*Journal of the Korean Data & Information Science Society*
*2012,* **23**(5), 939–948

http://dx.doi.org/10.7465/jkdi.2012.23.5.939
한국데이터정보과학회지

# 음의 연관성 규칙 생성을 위한 음의 기여 순수 신뢰도의 제안

박희창[1]

[1]창원대학교 통계학과

접수 2012년 8월 8일, 수정 2012년 8월 28일, 게재확정 2012년 9월 3일

**요약**

데이터 마이닝 기법들 중에서 가장 많이 활용되고 있는 연관성 규칙은 방대한 데이터베이스에서 항목간의 관계를 흥미도 측도에 의해 명확히 수치화함으로써 그들간의 관련성을 표시해주는 기법이다. 양의 연관성 규칙 마이닝이 임의의 한 항목이 발생하면 다른 항목도 발생한다는 규칙을 생성하기 위한 기법인 반면에, 음의 연관성 규칙은 어느 항목이 발생하면 다른 항목은 발생하지 않는다는 규칙을 찾아내는 기법이다. 음의 연관성 규칙은 양의 연관성 규칙의 활용과 마찬가지로 고객의 구매 경향 및 마케팅 정책을 제시할 수 있고 교차판매와 매장 진열 등과 같이 타겟 마케팅에 활용 가능하다. 양의 연관성 규칙에 음의 연관성 규칙을 추가하게 되면 어떤 제품을 판매하기 위해서는 그 제품만 마케팅 하는 것뿐만 아니라 더 나아가 그 제품이 아닌 어느 제품을 마케팅 하는 것이 필요한지를 판단할 수 있다. 본 논문에서는 기존의 음의 신뢰도의 단점을 보완할 수 있는 음의 기여 순수 신뢰도를 제안한 후, 이에 대해 흥미도 측도가 가져야 할 조건들을 조사하였으며, 예제 데이터를 활용하여 음의 기여 순수 신뢰도의 유용성을 고찰하였다.

주요용어: 음의 기여 순수 신뢰도, 음의 순수 신뢰도, 음의 신뢰도, 흥미도 측도.

## 1. 서론

현대의 기업이나 조직은 최적의 전략수립과 올바른 의사결정을 위한 의미 있는 고급정보를 확보하기 위해 다양한 경영기법들을 개발하여 적용하고 있는데 그 중의 하나가 데이터 마이닝 기법이다. 이는 방대한 양의 데이터 속에서 쉽게 드러나지 않는 유용한 정보를 찾아내는 기법이다. 데이터 마이닝 기법들 중에서도 연관성 규칙은 가장 많이 활용되는 기법으로 방대한 데이터베이스 내에서 항목들 간의 관계를 지지도 (support), 신뢰도 (confidence), 향상도 (lift) 등의 흥미도 측도 (interestingness measure)를 바탕으로 관련성 여부를 측정한다. 연관성 규칙 기법은 Agrawal 등 (1993)이 처음으로 연구를 수행하였으며, 지금까지 국내외적으로 많은 학자들이 연관성 규칙과 관련된 연구를 수행하고 있다. Cho와 Park (2011)에 따르면 수행된 많은 연구들 중에서 연관성규칙 생성에 대한 수행속도를 향상시키기 위한 대표 연구로는 Han 등 (2000), Pei 등 (2000), Saygin 등 (2002) 등이 있으며, 제약조건을 가지는 항목으로 구성된 트랜잭션 데이터베이스에서 빈발항목을 찾는 대표 연구로는 Han과 Fu (1999), Liu 등 (1999)이 있다. 또한 연관성 규칙에 대한 최근 국내 연구로는 Choi와 Park (2008), Lim 등 (2010), Park (2010a, 2010b, 2011a, 2011b, 2011c) 등이 있다.

한편, Han과 Kamber (2006)은 연관성 규칙을 여러 형태로 분류하였는데 응용 결과에 따라 순차연관성 규칙 (sequences association rule)과 음의 연관성 규칙 (negative association rule)으로 분류한

