Journal of the Korean Data & Information Science Society
2013, **24**(4), 835–845

http://dx.doi.org/10.7465/jkdi.2013.24.4.835
한국데이터정보과학회지

기준 확인 측도와 연관성 평가기준과의 관계 탐색

박희창[1]

[1]창원대학교 통계학과

접수 2013년 6월 18일, 수정 2013년 7월 8일, 게재확정 2013년 7월 13일

요 약

데이터 마이닝닝 기법들 중에서 연관성 규칙 마이닝 (association rule mining)은 대용량의 사건 발생 기록 데이터로부터 항목 간의 연관성을 측정하는 기법이다. 이 기법은 매우 방대한 양의 상품 또는 서비스 거래 기록 데이터로부터 항목들 간의 연관성을 측정하는 기법으로 제조업, 유통업, 보험업, 의료 및 교육 분야 등 많은 분야에 적용되고 있다. 의미 있는 연관성 규칙을 탐색하기 위한 흥미도 측도는 크게 객관적 흥미도 측도와 주관적 흥미도 측도, 그리고 의미론적 흥미도 측도로 분류할 수 있다. 이와는 별개로 기준 확인 또는 증거 지원과 관련된 측도들을 개발하기 위해 많은 시도가 있었으나 기준 확인 측도에 대한 연관성 평가 기준 조건 충족 여부나 기본적인 연관성 평가 측도인 지지도, 신뢰도, 그리고 향상도 등과의 관계는 아직 규명되지 않았다. 이에 본 논문에서는 가장 많이 활용되고 있는 비대칭적 기준 확인 측도에 대해 흥미도 측도의 기준에 대한 조건 충족 여부를 검토하는 동시에 기본적인 연관성 평가 측도들과의 관계를 수식을 통해 유도한 후, 예제를 통해 연관성 규칙의 관점에서 기준 확인 측도의 유용성을 살펴보았다. 그 결과, 본 논문에서 고려한 모든 기준 확인 측도들이 흥미도 측도의 기준에 대한 조건들을 모두 만족하였다. 또한 이들을 기본적인 연관성 평가 기준인 지지도, 신뢰도, 그리고 향상도와의 관계를 식을 통해 규명한 동시에 방향성과 행태적 해석 가능성을 예제를 통해 확인할 수 있었다. 특히 이들 측도 중에서 Kemeny와 Oppenheim이 제안한 측도와 Rips가 제안한 측도가 가장 바람직한 연관성 평가 기준으로 활용할 수 있다는 사실을 확인할 수 있었다.

주요용어: 기준 확인 측도, 신뢰도, 연관성 규칙, 지지도, 향상도, 흥미도 측도.

1. 서론

연관성 규칙 마이닝 (association rule mining)은 데이터 마이닝 기법들 중에서 많이 연구되고 있는 분야로, 매우 방대한 양의 상품 또는 서비스 거래 기록 데이터로부터 항목들 간의 연관성을 측정하는 기법으로 유통업, 제조업, 보험업, 의료 및 교육 분야 등 많은 분야에 적용되고 있다 (Park, 2012b). 이러한 연관성 규칙 마이닝 기법에서 측정의 기준이 되는 것은 어떤 상품들이 얼마나 자주 함께 구매되었는가 하는 빈도 (frequency)이다. 이 기법은 동시에 발생하는 여러 항목들을 생성된 규칙의 집합으로 나타냄으로써 항목들 간의 상호 연관성들을 쉽게 파악할 수 있는 정성적인 의미로 해석할 수 있으며, 연관성 평가 측도를 이용함으로써 정량적으로도 분석이 가능하다 (Srikant와 Agrawal, 1995). Agrawal 등 (1993)에 의해 처음으로 소개된 연관성 규칙 기법은 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

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

대표 연구로는 Han과 Fu (1999), Liu 등 (1999)이 있다. 또한 연관성 규칙에 대한 최근 국내 연구로는 Lim 등 (2010), Choi와 Park (2011), Park (2011a, 2011b, 2011c, 2012a, 2012b) 등이 있다. 이러한 연관성 규칙은 항목집합들 간의 지지도 (support), 신뢰도 (confidence), 향상도 (lift) 등의 흥미도 측도 (interestingness measure)를 바탕으로 관련성 여부를 측정한다. 일반적으로 연관성 규칙을 생성할 때 우선적으로 사용자가 지정한 최소 지지도의 조건을 만족하는 빈발항목집합을 생성한다. 그런 후, 향상도가 1이상인 것 중에서 최저 신뢰도의 조건을 만족하는 규칙을 연관성 규칙으로 채택하게 된다 (Park, 2011a).

