original characteristics grasping. *Journal of the Korean Data Analysis Society*, **11**, 859-869.

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-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.

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.

Journal of the Korean Data & Information Science Society
2011, **22**(2), 235–243

한국데이터정보과학회지

The proposition of attributably pure confidence in association rule mining

Hee Chang Park[1]

[1]Department of Statistics, Changwon National University

Received 2 February 2011, revised 12 March 2011, accepted 17 March 2011

Abstract

The most widely used data mining technique is to explore association rules. This technique has been used to find the relationship between each set of items based on the association thresholds such as support, confidence, lift, etc.There are many interestingness measures as the criteria for evaluating association rules. Among them, confidence is the most frequently used, but it has the drawback that it can not determine the direction of the association. The net confidence measure was developed to compensate for this drawback, but it is useless in the case that the value of positive confidence is the same as that of negative confidence.This paper propose a attributably pure confidence to evaluate association rules and then describe some properties for a proposed measure. The comparative studies with confidence, net confidence, and attributably pure confidence are shown by numerical example. The results show that the attributably pure confidence is better than confidence or net confidence.

Keywords: Association threshold, attributably pure confidence, confidence, net confidence.

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