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

바 있다. 일반적으로 연관성 규칙 마이닝에서는 어느 항목이 발생하면 다른 항목도 발생한다는 규칙을 발견하는 기법인 반면에, 음의 연관성 규칙 마이닝은 어느 항목이 발생하면 다른 항목은 발생하지 않는다는 규칙을 찾아내는 것이다. 음의 연관 규칙은 양의 연관 규칙과 같이 흔하게 나타나는 규칙은 아니지만 지지도와 신뢰도가 음의 연관 규칙에서 훨씬 높게 나타난다면 오히려 음의 연관규칙에서 찾아낸 규칙이 훨씬 더 가치 있다고 볼 수 있다 (Lee 등, 2003). 음의 연관성 규칙에 대한 최근 연구로는 Sim 등 (2008), Shang 등 (2008), 그리고 Bala (2009) 등이 있다. 연관성 규칙은 전항 항목을 고정시키고 후항 항목을 마케팅 하는 반면에 음의 연관성 규칙을 추가로 생성하게 되면 후항 항목을 고정시키고 전항 항목을 마케팅 하는 전략도 가능하게 된다 (Hwang과 Kim, 2003). 이와 같은 음의 연관성 규칙을 생성하기 위한 평가기준으로 음의 신뢰도 (negative confidence)가 있다. 그러나 기존의 음의 신뢰도는 양의 값만을 가지므로 연관성의 방향을 알 수 없는 동시에 양의 연관성을 가지는 연관성 규칙을 음의 연관성 규칙으로 선택하게 되는 오류를 범할 수 있다. Park (2011a)은 이러한 단점을 보완할 수 있는 음의 순수 신뢰도 (negatively pure confidence)를 제안한 바 있다. 음의 순수 신뢰도는 순수하게 특정 항목에 의해서만 결과가 얼마인가를 나타내주는 측도인 동시에 그 부호에 의해 음의 관련성과 양의 관련성을 판단할 수는 있으나 수식에 포함된 두 확률값의 크기가 동일하면 순수 신뢰도의 값도 동일하게 되어 관련성의 강도를 구분하지 못하는 단점을 가지고 있다.

본 논문에서는 이러한 단점을 보완할 수 있는 음의 순수 연관성 규칙 (negatively pure association rule)의 측도인 음의 기여 순수 신뢰도 (negatively attributable and pure confidence)를 제안하고자 한다. 논문의 2절에서는 음의 기여 순수 신뢰도를 정의하고, Piatetsky-Shapiro (1991)가 제안한 흥미도 측도가 가져야 할 조건들을 조사한다. 3절에서는 모의실험 데이터를 활용하여 기존의 음의 신뢰도와 음의 순수 신뢰도와의 비교를 통해 음의 기여 순수 신뢰도의 유용성에 대해 살펴본 후, 마지막으로 4절에서 결론을 내리고자 한다.

## 2. 음의 기여 순수 신뢰도의 제안

### 2.1. 음의 연관성 규칙

음의 연관성 규칙은 기존의 연관 규칙에서 'not'의 개념이 들어간 것이라고 볼 수 있다. 음의 연관성 규칙을 기호로 나타내면 $A \to B^c$, $A^c \to B$ 형태로 표현할 수 있으며, 상첨자 $c$는 'not'의 의미이다. 음의 연관성 규칙은 양의 연관성 규칙과 같이 흔하게 나타나는 규칙이 아니다. 그러나 만일 생성된 규칙 중 지지도와 신뢰도가 음의 연관성 규칙 쪽에서 훨씬 높게 나타난다면 오히려 음의 연관 규칙에서 찾아낸 규칙이 훨씬 더 가치가 있을 것이다. 음의 연관성 규칙을 평가하는 기준들을 수식으로 표현하기 위해 다음과 같은 분할표를 고려하기로 한다.

**Table 2.1** $2 \times 2$ contingency table

| | | $B$ | | Total |
|---|---|---|---|---|
| | | 1 | 0 | |
| $A$ | 1 | $n_{11}$ | $n_{10}$ | $n_{1.}$ |
| | 0 | $n_{01}$ | $n_{00}$ | $n_{0.}$ |
| Total | | $n_{.1}$ | $n_{.0}$ | $n$ |

'A이면 not B이다.' 또는 'not A이면 B이다.'로 정의되는 음의 연관성 규칙의 평가기준인 음의

지지도 $NS$, 음의 신뢰도 $NC$, 그리고 음의 향상도 $NL$은 다음과 같다 (Park, 2011a).

$$NS_1(A \Rightarrow B^c) = P(A \cap B^c) = n_{10}/n, NS_2(A^c \Rightarrow B) = P(A^c \cap B) = n_{01}/n$$
$$NC_1(A \Rightarrow B^c) = P(B^c|A) = n_{10}/n_{1.}, NC_2(A^c \Rightarrow B) = P(B|A^c) = n_{01}/n_{0.}$$
$$NL_1(A \Rightarrow B^c) = P(B^c|A)/P(B^c) = n_{10}n/(n_{1.}n_{.0}), NL_2(A^c \Rightarrow B) = P(B|A^c)/P(B) = n_{01}n/(n_{0.}n_{.1})$$

여기서 $A^c$와 $B^c$의 의미는 각각 $A$와 $B$가 일어나지 않음을 의미한다. 연관성 규칙 마이닝에서 음의 연관성 규칙을 추가하게 되면 어떤 제품을 판매하기 위해서는 그 제품만 마케팅 하는 것뿐만 아니라 더 나아가 그 제품이 아닌 어느 제품을 마케팅 하는 것이 필요한 지를 판단할 수 있다. 그러나 음의 연관성 규칙에서는 음의 연관성 측정을 위한 기존의 신뢰도는 항상 양의 값을 갖게 되어 연관성의 방향을 알 수 없다. Park (2011a)은 이러한 단점을 보완할 수 있는 음의 순수 신뢰도를 제안한 바 있다.