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. 여기에 Geng과 Hamilton (2006)은 의미론적 흥미도 측도 (semantic measures of interestingness)의 개념을 추가하였다. 이러한 흥미도 측도에 관해서는 많은 학자들에 의해 연구가 수행되었으며, 대표적인 연구로는 Hilderman과 Hamilton(2000)이 객관적 흥미도 측도들을 데이터마이닝에 응용하였으며, Liu 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.

다른 관점에서 과학자들은 기준 확인 또는 증거 지원과 관련된 측도들을 개발하기 위해 많은 시도가 있었다. 기준 확인 측도 (confirmation measure)는 확률이론에 바탕을 두고 있으며, 임의의 한 증거와 가설이 확률적으로 양의 종속이면 그 증거가 가설을 확인 또는 지원해주는 정도가 양의 값을 가지며, 그들이 음으로 종속적이면 그 정도는 음의 값을 가진다. 반면에 그들이 확률적으로 독립이면 그 값은 0이 된다. 흥미도 측도의 관점에서 기준 확인 측도의 바람직한 성질들에 대한 연구가 진행되어 왔으며, 실제적으로 기준 확인 측도와 흥미도 측도는 많이 관련되어 있는 것으로 나타났다. 그 이유는 많은 흥미도 측도들이 기준 확인 측도가 될 수 있으며, 확률적 독립/종속이 연관성 규칙의 관점에서 논의되고 있기 때문이다 (Glass, 2013). 그러나 기준 확인 측도에 대한 활용이나 측도 개발에 대한 연구는 많이 수행되고 있으나 기준 확인 측도에 대한 연관성 평가 기준에 대한 만족 여부나 기본적인 연관성 평가 측도인 지지도, 신뢰도, 그리고 향상도 등과의 관계는 아직 규명되지 않고 있다. 이에 본 논문에서는 가장 많이 활용되고 있는 비대칭적 기준 확인 측도에 대해 Piatetsky-Shapiro (1991)가 제안한 흥미도 측도의 기준에 대한 조건 충족 여부를 검토하는 동시에 기본적인 연관성 평가 측도들과의 관계를 수식을 통해 유도한 후, 예제를 통해 연관성 규칙의 관점에서 기준 확인 측도의 유용성을 살펴보고자 한다.

2. 기본적인 비대칭 기준 확인 측도

본 논문에서는 가장 많이 활용되고 있는 기본적인 기준 확인 측도들에 대해 연관성 평가 기준으로서의 적용가능성을 평가하고자 한다. Glass (2013)와 Crupi (2007)이 기술한 바와 같이 가정을 H로 하고, 증거 또는 결과를 E로 하였을 때 기본적인 비대칭적 기준 확인 측도에는 Eells (1982)와 Jeffrey (1992)이 제안한 $EJ(E,H)$, Mortimer (1988)의 $MO(E,H)$, Nozick (1981)의 $NO(E,H)$, Christensen (1999)의 $CH(E,H)$, Good (1984)의 $GO(E,H)$, Kemeny and Oppenheim (1952)의 $KO(E,H)$, 그

리고 Rips (2001)의 $RI(E,H)$등이 있다.

$$EJ(E,H) = P(H|E) - P(H)$$
$$MO(E,H) = P(E|H) - P(E)$$
$$NO(E,H) = P(E|H) - P(E|H^c)$$
$$CH(E,H) = P(H|E) - P(H|E^c$$
$$RI(E,H) = 1 - \frac{P(H^c|E)}{P(H^c)}$$
$$GO(E,H) = \log\left[\frac{P(E|H)}{P(E|H^c)}\right]$$
$$KO(E,H) = \frac{P(E|H) - P(E|H^c)}{P(E|H) + P(E|H^c)}$$

