Journal of the Korean Data & Information Science Society
2011, **22**(2), 179–188

한국데이터정보과학회지

음의 순수 연관성 규칙 평가 기준의 제안

박희창[1]

[1]창원대학교 통계학과

접수 2011년 1월 16일, 수정 2011년 2월 11일, 게재확정 2011년 2월 14일

요 약

연관성 규칙은 방대한 데이터베이스에서 항목간의 관계를 명확히 수치화함으로써 그들간의 관련성을 표시해주는 기법으로 데이터 마이닝 기법들 중에서 가장 많이 활용되고 있다. 어느 항목이 발생하면 다른 항목도 발생한다는 규칙을 발견하기 위한 기법이 연관성 규칙이라면 음의 연관성 규칙 마이닝은 어느 항목이 발생하면 다른 항목도 발생하지 않는다는 규칙을 찾아내는 기법이다. 기존의 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 음의 연관성 규칙의 단점을 보완할 수 있는 음의 순수 연관성 규칙의 측도들을 제시하고 흥미도 측도가 가져야 할 조건들을 조사하였으며, 예제 데이터를 활용하여 음의 순수 연관성 규칙의 유용성에 대해 살펴보았다.

주요용어: 음의 순수 신뢰도, 음의 순수 연관성 규칙, 음의 순수 지지도, 음의 순수 향상도, 흥미도 측도.

1. 서론

오늘날 많은 조직이 최적의 전략이나 의사결정을 위한 의미 있는 고급정보를 확보하기 위해 데이터 마이닝 기법을 활용하고 있다. 데이터 마이닝 기법들 중에서도 연관성 규칙은 가장 많이 활용되고 있는 기법으로 방대한 데이터베이스에서 항목간의 관계를 명확히 수치화함으로써 관련성을 표시하여 주기 때문에 현장에서 직접 적용이 가능하다. 일반적으로 연관성 규칙에서는 항목간의 연관성을 반영하는 기준인 지지도 (support), 신뢰도 (confidence), 향상도 (lift) 등의 흥미도 측도 (interestingness measure)를 바탕으로 관련성 여부를 측정한다. 연관성 규칙은 Agrawal 등 (1993)에 의해 처음 소개된 이후, 국내외 많은 학자들이 연관성 측정에 관한 연구를 수행한 바 있으며, 특히 주목할 만한 연구로는 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)이 있으며, 국내연구로는 Cho와 Park (2007, 2008), Choi와 Park (2008), 그리고 Park (2008) 등이 있다. 한편, Han과 Kamber (2006)은 연관성 규칙을 여러 형태로 분류하였는데 먼저 데이터의 유형에 따라 불리언 연관성 규칙 (boolean association rule)과 정량적 연관성 규칙 (quantitative association rule)으로 분류하였고, 데이터의 차원 수에 따라 1차원 연관성 규칙과 다중차원 연관성 규칙으로 분류하였다. 그리고 규칙에 포함된 요약의 수준에 따라 단일 수준 연관성 규칙 (single-level association rule)과 다수준 연관성 규칙 (multi-level association rule)으로, 응용 결과에 따라 순차 연관성 규칙 (sequences association rule)과 음의 연관성 규칙 (negative association rule)으로 분류한 바 있다. 일반적으로 연