$$NC_{pure}(A \Rightarrow B) = \begin{cases} P(B^c|A) - P(B^c|A^c) \\ P(B|A^c) - P(B|A) \end{cases} \tag{2.1}$$

식 (2.1)의 첫 번째 식을 편의상 $NC_{pure1}(A \Rightarrow B^c)$으로 표기하고, 두 번째 식을 $NC_{pure2}(A^c \Rightarrow B)$으로 표기한다. 음의 연관성 규칙에서의 기존의 신뢰도는 항목 집합 A가 포함된 거래 비율 중 항목 집합 A는 포함되고 항목 집합 B는 포함되지 않은 거래의 비율 또는 항목 집합 A가 포함되지 않은 거래 비율 중 항목 집합 B는 포함되고 항목 집합 A는 포함되지 않은 거래의 비율이다. 반면에 음의 순수 신뢰도는 A가 포함된 거래 중 B가 포함되지 않은 거래의 비율과 A가 포함되지 않은 거래 비율 중 B도 포함되지 않은 거래의 비율의 차이 또는 A가 포함되지 않은 거래 중 B가 포함된 거래의 비율과 A가 포함된 거래 비율 중 B가 포함된 거래의 비율의 차이를 의미한다. 위에서 기술한 바와 같이 기존의 음의 신뢰도는 양 또는 역의 연관성 방향(양, 역, 음)을 알 수 없을 뿐만 아니라 음의 연관성을 가지는 연관성 규칙을 의미 있는 양 또는 역의 관계를 가지는 규칙으로 선택하게 되는 오류를 범할 수 있다. 하지만 음의 순수 신뢰도는 순수하게 특정 요인에 의해서만 결과가 얼마인가를 나타내주는 측도이며, 부호에 의해 양 또는 역의 관련성과 음의 관련성을 판단할 수 있다.

### 2.2. 음의 기여 순수 신뢰도

양의 연관성 규칙 생성을 위한 순수 신뢰도와 마찬가지로 음의 순수 신뢰도는 부호에 의해 양의 관련성과 음의 관련성을 판단할 수 있기는 하지만, $P(B^c|A)$와 $P(B^c|A^c)$의 값 (또는 $P(B|A^c)$와 $P(B|A)$의 값)이 어떤 값을 가지더라도 두 값의 차이가 동일하면 순수 신뢰도의 값도 동일하게 되는 단점을 가지고 있다. 이를 좀 더 구체적으로 알아보기 위해 Table 2.2와 Table 2.3과 같은 가상의 분할표를 이용하고자 한다.

**Table 2.2** Virtual contingency table(1)

| | | B | | total |
|---|---|---|---|---|
| | | 1 | 0 | |
| A | 1 | 50 | 50 | 100 |
| | 0 | 70 | 30 | 100 |
| total | | 120 | 80 | 200 |

**Table 2.3** Virtual contingency table(2)

| | | B | | total |
|---|---|---|---|---|
| | | 1 | 0 | |
| A | 1 | 60 | 40 | 100 |
| | 0 | 80 | 20 | 100 |
| total | | 140 | 60 | 200 |

먼저 Table 2.2와 Table 2.3으로부터 음의 신뢰도 $NC_1(A \Rightarrow B^c) = P(B^c|A)$를 계산하면 각각 0.5와 0.4이며, $P(B^c|A^c)$는 각각 0.3과 0.2로 나타났다. 이로부터 음의 순수 신뢰도는 두 표에서 모두 0.2로 동일하게 나타났으므로 음의 순수 신뢰도만을 가지고는 이 두 경우의 차이를 설명할 수 없다. 또한 Table 2.2는 $P(B^c|A)$이 $P(B^c|A^c)$에 비해 1.25배의 크기이고, Table 2.3은 1.50배의 크기이나 음의 순수 신뢰도로는 이러한 크기를 고려할 수 없다. 이러한 문제를 보완하기 위해 본 논문에서는 다음과 같은 음의 기여 순수 신뢰도 $NAC_{pure}$를 제안하고자 한다. 편의상 식 (2.2)의 첫 번째 식을 $NAC_{pure1}(A \Rightarrow B^c)$으로 표기하고, 두 번째 식을 $NAC_{pure2}(A^c \Rightarrow B)$으로 표기한다.

$$NAC_{pure}(A \Rightarrow B) = \begin{cases} \dfrac{P(B^c|A) - P(B^c|A^c)}{P(B^c|A)} \\ \dfrac{P(B|A^c) - P(B|A)}{P(B|A^c)} \end{cases} \tag{2.2}$$