본 논문에서는 연관성 평가 기준으로서의 기준 확인 측도들을 고려하므로 $H \rightarrow E$의 형태로 재정리하면 $EJ(E,H)$와 $CH(E,H)$는 각각 $MO(E,H)$및 $NO(E,H)$과 같아지므로 고려할 필요가 없어지며, $RI(E,H)$는 다음과 같이 변형된 것을 연관성 평가 기준으로 탐색하고자 한다,

$$RI(E,H) = 1 - \frac{P(E^c|H)}{P(E^c)}$$

3. 비대칭 기준 확인 측도와 연관성 평가 측도와의 관계 규명

연관성 규칙을 평가하는 기준에는 지지도, 신뢰도, 향상도 등이 있다. 지지도 $S(A,B)$는 항목 집합 A와 B가 동시에 발생하는 거래의 비율로 $P(A \text{ and } B)$으로 계산된다. 신뢰도 $C(A \Rightarrow B)$는 항목 집합 A가 포함된 거래 비율 중 항목 집합 A와 항목 집합 B가 동시에 포함된 거래의 비율을 의미하며 $P(B\ |A)$으로 계산된다. 그리고 향상도 $L(A,B)$는 항목 집합 A를 구매한 경우 그 거래가 항목 집합 B를 포함하는 경우와 항목 집합 B가 임의로 구매되는 경우의 비를 의미하며, $P(B\ |A)/P(B)$로 계산된다.

이 절에서는 위에서 논의한 수식에서 H를 전항집합 A로 하고 E를 후항집합 B로 하여 기본적인 연관성 평가 기준인 지지도, 신뢰도, 그리고 향상도와의 관계식을 나타내면 다음과 같다.

$$MO(A,B) = \begin{cases} \dfrac{1}{P(A)}[S(A.B) - P(A)P(B)] \\ C(A \Rightarrow B) - P(B) \\ P(B)[L(A,B) - 1] \end{cases}$$

위의 식에서 보는 바와 같이 측도 $MO(A,B)$는 전항의 발생 확률 $P(A)$에 대해 지지도와 A와 B가 독립일 때 확률값의 차이를 나타내고 있는 동시에, 신뢰도와 후항 발생 확률 $P(B)$의 차이를 나타내고 있다. 또한 향상도와는 두 항목이 독립이면 향상도가 1이므로 향상도와 1의 차이를 $P(B)$의 크기에 따

라 나타내주는 측도가 된다.

$$NO(A,B) = \begin{cases} \dfrac{S(A,B) - P(A)P(B)}{P(A)[1-P(A)]} \\ \dfrac{C(A \Rightarrow B) - P(B)}{1-P(A)} \\ \dfrac{P(B)[L(A,B)-1]}{1-P(A)} \end{cases}$$

측도 $NO(A,B)$는 $P(A)P(A^c)$에 대해 지지도와 A와 B가 독립일 때의 확률 차이를 나타내고 있는 동시에, 전항의 비발생 확률 $P(A^c)$에 대해 신뢰도와 후항 발생 확률 $P(B)$의 차이를 나타내고 있다. 또한 향상도와는 두 항목이 독립이면 향상도가 1이므로 향상도와 1의 차이를 $P(B)$와 $P(A^c)$의 비에 따라 나타내주는 측도가 된다. 따라서 $NO(A,B)$는 $MO(A,B)$에 비해 분자에 $P(A^c)$를 더 고려한 측도로 생각할 수 있다.

$$RI(A,B) = \begin{cases} \dfrac{S(A,B) - P(A)P(B)}{P(A)[1-P(B)]} \\ \dfrac{C(A \Rightarrow B) - P(B)}{1-P(B)} \\ \dfrac{P(B)[L(A,B)-1]}{1-P(B)} \end{cases}$$

측도 $RI(A,B)$는 $P(A)P(B^c)$에 대해 지지도와 A와 B가 독립일 때의 확률 차이를 나타내고 있는 동시에, 후항의 비발생 확률 $P(B^c)$에 대해 신뢰도와 후항 발생 확률 $P(B)$의 차이를 나타내고 있다. 또한 향상도와는 두 항목이 독립이면 향상도가 1이므로 향상도와 1의 차이를 $P(B)$와 $P(B^c)$의 비에 따라 나타내주는 측도가 된다. 따라서 $RI(A,B)$는 $NO(A,B)$에서 $P(A^c)$대신 $P(B^c)$를 고려한 측도로 생각할 수 있다.

$$KO(A,B) = \begin{cases} \dfrac{S(A,B) - P(A)P(B)}{S(A,B)[1-2P(A)] + P(A)P(B)} \\ \dfrac{C(A \Rightarrow B) - P(B)}{C(A \Rightarrow B)[1-2P(A)] + P(B)} \\ \dfrac{L(A,B) - 1}{L(A,B)[1-2P(A)] + 1} \end{cases}$$

측도 $KO(A,B)$는 위에서 기술한 측도들과 같이 지지도와 A와 B가 독립일 때의 확률 차이, 신뢰도와 $P(B)$의 차이, 그리고 향상도와 1의 차이의 크기를 나타내주는 측도이나 분모의 값이 식에서 보는 바