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

관성 규칙 마이닝에서는 어느 항목이 발생하면 다른 항목도 발생한다는 규칙을 발견하는 기법인 반면에, 음의 연관성 규칙 마이닝은 어느 항목이 발생하면 다른 항목도 발생하지 않는다는 규칙을 찾아내는 것이다. 음의 연관성 규칙에 대한 연구로는 Yuan 등 (2002), Koh와 Pears (2007), Sharma 등 (2007), Sim 등 (2008), Shang 등 (2008), 그리고 Bala (2009)이 제시한 바 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 (황준현, 김재련, 2003). 예를 들어 치아의 청결을 위해 치실을 구매하는 사람은 치간 칫솔을 구매하지 않는 경향이 많다는 정보가 의미 있는 동시에 치간 칫솔을 구매하는 사람은 치실을 구매하지 않는 경향이 많다는 정보가 의미 있다고 하면 치실이나 치간 칫솔을 사용하는 사람들은 치아의 건강에 많은 관심을 가진다고 볼 수 있다. 따라서 치간 칫솔을 팔기 위해서는 치실을 구매한 사람들에게 치간 칫솔의 유용성을, 그리고 치실을 팔기 위해서는 치간 칫솔을 구매한 사람들에게 유용성을 홍보하여 구매를 유도하는 마케팅 전략도 필요하다. 이러한 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 지지도와 신뢰도는 방향성이 없으며, 향상도는 방향성이 없을 뿐만 아니라 범위의 제한도 없다. 본 논문에서는 이러한 단점을 보완할 수 있는 음의 순수 연관성 규칙 (negatively pure association rule)의 측도인 음의 순수 지지도 (negatively pure support), 음의 순수 신뢰도 (negatively pure confidence), 그리고 음의 순수 향상도 (negatively pure lift)를 정의함으로써 음의 순수 연관성 규칙을 제안하고자 한다. 2절에서는 음의 순수 연관성 규칙을 정의하고, Piatetsky-Shapiro (1991)가 제안한 흥미도 측도가 가져야 할 조건들을 조사한 후 측도들의 성질을 규명한다. 3절에서는 모의실험 데이터를 활용한 음의 연관성 규칙과의 비교를 통해 음의 순수 연관성 규칙의 유용성에 대해 살펴본 후, 마지막으로 4절에서 결론을 내리고자 한다.

2. 음의 순수 연관성 규칙의 평가 기준

′A이면 not B이다.′ 또는 ′not A이면 B이다.′로 정의되는 음의 연관성 규칙의 평가기준은 다음과 같다.

- 음의 지지도 : $NS(X \Rightarrow Y^c) = P(X \cap Y^c)$, $NS(X^c \Rightarrow Y) = P(X^c \cap Y)$
- 음의 신뢰도 : $NC(X \Rightarrow Y^c) = P(Y^c|X)$, $NC(X^c \Rightarrow Y) = P(Y|X^c)$
- 음의 향상도 : $NL(X \Rightarrow Y^c) = P(Y^c|X)/P(Y^c)$, $NL(X^c \Rightarrow Y) = P(Y|X^c)/P(Y)$

여기서 X^c와 Y^c의 의미는 각각 X와 Y가 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 지지도와 신뢰도의 범위는 [0, 1]이나 방향성이 없으며, 향상도는 방향성이 없을 뿐만 아니라 범위의 제한도 없다. 본 절에서는 이러한 단점을 보완할 수 있는 음의 순수 연관성 규칙의 측도인 음의 순수 지지도, 음의 순수 신뢰도, 그리고 음의 순수 향상도를 정의함으로써 음의 순수 연관성 규칙을 제안하고자 한다.

이를 위해 음의 순수 연관성 규칙의 평가기준들에 대한 흥미도 측도의 조건 만족 여부를 조사하여야 하는데 Piatetsky-Shapiro가 제안한 흥미도 측도의 조건들은 다음과 같이 변형된다.

[조건 1] $P(X \cap Y^c) = P(X)P(Y^c)$ 또는 $P(X^c \cap Y) = P(X^c)P(Y)$이면 흥미도 측도의 값은 0이 된다.

[조건 2] 흥미도 측도는 $P(Y^c)$ 또는 $P(X^c)$의 값에 따라 단조 감소한다.

[조건 3] 흥미도 측도는 $P(X \cap Y^c)$ 또는 $P(X^c \cap Y)$의 값에 따라 단조 증가한다.

2.1. 음의 순수 지지도

음의 지지도는 항목 집합 X는 발생하고 항목 집합 Y는 발생하지 않는 거래의 비율, 또는 X는 발생하지 않고 Y는 발생하는 거래의 비율로 나타낸다. 반면에 음의 순수 지지도는 항목 집합 X는 발생하고 항목 집합 Y는 발생하지 않는 거래의 비율과 X와 Y가 동시에 발생하지 않는 거래 비율의 차이, 또는 X는 발생하지 않고 항목 집합 Y는 발생하는 거래의 비율과 X와 Y가 동시에 발생하는 거래 비율의 차이를 의미하며, 다음과 같이 정의한다.