양의 연관성 규칙 마이닝에서 기여 순수 신뢰도는 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 (Park, 2011b). 이를 본 논문에서는 음의 연관성 규칙을 탐색하기 위해 Park (2011b)에서 고려한 순수 신뢰도와 기여 순수 신뢰도와의 관계식을 음의 연관성 규칙에 부합하도록 변형하여 나타낸 것이다. 위의 Table 2.2와 Table 2.3에서 기여 순수 신뢰도를 계산하면 각각 0.4와 0.5로 나타났다. 따라서 음의 기여 순수 신뢰도는 $P(B^c|A)$와 $P(B^c|A^c)$와의 차이를 $P(B^c|A)$에 대해 상대적으로 나타낸 것으로, 이를 이용하면 $P(B^c|A)$와 $P(B^c|A^c)$의 크기를 반영할 수 있게 된다. 또한 Table 2.2와 Table 2.3으로부터 향상도(lift)를 계산하면 0.833과 0.857로 두 값 모두 1 보다 작게 나타나고 있어서 향상도를 이용했을 때와 같이 동일한 방향의 결과를 얻을 수 있다.

다음으로는 본 논문에서 고려하는 음의 기여 순수 신뢰도 $NAC_{pure1}$과 $NAC_{pure2}$가 Piatetsky-Shapiro가 제안한 흥미도 측도의 세 가지 조건을 충족하는지의 여부를 조사하고자 한다. 그런데 이 조건은 양의 연관성 규칙을 위한 흥미도 측도를 평가하는 것이므로 Park (2011a)이 제안한 음의 연관성 규칙에 대한 흥미도 측도의 조건에 대해 충족하는지의 여부를 조사하면 다음과 같다. 위의 조건 중에서 둘 중 하나만 증명하면 다른 하나는 동일하게 증명할 수 있으므로 $NAC_{pure1}$인 $NAC_{pure}(A \Rightarrow B^c)$의 경우에 대해서만 증명하기로 한다.

[조건 1] $P(A \cap B^c) = P(A)P(B^c)$이면 $NAC_{pure}(A \Rightarrow B^c)$의 값은 0이 되고, $P(A^c \cap B) = P(A^c)P(B)$이면 $NAC_{pure}(A^c \Rightarrow B)$의 값은 0이 된다.

(증명) : $P(A \cap B^c) = P(A)P(B^c)$이면 $P(B^c|A) = P(B^c)$이고, $P(B^c|A^c) = P(B^c)$가 되므로 정의된 식 (2.3)으로부터 $NAC_{pure}(A \Rightarrow B^c)$의 값은 0이 된다.

[조건 2] $NAC_{pure}(A \Rightarrow B^c)$는 $P(B^c)$의 값에 따라 단조 감소하고 $NAC_{pure}(A^c \Rightarrow B)$는 $P(B)$의 값에 따라 한다.

(증명) : 식 (2.2)의 $NAC_{pure}(A \Rightarrow B^c)$를 정리하면 다음의 식을 얻을 수 있다.

$$NAC_{pure}(A \Rightarrow B^c) = \frac{P(A \cap B^c) - P(A)P(B^c)}{[1 - P(A)]P(A \cap B^c)} \tag{2.3}$$

이로부터 $P(B^c)$의 값이 증가함에 따라 $NAC_{pure}(A \Rightarrow B^c)$는 단조 감소하는 것을 알 수 있다.

[조건 3] $NAC_{pure}(A \Rightarrow B^c)$는 $P(A \cap B^c)$의 값에 따라 단조 증가하고, $NAC_{pure}(A^c \Rightarrow B)$는 $P(A^c \cap B)$의 값에 따라 단조 증가한다.

(증명) : 식 (2.3)을 다시 정리하면 다음과 같이 표현할 수 있다.

$$NAC_{pure}(A \Rightarrow B^c) = \frac{1 - P(A)P(B^c)/P(A \cap B^c)}{[1 - P(A)]} \tag{2.4}$$

식 (2.4)로부터 $P(A \cap B^c)$의 값이 증가함에 따라 $NAC_{pure}(A \Rightarrow B^c)$는 단조 증가하는 것을 알 수 있다.

## 3. 적용 예제

본 절에서는 기존의 음의 신뢰도 및 음의 순수 신뢰도와 본 논문에서 제안하는 음의 기여 순수 신뢰도를 비교하기 위해 Park (2011a)에서 고려한 예제를 변형하여 이용하고자 한다. 이 예제는 데이터베이스에 있는 총 트랜잭션의 수 ($t$)를 100명으로 하고, 항목 집합 A는 구매한 치간 칫솔을 구매한 (1) 사람 수와 치간 칫솔을 구매하지 않은 (0) 사람 수를 각각 50명으로 하였다. 또한 항목 집합 B를 치실을 구매한 (1) 사람 수를 70명으로 하고 치실을 구매하지 않은 (0) 사람의 수를 30명으로 하였다. 이 논문에서는 음의 연관성 규칙을 평가하기 위한 측도들을 비교하는 것이므로 먼저 $n(A \cap B^c)$인 $b$의 증감 경향에 따라 측도들의 변화하는 양성을 비교하고자 한다. 이를 정리하면 Table 3.1과 같다.