와 같이 위의 측도들에 비해 좀 더 복잡한 수식으로 나타나고 있다.

$$GO(A,B) = \begin{cases} \log[\dfrac{S(A,B) - P(A)S(A,B)}{P(A)P(B) - P(A)S(A,B)}] \\ \log[\dfrac{C(A \Rightarrow B) - P(A)C(A \Rightarrow B)}{P(B) - P(A)C(A \Rightarrow B)} \\ \log[\dfrac{L(A,B) - P(A)L(A,B)}{1 - P(A)L(A,B)}] \end{cases}$$

측도 $GO(A,B)$는 $NO(A,B)$를 구성하고 있는 각 항의 차이를 비로 나타낸 후 로그함수를 적용한 것으로 지지도와 $P(A)P(B)$의 차이, 신뢰도와 $P(B)$의 차이, 그리고 향상도와 1의 차이의 크기에 따라 그 값이 결정된다.

4. 비대칭 기준 확인 측도의 연관성 평가 기준 조건 탐색

본 논문에서 제안한 비대칭 기준 확인 측도들이 Piatetsky-Shapiro (1991)가 제안한 흥미도 측도의 조건을 충족하는지의 여부를 조사하면 다음과 같다.

[조건 1] $P(A \text{ and } B) = P(A)P(B)$이면 비대칭 기준 확인 측도는 0이 된다.

(증명) 위의 식들에서 첫 번째 수식으로부터 모두 0이 되는 것을 알 수 있다. 특히 측도 $GO(E,H)$는 첫 번째 식에서 $P(A \text{ and } B) = P(A)P(B)$이면 로그 안의 값이 1이 되므로 이 측도는 0이 된다.

[조건 2] 비대칭 기준 확인 측도는 $P(A)$와 $P(B)$가 고정되어 있을 때, $P(A \text{ and } B)$의 값에 따라 단조 증가한다.

(증명) $P(A)$와 $P(B)$가 고정되어 있을 때, $conf(A \Rightarrow B) = P(A \text{ and } B)/P(A)$이므로 $P(A \text{ and } B)$가 증가한다는 것은 $conf(A \Rightarrow B)$가 증가한다는 의미이므로 위의 측도들의 두번째 식으로부터 $P(A \text{ and } B)$의 값이 증가함에 따라 비대칭 기준 확인 측도들이 증가한다는 것을 쉽게 알 수 있다. 특히 $KO(A,B)$는 두 번째 식을 $conf(A \Rightarrow B)$에 대해 편미분하면 다음과 같이 나타나서 0보다 큰 값이 되므로 이 조건이 만족함을 알 수 있다.

$$\frac{\partial KO(A,B)}{\partial C(A \Rightarrow B)} = \frac{2P(B)[1 - P(A)]}{[C(A \Rightarrow B)(1 - 2P(A)) + P(B)]^2}$$

또한 $GO(A,B)$도 두 번째 식의 로그 안의 수식을 $conf(A \Rightarrow B)$에 대해 편미분하면 다음과 같이 나타나서 0보다 큰 값이 되므로 이 조건이 만족함을 알 수 있다.

$$\frac{\partial GO(A,B)}{\partial C(A \Rightarrow B)} = \frac{[1 - P(A)][1 + P(A)C(A \Rightarrow B)]}{[P(B) - P(A)C(A \Rightarrow B)]^2}$$

[조건 3] 비대칭 기준 확인 측도는 $P(B)$의 값에 따라 단조 감소한다.

(증명) 측도 $MO(A,B)$와 $NO(A,B)$, 그리고 $GO(A,B)$는 첫 번째 수식 또는 두 번째 수식에 의해 $P(B)$가 증가하면 감소하고 있는 것을 쉽게 알 수 있다. 측도 $RI(A,B)$는 두 번째 식을 $P(B)$에 대해 편미분하면 다음과 같이 음의 값을 가지게 되므로 $P(B)$가 증가하면 감소하고 있는 것을 알 수 있다.

$$\frac{\partial RI(A,B)}{\partial P(B)} = -\frac{1 - C(A \Rightarrow B)}{[1 - P(B)]^2}$$

또한 $KO(A,B)$도 첫 번째 식을 $P(B)$에 대해 편미분하면 다음과 같이 나타나서 0보다 작은 값이 되므로 이 조건이 만족함을 알 수 있다.