$$NS_{pure}(X \Rightarrow Y) = \begin{cases} P(X \cap Y^c) - P(X^c \cap Y^c) \\ P(X^c \cap Y) - P(X \cap Y) \end{cases} \tag{2.1}$$

식 (2.1)의 우변 첫 번째 식을 $NS_{pure}(X \Rightarrow Y^c)$으로, 두 번째 식은 $NS_{pure}(X^c \Rightarrow Y)$으로 나타내기로 한다. 음의 순수 지지도는 Piatetsky-Shapiro가 제안한 흥미도 측도의 조건이 충족되지는 않으나, 규칙 생성을 위한 첫 번째 단계가 최소 지지도를 만족시키는 빈발항목집합 생성과정이므로 의미 있는 규칙 발견을 위해 필요한 측도이며, 다음과 같은 성질을 가지고 있다.

[성질 1] $NS_{pure}(X \Rightarrow Y)$가 양의 값을 가지면 두 항목집합 X와 Y가 음의 연관성 규칙이 성립하는 반면에 $NS_{pure}(X \Rightarrow Y)$가 음이면 X와 Y가 양 또는 역의 연관성을 가 지며, $NS_{pure}(X \Rightarrow Y)$가 0이면 X와 Y가 서로 독립적인 관계임을 의미한다.

(설명): 식 (2.1)의 첫 번째 경우만을 기술하기로 한다. $NS_{pure}(X \Rightarrow Y^c) > 0$이면 $P(X \cap Y^c) > P(X^c \cap Y^c)$이므로 X는 포함되나 Y가 포함되지 않는 거래의 비율이 X와 Y가 동시에 발생하지 않는 거래의 비율보다 크므로 두 항목 간에는 음의 연관성이 존재한다고 볼 수 있다. 만약 $NS_{pure}(X \Rightarrow Y^c) < 0$이면 $P(X \cap Y^c) < P(X^c \cap Y^c)$이므로 X는 포함되나 Y가 포함되지 않는 거래의 비율이 X와 Y가 동시에 발생하지 않는 거래의 비율보다 작으므로 두 항목 간에는 역의 연관성이 존재한다고 볼 수 있다. 그리고 $NS_{pure}(X \Rightarrow Y^c) = 0$이면 $P(X \cap Y^c) = P(X^c \cap Y^c)$이므로 X는 포함되나 Y가 포함되지 않는 거래의 비율과 X와 Y가 동시에 발생하지 않는 거래의 비율이 같으므로 X와 Y는 서로 독립관계라고 볼 수 있다.

[성질 2] $NS_{pure}(X \Rightarrow Y^c)$값의 도메인은 $[-1,\ 1]$이다.

(설명): $NS_{pure}(X \Rightarrow Y^c)$의 값이 1이라는 것은 $P(X \cap Y^c)$가 1의 값을 가지므로 X는 포함되나 Y가 포함되지 않는 거래의 비율은 1이고 X와 Y가 동시에 발생하지 않는 거래의 비율이 0이라는 의미이다. 그리고 $NS_{pure}(X \Rightarrow Y^c)$ 값이 -1이라는 의미는 X는 포함되나 Y가 포함되지 않는 거래의 비율이 0이고 X와 Y가 동시에 발생하지 않는 거래의 비율은 1이 된다는 것을 의미한다.

[성질 3] $NS_{pure}(X \Rightarrow Y^c)$와 $NS_{pure}(X^c \Rightarrow Y)$의 값은 동일하지 않다.

(설명): 식 (2.1)을 통해서 알 수 있는 바와 같이 $NS_{pure}(X \Rightarrow Y)$ 는 $X \Rightarrow Y^c$와 $X^c \Rightarrow Y$의 값이 다르므로 그 크기와 부호를 이용하여 양과 음의 관련성을 파악할 수 있다.