**Table 3.1** Simulation data(1)

| | | B | | total |
|---|---|---|---|---|
| | | 1 | 0 | |
| A | 1 | $50 - b$ | $b$ | 50 |
| | 0 | $20 + b$ | $30 - b$ | 50 |
| total | | 70 | 30 | 100 |

이 표에서 불일치 발생 빈도 $b$가 취할 정수 값의 범위를 정하면 다음과 같다.

$$0 \leq b \leq 30$$

이로부터 $b$값의 변화에 따른 음의 신뢰도, 음의 순수 신뢰도, 그리고 음의 기여 순수 신뢰도를 미니탭 16의 통계 소프트웨어를 이용하여 계산하면 다음의 Table 3.2와 같은 결과를 얻을 수 있다. 여기서 $a = n(A = 1, B = 1), b = n(A = 1, B = 0), c = n(A = 0, B = 1)$, 그리고 $d = n(A = 0, B = 0)$을 의미한다. 이 표로부터 알 수 있는 바와 같이 불일치 발생 빈도 $b$의 값이 증가함에 따라 음의 신뢰도, 음의 순수 신뢰도, 그리고 음의 기여 순수 신뢰도는 모두 증가하고 있다. 또한 음의 신뢰도 $NC_1$과 $NC_2$는 모두 양의 값을 가지므로 방향이 없으나 음의 순수 신뢰도 $NC_{pure1}$과 $NC_{pure2}$, 그리고 음의 기여 순수 신뢰도 $NAC_{pure1}$과 $NAC_{pure2}$는 그 부호에 의해 연관성 규칙의 방향을 알 수 있어서 두 항목 간에 양의 연관성 규칙이 있는지 아니면 음의 연관성 규칙이 있는지를 파악할 수 있다. 그리고 $b$가 한 단위 증가함에 따라 음의 기여 순수 신뢰도의 증가폭이 가장 크고, 그 다음이 음의 순수 신뢰도이며, 음의 신뢰도의 증가폭이 가장 작다. 따라서 음의 연관성 규칙 평가기준으로 음의 기여 순수

신뢰도를 사용하는 것이 가장 구별이 잘 된다고 할 수 있다. Table 3.2를 좀 더 구체적으로 살펴보면, $a$=39, $b$=11, $c$=31, 그리고 $d$=19인 경우 만약 최저 신뢰도의 기준값이 0.2이라고 하면 기존의 음의 신뢰도 값이 0.22이므로 음의 연관성이 있다고 할 수 있다. 그러나 이 연관성 규칙은 흥미로운 규칙으로 보기 어렵다. 왜냐하면 $P(B^c|A) = 0.22$는 전체 트랜잭션을 선택했을 때 $B$가 포함되지 않을 확률인 $P(B^c) = 0.3$보다 작기 때문이다. 오히려 이 규칙은 양의 연관성이 있는 것으로 판단하여야 할 것이다.

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

| $a$ | $b$ | $c$ | $d$ | $NC_1$ | $NC_{pure1}$ | $NAC_{pure1}$ | $NC_2$ | $NC_{pure2}$ | $NAC_{pure2}$ |
|---|---|---|---|---|---|---|---|---|---|
| 45 | 5 | 25 | 25 | 0.1000 | -0.4000 | -4.0000 | 0.5000 | -0.4000 | -0.8000 |
| 44 | 6 | 26 | 24 | 0.1200 | -0.3600 | -3.0000 | 0.5200 | -0.3600 | -0.6923 |
| 43 | 7 | 27 | 23 | 0.1400 | -0.3200 | -2.2857 | 0.5400 | -0.3200 | -0.5926 |
| 42 | 8 | 28 | 22 | 0.1600 | -0.2800 | -1.7500 | 0.5600 | -0.2800 | -0.5000 |
| 41 | 9 | 29 | 21 | 0.1800 | -0.2400 | -1.3333 | 0.5800 | -0.2400 | -0.4138 |
| 40 | 10 | 30 | 20 | 0.2000 | -0.2000 | -1.0000 | 0.6000 | -0.2000 | -0.3333 |
| 39 | 11 | 31 | 19 | 0.2200 | -0.1600 | -0.7273 | 0.6200 | -0.1600 | -0.2581 |
| 38 | 12 | 32 | 18 | 0.2400 | -0.1200 | -0.5000 | 0.6400 | -0.1200 | -0.1875 |
| 37 | 13 | 33 | 17 | 0.2600 | -0.0800 | -0.3077 | 0.6600 | -0.0800 | -0.1212 |
| 36 | 14 | 34 | 16 | 0.2800 | -0.0400 | -0.1429 | 0.6800 | -0.0400 | -0.0588 |
| 35 | 15 | 35 | 15 | 0.3000 | 0.0000 | 0.0000 | 0.7000 | 0.0000 | 0.0000 |
| 34 | 16 | 36 | 14 | 0.3200 | 0.0400 | 0.1250 | 0.7200 | 0.0400 | 0.0556 |
| 33 | 17 | 37 | 13 | 0.3400 | 0.0800 | 0.2353 | 0.7400 | 0.0800 | 0.1081 |
| 32 | 18 | 38 | 12 | 0.3600 | 0.1200 | 0.3333 | 0.7600 | 0.1200 | 0.1579 |
| 31 | 19 | 39 | 11 | 0.3800 | 0.1600 | 0.4211 | 0.7800 | 0.1600 | 0.2051 |
| 30 | 20 | 40 | 10 | 0.4000 | 0.2000 | 0.5000 | 0.8000 | 0.2000 | 0.2500 |
| 29 | 21 | 41 | 9 | 0.4200 | 0.2400 | 0.5714 | 0.8200 | 0.2400 | 0.2927 |
| 28 | 22 | 42 | 8 | 0.4400 | 0.2800 | 0.6364 | 0.8400 | 0.2800 | 0.3333 |