2.2. 음의 순수 신뢰도

음의 연관성 규칙에서의 신뢰도는 항목 집합 X가 포함된 거래 비율 중 항목 집합 X는 포함되고 항목 집합 Y는 포함되지 않은 거래의 비율 또는 항목 집합 X가 포함되지 않은 거래 비율 중 항목 집합 Y는 포함되고 항목 집합 X는 포함되지 않은 거래의 비율이다. 반면에 음의 순수 신뢰도는 X가 포함된 거래 중 Y가 포함되지 않은 거래의 비율과 X가 포함되지 않은 거래 비율 중 Y도 포함되지 않은 거래의 비율의 차이 또는 X가 포함되지 않은 거래 중 Y가 포함된 거래의 비율과 X가 포함된 거래 비율 중 Y가 포

함된 거래의 비율의 차이를 의미하며, 다음과 같이 정의한다.

$$NC_{pure}(X \Rightarrow Y) = \begin{cases} P(Y^c|X) - P(Y^c|X^c) \\ P(Y|X^c) - P(Y|X) \end{cases} \tag{2.2}$$

여기서도 위와 마찬가지로 식 (2.2)의 우변 첫 번째 식을 $NC_{pure}(X \Rightarrow Y^c)$으로, 두 번째 식은 $NC_{pure}(X^c \Rightarrow Y)$으로 나타내기로 한다.

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. 또한 순수 신뢰도의 값의 범위는 식 (2.2)로부터 [-1, 1]임을 알 수 있다.

음의 순수 신뢰도가 Piatetsky-Shapiro가 제안한 흥미도 측도의 조건 만족 여부에 대해 조사하면 다음과 같다. 위의 조건 중에서 둘 중 하나만 증명하면 다른 하나는 동일하게 증명할 수 있으므로 $NC_{pure}(X \Rightarrow Y^c)$의 경우에 대해서만 증명하기로 한다.

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

(증명): $P(X \cap Y^c) = P(X)P(Y^c)$이면 $P(Y^c|X) = P(Y^c)$ 이고, $P(Y^c|X^c) = P(Y^c)$가 되므로 $NC_{pure}(X \Rightarrow Y^c)$의 값은 0이 된다.

[조건 2] $NC_{pure}(X \Rightarrow Y^c)$는 $P(Y^c)$의 값에 따라 단조 감소하고 $NC_{pure}(X^c \Rightarrow Y)$는 $P(Y)$의 값에 따라 한다.

(증명): 식 (2.2)의 $NC_{pure}(X \Rightarrow Y^c)$를 정리하면 다음의 식을 얻을 수 있다.

$$NC_{pure}(X \Rightarrow Y^c) = \frac{P(X \cap Y^c) - P(X)P(Y^c)}{P(X)[1 - P(X)]} \tag{2.3}$$

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

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

(증명): 식 (2.3)으로부터 $P(X \cap Y^c)$의 값이 증가함에 따라 $NC_{pure}(X \Rightarrow Y^c)$는 단조 증가하는 것을 알 수 있다.

2.3. 음의 순수 향상도

음의 향상도는 항목 집합 X를 구매한 경우 그 거래가 항목 집합 Y를 포함하지 않는 경우와 Y가 구매되지 않는 경우의 비, 또는 항목 집합 X를 구매하지 않은 경우 그 거래가 항목 집합 Y를 포함하는 경우와 Y가 임의로 구매된 경우의 비를 의미한다. 반면에 음의 순수 향상도는 X가 포함된 거래 중 Y가 포함되지 않은 거래의 비율과 X가 포함되지 않은 거래 비율 중 Y가 포함되지 않은 거래의 비율의 차이와 Y가 임의로 구매되지 않는 경우의 비, 또는 X가 포함되지 않은 거래 중 Y가 포함된 거래의 비율과 X가 포함된 거래 비율 중 Y가 포함된 거래의 비율의 차이와 Y가 임의로 구매된 경우의 비를 의미하며, 다음

과 같이 정의된다.

$$NL_{pure}(X \Rightarrow Y) = \begin{cases} \dfrac{P(Y^c|X) - P(Y^c|X^c)}{P(Y^c)} \\ \dfrac{P(Y|X^c) - P(Y|X)}{P(Y)} \end{cases} \tag{2.4}$$

이 식에서도 우변의 첫 번째 식을 $NL_{pure}(X \Rightarrow Y^c)$으로, 두 번째 식은 $NL_{pure}(X^c \Rightarrow Y)$으로 나타내기로 한다.

음의 순수 향상도는 항목집합 X가 포함된 트랜잭션 (주문이나 판매와 같은 하나의 외부 거래를 기록하기 위한 일련의 처리 동작)에서 항목집합 Y가 함께 발견되지 않고 있다는 사실이 항목집합 X를 포함하지 않는 트랜잭션들만이 가지는 고유한 특성인지, 아니면 항목집합 X가 포함되지 않은 트랜잭션에서 항목집합 Y가 발견된 사실이 항목집합 X를 포함하지 않는 트랜잭션들만이 가지는 고유한 특성인지를 나타내는 것이다. 만약 항목집합 X가 포함되지 않은 트랜잭션에서도 항목집합 Y가 발견되지 않는 경우가 많으면 음의 순수 향상도의 값이 작아지며, 항목집합 X와 Y의 음의 연관성은 의미가 없게 된다. 또한 음의 순수 향상도의 범위는 $[-1/P(Y^c), 1/P(Y^c)]$ 또는 $[-1/P(Y), 1/P(Y)]$이다.

음의 순수 향상도를 Piatetsky-Shapiro가 제안한 흥미도 측도의 조건 만족 여부에 대해 조사하면 다음과 같다. 이 경우에도 위의 조건 중에서 둘 중 하나만 증명하면 다른 하나는 동일하게 증명할 수 있으므로 $NL_{pure}(X \Rightarrow Y^c)$의 경우에 대해서만 증명하기로 한다.

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

(증명): $P(X \cap Y^c) = P(X)P(Y^c)$이면 $P(Y^c|X) = P(Y^c)$이고, $P(Y^c|X^c) = P(Y^c)$가 되므로 $NC_{pure}(X \Rightarrow Y^c)$의 [조건 1]에서의 증명에서와 같이 $NL_{pure}(X \Rightarrow Y^c)$의 값은 0이 된다.

[조건 2] $NL_{pure}(X \Rightarrow Y^c)$ 는 $P(Y^c)$의 값에 따라, 그리고 $NL_{pure}(X^c \Rightarrow Y)$는 $P(Y)$의 값에 따라 단조 감소한다.

(증명): 식 (2.3)의 결과를 활용하여 $NL_{pure}(X \Rightarrow Y^c)$를 정리하면 다음과 같은 식을 얻을 수 있다.

$$NL_{pure}(X \Rightarrow Y^c) = \frac{P(X \cap Y^c)}{P(X)[1-P(X)]P(Y^c)} - \frac{P(X)}{P(X)[1-P(X)]} \tag{2.5}$$

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

[조건 3] $NL_{pure}(X \Rightarrow Y^c)$ 는 $P(X \cap Y^c)$의 값에 따라 단조 증가한다.

(증명): 식 (2.5)로부터 $P(X \cap Y^c)$의 값이 증가함에 따라 $NL_{pure}(X \Rightarrow Y^c)$는 단조 증가하는 것을 알 수 있다.

3. 적용 예제

본 절에서는 예제를 통하여 기존의 음의 연관성 규칙의 흥미도 측도와 음의 순수 연관성 규칙을 위한 흥미도 측도를 비교하고자 한다. 먼저 데이터베이스에 있는 총 트랜잭션의 수 (t)를 85명으로 하고, 항목 집합 X는 구매한 치간 칫솔을 구매한 (1) 사람 수를 45명으로 하고 치간 칫솔을 구매하지 않은 (0) 사람 수를 40명으로 하였다. 또한 항목 집합 Y를 치실을 구매한 (1) 사람 수를 70명으로 하고 치실을 구매하지 않은 (0) 사람의 수를 15명으로 하였다. 항목 집합 X와 Y가 동시에 발생한 빈도 수, 즉 치간 칫솔과 치실을 동시에 구매한 빈도수는 a명으로 하였다. 이를 정리하면 표 3.1과 같다.