따라서 기존의 음의 신뢰도를 사용하게 되면 양의 연관성 규칙을 음의 연관성 규칙으로 해석하는 오류를 범할 수 있다. 이 경우에 음의 순수 신뢰도의 값은 -0.16이고, 음의 기여 순수 신뢰도의 값은 -0.7273으로 나타나고 있으므로 두 측도에 의해서는 양의 연관성이 있는 것으로 판단할 수 있다. 또한 $a$=38, $b$=12, $c$=32, $d$=18인 경우와 $a$=37, $b$=13, $c$=33, $d$=17인 경우를 비교해보면 음의 순수 신뢰도는 각각 -0.12와 -0.08로 나타난 반면에 음의 기여 순수 신뢰도의 값은 각각 -0.5와 -0.3077로 나타났다. 따라서 음의 기여 순수 신뢰도의 값이 음의 순수 신뢰도 보다 절대값이 훨씬 크게 나타나는 동시에 두 경우의 차이는 음의 기여 순수 신뢰도가 더 크게 차이가 나타났다. 따라서 음의 기여 순수 신뢰도가 기존의 다른 측도들에 비해 음의 연관성 규칙의 유무를 더 확연하게 구분할 수 있다. 또한 $a$=35, $b$=15, $c$=35, $d$=15인 경우에는 음의 기여 순수 신뢰도는 모두 0으로 나타나고 있어서 이 값을 중심으로 양의 연관관계와 음의 연관관계를 구분할 수 있다.

이번에는 불일치 발생 빈도 $c$의 변화에 따른 음의 신뢰도, 음의 순수 신뢰도, 그리고 음의 기여 순수 신뢰도의 변화하는 양상을 알아보기 위해 Table 3.3과 같은 데이터를 활용하고자 한다.

**Table 3.3** Simulation data(2)

| | | B | | total |
|---|---|---|---|---|
| | | 1 | 0 | |
| A | 1 | $30 - c$ | $50 + c$ | 80 |
| | 0 | $c$ | $20 - c$ | 20 |
| total | | 30 | 70 | 100 |

이 표에서 불일치 발생 빈도 $c$가 취할 정수 값의 범위를 정하면 다음과 같다.

$$0 \leq c \leq 20$$

Table 3.3으로부터 $c$의 값이 증가함에 따라 음의 신뢰도, 음의 순수 신뢰도, 음의 기여 순수 신뢰도를 계산하면 다음의 Table 3.4와 같은 결과를 얻을 수 있다. Table 3.4로부터 알 수 있는 바와 같이 불일치 발생 빈도 $c$의 값이 증가함에 따라 음의 신뢰도, 음의 순수 신뢰도, 그리고 음의 기여 순수 신뢰도는 모두 증가하고 있다. Table 3.2에서와 마찬가지로 음의 신뢰도는 모두 양의 값을 가지므로 방향이 없으나 음의 순수 신뢰도와 음의 기여 순수 신뢰도는 모두 그 부호에 의해 연관성 규칙의 방향을 파악할 수 있다. 또한 $c$가 한 단위 증가함에 따라 음의 기여 순수 신뢰도의 증가폭이 가장 크고, 그 다음이 음의 순수 신뢰도이며, 음의 신뢰도의 증가폭이 가장 작다. 따라서 음의 연관성 규칙 평가기준으로 음의 기여 순수 신뢰도를 사용하는 것이 가장 구별이 잘 된다고 할 수 있다. 이 표에서도 $a$=27, $b$=53, $c$=3, $d$=17인 경우 만약 최저 신뢰도의 기준값이 0.5라고 하면 기존의 음의 신뢰도 값이 0.6625이므로 음의 연관성이 있다고 할 수 있다. 그러나 이 음의 연관성 규칙은 흥미로운 규칙으로 보기 어렵다. 왜냐하면 $P(B^c|A) = 0.6625$는 $P(B^c) = 0.7$보다 작기 때문이다. 이 경우에 음의 순수 신뢰도의 값은 -0.1875이고, 음의 기여 순수 신뢰도의 값은 -0.2830으로 나타나고 있으므로 두 측도에 의해서는 양의 연관성이 있는 것으로 판단할 수 있다. 또한 음의 기여 순수 신뢰도의 값이 음의 순수 신뢰도 보다 절대값이 크게 나타나고 있고 변화량도 더 크므로 음의 기여 순수 신뢰도가 음의 연관성 규칙의 유무를 더 확연하게 구분해준다고 할 수 있다.