표 3.1 예제 데이터

		Y		합계
		1	0	
X	1	a	$45-a$	45
	0	$70-a$	$a-30$	40
합계		70	15	85

이 표에서 a가 취할 정수 값의 범위를 정하면 다음과 같다.

$$30 \leq a \leq 45 \tag{3.1}$$

이로부터 발생빈도에 따른 음의 신뢰도와 음의 순수 신뢰도 $NS_{pure}(X \Rightarrow Y^c)$를 계산하면 다음의 표 3.2와 같은 결과를 얻을 수 있다. 여기서 a=n(X=1, Y=1), b=n(X=1, Y=0), c=n(X=0, Y=1), 그리고 d=n(X=0, Y=0)을 의미한다. 이 표로부터 알 수 있는 바와 같이 음의 신뢰도는 모두 양의 값을 가지므로 방향이 없고, 음의 순수 신뢰도는 그 부호에 의해 연관성 규칙의 방향을 알 수 있다. 이 표에서 a=40, b=5, c=30, 그리고 d=10인 경우, 만약 최저 신뢰도의 기준값이 0.1이라고 하면 기존의 음의 신뢰도 값이 0.111이므로 음의 연관성이 있다고 할 수 있다. 그러나 이 연관성 규칙은 흥미로운 규칙으로 보기 어렵다. 왜냐하면 $P(Y^c|X) = 0.111$이 전체 트랜잭션을 선택했을 때 Y가 포함되지 않을 확률인 $P(Y^c) = 0.176$보다 작기 때문이다. 오히려 이 규칙은 양의 연관성이 있는 것으로 판단하여야 할 것이다. 따라서 기존의 음의 신뢰도를 사용하게 되면 양의 연관성 규칙을 음의 연관성 규칙으로 해석하는 오류를 범할 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, 순수 신뢰도는 그 부호에 의해 연관성 규칙의 방향을 알 수 있다는 사실을 확인하였다. 일반적으로 음의 순수 신뢰도 $NC_{pure}(X \Rightarrow Y^c)$의 범위는 [-1, 1]이나 본 예제에서는 a가 취할 정수 값의 범위가 제한되어 있으므로 [-0.643, 0.571]의 값을 갖게 된다.

표 3.2 음의 신뢰도와 음의 순수 신뢰도의 비교

a	b	c	d	NC	NC_{pure}
30	15	40	0	0.333	0.333
31	14	39	1	0.311	0.286
32	13	38	2	0.289	0.239
33	12	37	3	0.267	0.192
34	11	36	4	0.244	0.144
35	10	35	5	0.222	0.097
36	9	34	6	0.200	0.050
37	8	33	7	0.178	0.003
38	7	32	8	0.156	-0.044
39	6	31	9	0.133	-0.092
40	5	30	10	0.111	-0.139
41	4	29	11	0.089	-0.186
42	3	28	12	0.067	-0.233
43	2	27	13	0.044	-0.281
44	1	26	14	0.022	-0.328
45	0	25	15	0.000	-0.375

표 3.1로부터 발생빈도에 따른 음의 향상도와 음의 순수 향상도 $NL_{pure}(X \Rightarrow Y^c)$를 계산하면 표 3.3과 같다. Park (2009)이 지적한 바와 같이 향상도는 모두 양의 값을 가지므로 방향이 없으며, 향상도가 취할 수 있는 값의 범위를 알 수 없다. 이와 마찬가지로 음의 향상도도 항상 양의 값을 가지므로 방향

성이 없어서 행태적으로 해석하기가 곤란하다. 반면에 음의 순수 향상도 $NL_{pure}(X \Rightarrow Y^c)$는 그 부호에 의해 연관성 규칙의 방향을 알 수 있는 동시에 $NL_{pure}(X \Rightarrow Y^c)$와 $NL_{pure}(X^c \Rightarrow Y)$의 값이 다르므로 항목집합의 선행여부에 따라 값이 변함을 알 수 있다. 또한 일반적으로 $NL_{pure}(X \Rightarrow Y^c)$가 취할 수 있는 값의 범위는 $[-1/P(Y^c), 1/P(Y^c)]$이나 본 예제에서는 a가 취할 정수 값의 범위가 제한되어 있으므로 [-5.667, 5.667]의 값을 갖게 된다. 따라서 $NL_{pure}(X \Rightarrow Y^c)$는 특정 연관성 규칙에서 그 크기를 확인할 수 있으므로 행태적인 해석이 가능하며, $P(X \cap Y^c)$의 값이 증가함에 따라 $NL_{pure}(X \Rightarrow Y^c)$는 단조 증가하는 것을 알 수 있다.