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

| $a$ | $b$ | $c$ | $d$ | $NC_1$ | $NC_{pure1}$ | $NAC_{pure1}$ | $NC_2$ | $NC_{pure2}$ | $NAC_{pure2}$ |
|---|---|---|---|---|---|---|---|---|---|
| 29 | 51 | 1 | 19 | 0.6375 | -0.3125 | -0.4902 | 0.0500 | -0.3125 | -6.2500 |
| 28 | 52 | 2 | 18 | 0.6500 | -0.2500 | -0.3846 | 0.1000 | -0.2500 | -2.5000 |
| 27 | 53 | 3 | 17 | 0.6625 | -0.1875 | -0.2830 | 0.1500 | -0.1875 | -1.2500 |
| 26 | 54 | 4 | 16 | 0.6750 | -0.1250 | -0.1852 | 0.2000 | -0.1250 | -0.6250 |
| 25 | 55 | 5 | 15 | 0.6875 | -0.0625 | -0.0909 | 0.2500 | -0.0625 | -0.2500 |
| 24 | 56 | 6 | 14 | 0.7000 | 0.0000 | 0.0000 | 0.3000 | 0.0000 | 0.0000 |
| 23 | 57 | 7 | 13 | 0.7125 | 0.0625 | 0.0877 | 0.3500 | 0.0625 | 0.1786 |
| 22 | 58 | 8 | 12 | 0.7250 | 0.1250 | 0.1724 | 0.4000 | 0.1250 | 0.3125 |
| 21 | 59 | 9 | 11 | 0.7375 | 0.1875 | 0.2542 | 0.4500 | 0.1875 | 0.4167 |
| 20 | 60 | 10 | 10 | 0.7500 | 0.2500 | 0.3333 | 0.5000 | 0.2500 | 0.5000 |
| 19 | 61 | 11 | 9 | 0.7625 | 0.3125 | 0.4098 | 0.5500 | 0.3125 | 0.5682 |
| 18 | 62 | 12 | 8 | 0.7750 | 0.3750 | 0.4839 | 0.6000 | 0.3750 | 0.6250 |
| 17 | 63 | 13 | 7 | 0.7875 | 0.4375 | 0.5556 | 0.6500 | 0.4375 | 0.6731 |
| 16 | 64 | 14 | 6 | 0.8000 | 0.5000 | 0.6250 | 0.7000 | 0.5000 | 0.7143 |
| 15 | 65 | 15 | 5 | 0.8125 | 0.5625 | 0.6923 | 0.7500 | 0.5625 | 0.7500 |
| 14 | 66 | 16 | 4 | 0.8250 | 0.6250 | 0.7576 | 0.8000 | 0.6250 | 0.7813 |

마지막으로 동시 비발생 빈도의 변화에 따른 음의 신뢰도, 음의 순수 신뢰도, 그리고 음의 기여 순수 신뢰도의 변화하는 양상을 알아보았다. 위의 경우와 마찬가지로 음의 신뢰도는 모두 양의 값을 가지므로 방향이 없으나 음의 순수 신뢰도와 음의 기여 순수 신뢰도는 모두 그 부호에 의해 연관성 규칙의 방향을 파악할 수 있었다. 뿐만 아니라 음의 기여 순수 신뢰도의 값이 음의 순수 신뢰도에 비해 절대값이 크고, 동시 비발생 빈도의 증가에 따른 변화량도 더 크므로 음의 기여 순수 신뢰도가 음의 연관성 규칙의 유무를 더 확연하게 구분해준다는 사실도 확인할 수 있었다. 이 경우에도 음의 기여 순수 신뢰도의 값이 음의 순수 신뢰도 보다 절대값이 크게 나타나므로 음의 연관성 규칙의 유무를 더 확연하게 구분할 수 있다. 또한 다른 셀의 값의 변화에 따른 신뢰도와 순수 신뢰의 값을 비교하기

위해 각 셀의 값을 바꾸어 가면서 실험해본 결과, 신뢰도는 모두 양의 값을 가지므로 방향이 없으며, 음의 순수 신뢰도와 음의 기여 순수 신뢰도는 그 부호에 의해 연관성 규칙의 방향을 알 수 있다는 사실을 확인하였다.