표 3.3 음의 향상도와 음의 순수 향상도의 비교

a	b	c	d	NL	NL_{pure}
30	15	40	0	1.889	1.889
31	14	39	1	1.763	1.621
32	13	38	2	1.637	1.354
33	12	37	3	1.511	1.086
34	11	36	4	1.385	0.819
35	10	35	5	1.259	0.551
36	9	34	6	1.133	0.283
37	8	33	7	1.007	0.016
38	7	32	8	0.881	-0.252
39	6	31	9	0.756	-0.519
40	5	30	10	0.630	-0.787
41	4	29	11	0.504	-1.055
42	3	28	12	0.378	-1.322
43	2	27	13	0.252	-1.590
44	1	26	14	0.126	-1.857
45	0	25	15	0.000	-2.125

마지막으로 표 3.1로부터 발생빈도에 따른 음의 지지도와 음의 순수 지지도 $NS_{pure}(X \Rightarrow Y^c)$를 계산하면 다음 표 3.4와 같다.

표 3.4 음의 지지도와 음의 순수 지지도의 비교

a	b	c	d	NS	NS_{pure}
30	15	40	0	0.176	0.176
31	14	39	1	0.165	0.153
32	13	38	2	0.153	0.129
33	12	37	3	0.141	0.106
34	11	36	4	0.129	0.082
35	10	35	5	0.118	0.059
36	9	34	6	0.106	0.035
37	8	33	7	0.094	0.012
38	7	32	8	0.082	-0.012
39	6	31	9	0.071	-0.035
40	5	30	10	0.059	-0.059
41	4	29	11	0.047	-0.082
42	3	28	12	0.035	-0.106
43	2	27	13	0.024	-0.129
44	1	26	14	0.012	-0.153
45	0	25	15	0.000	-0.176

일반적으로 음의 지지도는 [0, 1]의 값을 취하는 반면에 음의 순수 지지도는 [-1. 1]의 값을 갖게 된다. 그러나 본 예제에서는 a가 취할 정수 값의 범위가 제한되어 있으므로 음의 지지도는 [0, 0.176]의 값을,

그리고 음의 순수 지지도는 [-0.176, 0.176]의 값을 갖는다. 또한 위의 표에서 a=39, b=6, c=31, 그리고 d=9이면 지지도는 0.071이 되는 반면에 순수 지지도 $NS_{pure}(X \Rightarrow Y^c)$는 -0.035의 값을 갖게 된다. 이들 값이 의미가 있다고 가정한다면 음의 지지도가 0.071라는 의미는 하나의 트랜잭션에 항목 X는 포함되고 항목 Y는 포함되지 않을 확률이 0.071라는 의미이고, 음의 순수 지지도의 값이 -0.035이라는 의미는 항목 X는 포함되지 않고 항목 Y만 포함될 확률에서 두 항목이 동시에 포함되지 않을 확률을 뺀 값이 -0.035라는 의미이다. 이러한 경우에는 동시에 두 항목이 포함되지 않을 확률보다 X는 포함되지 않고 Y만 포함될 확률이 더 작으므로 음의 연관성 규칙 생성을 위한 전제조건으로부터 배제하는 것이 바람직한 것으로 판단된다.

따라서 음의 연관성 규칙을 탐색하기 위해 음의 순수 연관성 측도를 이용하는 경우에는 첫 번째 단계에서 음의 순수 지지도의 값이 양의 특정한 값 이상이 되도록 하여 원하는 수준의 음의 순수 신뢰도의 값을 만족하는 연관성 규칙을 발견하는 것이 바람직하다.