## 4. 결론

음의 연관성 규칙 마이닝은 어느 항목이 발생하면 다른 항목은 발생하지 않는다는 규칙을 생성하는 것이다. 기존의 음의 신뢰도는 양의 연관성을 가지는 연관성 규칙을 음의 연관성 규칙으로 선택하게 되는 오류를 범할 수 있다. 또한 음의 순수 신뢰도는 순수하게 특정 항목에 의해서만 결과가 얼마인가를 나타내주는 측도인 동시에 그 부호에 의해 음의 관련성과 양의 관련성을 판단할 수는 있으나 수식에 포함된 두 확률값의 크기가 동일하면 순수 신뢰도의 값도 동일하게 되는 단점을 가지고 있다. 이러한 문제를 해결하기 위해 본 논문에서는 음의 기여 순수 신뢰도를 제안하는 동시에 이에 대한 흥미도 측도 조건의 충족여부도 알아보았다.

또한 예제를 통하여 살펴본 결과, 불일치 발생 빈도의 값이 증가함에 따라 음의 신뢰도, 음의 순수 신뢰도, 그리고 음의 기여 순수 신뢰도는 모두 증가하고 있다는 사실을 확인할 수 있었다. 또한 음의 신뢰도는 모두 양의 값을 가지므로 방향이 없으나 음의 순수 신뢰도와 음의 기여 순수 신뢰도는 모두 그 부호에 의해 연관성 규칙의 방향을 알 수 있었다. 불일치 발생 빈도의 값이 한 단위 증가함에 따라 음의 기여 순수 신뢰도의 증가폭이 가장 크고, 그 다음이 음의 순수 신뢰도이며, 음의 신뢰도의 증가폭이 가장 작다. 따라서 음의 연관성 규칙 평가기준으로 음의 기여 순수 신뢰도를 사용하는 것이 가장 구별이 잘 된다는 사실도 확인할 수 있었다. 동시 비발생 빈도의 변화에 따른 음의 신뢰도, 음의 순수 신뢰도, 그리고 음의 기여 순수 신뢰도의 변화하는 양상에 대해서도 살펴보았는데 불일치 발생 빈도의 경우와 유사한 결과를 얻을 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.

Bala, P. K. (2009). A technique for mining negative association rules. *Proceedings of the 2nd Bangalore Annual Compute Conference*, 23-23.

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.

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. and Kamber, M. (2006). *Data mining : Concepts and techniques*, Morgam Kaufmann, USA.

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.

Hwang, J. and Kim, J. (2003). Target marketing using inverse association rule. *Journal of Intelligence and Information Systems*, **9**, 195-209.

Lee, J., Park, S., Kang, Y., Park, S. and Lee, J. (2003). Finding negative association rules with Boolean analyzer. *Proceedings of the Korean Institute of Information Scientists and Engineers*, **30**, 187-189.

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.

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.

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.

Saygin Y., Vassilios S. V. and Clifton C.(2002). Using unknowns to prevent discovery of association rules. *Proceedings of 2002 Conference on Research Issues in Data Engineering*, 45-54.

Shang, S., Dong, X., Geng, R. and Zhao, L. (2008). Mining negative association rules in multi-database. *Proceedings of the Fifth International Conference on Fuzzy Systems and Knowledge Discovery*, 596-599.

Sim, A., Indrawan, M. and Srinivasan, B. (2008). The importance of negative associations and the discovery of association rule pairs. *International Journal of Business Intelligence and Data Mining*, **3**, 158-176.

*Journal of the Korean Data & Information Science Society*
*2012,* **23**(5), 939–948

http://dx.doi.org/10.7465/jkdi.2012.23.5.939
한국데이터정보과학회지

# Negatively attributable and pure confidence for generation of negative association rules

Hee Chang Park[1]

[1]Department of Statistics, Changwon National University

Received 8 August 2012, revised 28 August 2012, accepted 3 September 2012

**Abstract**

The most widely used data mining technique is to explore association rules. This technique has been used to find the relationship between items in a massive database based on the interestingness measures such as support, confidence, lift, etc. Association rules are frequently used by retail stores to assist in marketing, advertising, floor placement, and inventory control.In general, association rule technique generates the rule, 'If A, then B.', whereas negative association rule technique generates the rule, 'If A, then not B.', or 'If not A, then B.'. We can determine whether we promote other products in addition to promote its products only if we add negative association rules to existing association rules. In this paper, we proposed the negatively attributable and pure confidence to overcome the problems faced by negative association rule technique, and then we checked three conditions for interestingness measure. The comparative studies with negative confidence, negatively pure confidence, and negatively attributable and pure confidence are shown by numerical examples. The results show that the negatively attributable and pure confidence is better than negative confidence and negatively pure confidence.

*Keywords*: Interestingness measure, negatively attributable and pure confidence, negative confidence, negatively pure confidence.

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