4. 결론

음의 연관성 규칙 마이닝은 어느 항목이 발생하면 다른 항목도 발생하지 않는다는 규칙을 찾아내는 것이다. 본 논문에서는 특정 항목의 고유한 특성을 파악하기 위한 음의 순수 연관성 규칙을 제안하였다. 제안된 흥미도 측도인 음의 순수 신뢰도, 음의 순수 향상도, 그리고 음의 순수 지지도에 대해 여러 가지 성질을 조사하는 동시에 흥미도 측도의 조건에 대한 충족여부도 알아보았다. 그리고 예제 데이터를 이용하여 이들 세 가지 측도와 기존의 음의 연관성 규칙을 위한 흥미도 측도를 비교하였다.

기존의 음의 신뢰도는 양의 연관성을 가지는 연관성 규칙을 의미 있는 음의 관계를 가지는 규칙으로 선택하게 되는 오류를 범할 수 있다. 반면에 음의 순수 신뢰도는 순수하게 특정 요인에 의해서만 결과가 얼마인가를 나타내주는 측도인 동시에 그 부호에 의해 음의 관련성과 양의 관련성을 판단할 수 있다는 사실을 예제를 통하여 알게 되었다. 또한 음의 향상도는 모두 양의 값을 가지므로 방향이 없으며, 음의 향상도가 취할 수 있는 값의 범위를 알 수 없으므로 행태적인 해석을 하기가 어렵다. 그러나 음의 순수 향상도는 부호에 의해 연관성 규칙의 방향을 알 수 있는 동시에 항목집합의 선행여부에 따라 값이 다르므로 항목집합의 선행여부에 따라 값이 변한다는 사실을 확인할 수 있었다. 마지막으로 음의 순수 지지도는 흥미도 측도의 조건이 충족되지는 않으나, 규칙 생성을 위한 첫 번째 단계가 최소 지지도를 만족시키는 빈발항목집합 생성과정이므로 의미 있는 규칙 발견을 위해 필요한 측도인 것을 예제를 통하여 확인하였다.

일반적으로 음의 연관성 규칙의 평가기준을 적용하는 경우에는 구매건수가 증가함에 따라 연관성 규칙의 수가 기하급수적으로 증가하게 되나 음의 순수 연관성 규칙의 평가기준을 적용하게 되면 양의 값을 가지는 연관성 규칙만을 고려함으로써 의미 있는 연관성 규칙의 수가 많이 감소하게 된다. 따라서 항목간의 음의 연관성 측정을 위해 본 논문에서 제안한 음의 순수 연관성 규칙을 적용하게 되면 음의 연관성의 정도를 보다 정확하게 평가함으로써 올바른 의사결정에 도움을 줄 수 있을 것이다. 향후에는 본 연구의 결과를 실제 데이터에 적용해 봄으로써 어떤 특성을 가지는 데이터베이스에 적합한지에 대한 논의가 필요할 것으로 사료된다.

참고문헌

황준현, 김재련 (2003). 역 연관규칙을 이용한 타겟 마케팅. <한국지능정보시스템학회논문지>, **9**, 195-209.

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.

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

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. (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. 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.

Koh, Y. S. and Pears, R. (2007). Efficiently finding negative association rules without support threshold. *Advances in Artificial Intelligence, Lecture Notes in Computer Science*, **4830**, 710-714.

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 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.

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.

Sharma, S., Sharma, S. and Agrawal, J. (2007). GA optimized negative association rule mining. *International Journal of Soft Computing*, **2**, 124-128.

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.

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.

Yuan, X., Buckles, B. P., Yuan, Z. and Zhang, J. (2002). Mining negative association rules. *Proceedings of the Seventh International Symposium on Computers and Communications*, 623-628.

Journal of the Korean Data & Information Science Society
2011, **22**(2), 179–188

한국데이터정보과학회지

Proposition of negatively pure association rule threshold

Hee Chang Park[1]

[1]Department of Statistics, Changwon National University

Received 16 January 2011, revised 11 February 2011, accepted 14 February 2011

Abstract

Association rule represents the relationship between items in a massive database by quantifying their relationship, and is used most frequently in data mining techniques. 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 pure association rules by negatively pure support, negatively pure confidence, and negatively pure lift to overcome the problems faced by negative association rule technique. In checking the usefulness of this technique through numerical examples, we could find the direction of association by the sign of the negatively pure association rule measure.

Keywords: Interestingness measure, negatively pure association rule, negatively pure confidence, negatively pure lift, negatively pure support.

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