$$\frac{\partial KO(A,B)}{\partial P(B)} = -\frac{P(A)[1+S(A,B)-P(A)P(B)]}{[S(A,B)(1-2P(A))+P(A)P(B)]^2}$$

5. 모의실험

본 절에서는 신뢰도와 기준 확인 측도 간의 관계와 기준 확인 측도의 유용성을 예제를 통해 고찰하고자 하며, 이를 위해 항목 A, B에 대해 Park (2012a)에서와 동일한 예제를 적용하였다. 이를 구체적으로 기술하면 먼저 데이터베이스에 있는 총 트랜잭션의 수 (t)를 100명으로 하고, 항목 A는 구매한 냉장고의 금액을 기준으로 100만원 이상 (1) 구매한 사람 수를 50명으로 하고 100만원 미만 (0)을 구매한 사람 수를 50명으로 하였다. 또한 항목 B를 결제 방식을 기준으로 신용 카드로 결제 (1)한 사람 수를 30명으로 하고 신용 카드 이외의 방법으로 결제 (0)한 사람의 수를 70명으로 하였다. 항목 A와 B가 동시에 발생한 빈도 수, 즉 100만원 이상의 냉장고를 구매하면서 신용카드로 결제한 빈도수는 a명으로 하였다. 이를 정리하면 Table 5.1과 같다. 이 표에서 a가 취할 수 있는 범위는 $0 \le a \le 30$이다. Table 5.1로부터 동시발생빈도 (a)에 따른 기준 확인 측도들을 계산하면 다음의 Table 5.2와 같은 결과를 얻을 수 있다. 이에 대한 계산은 미니탭 16의 계산기 기능을 이용하였다. 이 표에서 $b = P(A=1, B=0)$, $c = P(A=0, B=1)$, $d = P(A=0, B=0)$이며, $supp = P(A \text{ and } B)$이고 $conf = P(B|A)$을 의미한다.

Table 5.1 Simulation data(1)

		B		Total
		1	0	
A	1	a	$50-a$	50
	0	$30-a$	$a+20$	50
Total		30	70	100

Table 5.2 Output of confirmation measures by simulation data(1)

a	b	c	d	$supp$	$conf$	$lift$	MO	GO	NO	KO	RI
1	49	29	21	0.010	0.020	0.040	-0.280	-3.367	-0.560	-0.933	-0.400
2	48	28	22	0.020	0.040	0.080	-0.260	-2.639	-0.520	-0.867	-0.371
3	47	27	23	0.030	0.060	0.120	-0.240	-2.197	-0.480	-0.800	-0.343
4	46	26	24	0.040	0.080	0.160	-0.220	-1.872	-0.440	-0.733	-0.314
5	45	25	25	0.050	0.100	0.200	-0.200	-1.609	-0.400	-0.667	-0.286
6	44	24	26	0.060	0.120	0.240	-0.180	-1.386	-0.360	-0.600	-0.257
7	43	23	27	0.070	0.140	0.280	-0.160	-1.190	-0.320	-0.533	-0.229
8	42	22	28	0.080	0.160	0.320	-0.140	-1.012	-0.280	-0.467	-0.200
9	41	21	29	0.090	0.180	0.360	-0.120	-0.847	-0.240	-0.400	-0.171
10	40	20	30	0.100	0.200	0.400	-0.100	-0.693	-0.200	-0.333	-0.143
11	39	19	31	0.110	0.220	0.440	-0.080	-0.547	-0.160	-0.267	-0.114
12	38	18	32	0.120	0.240	0.480	-0.060	-0.405	-0.120	-0.200	-0.086
13	37	17	33	0.130	0.260	0.520	-0.040	-0.268	-0.080	-0.133	-0.057
14	36	16	34	0.140	0.280	0.560	-0.020	-0.134	-0.040	-0.067	-0.029
15	35	15	35	0.150	0.300	0.600	0.000	0.000	0.000	0.000	0.000
16	34	14	36	0.160	0.320	0.640	0.020	0.134	0.040	0.067	0.029
17	33	13	37	0.170	0.340	0.680	0.040	0.268	0.080	0.133	0.057
18	32	12	38	0.180	0.360	0.720	0.060	0.405	0.120	0.200	0.086
19	31	11	39	0.190	0.380	0.760	0.080	0.547	0.160	0.267	0.114
20	30	10	40	0.200	0.400	0.800	0.100	0.693	0.200	0.333	0.143

이 표에서 보는 바와 같이 신뢰도 $conf$의 값이 증가할수록 본 논문에서 고려하는 모든 측도들이 증가하는 양상을 보이고 있다. 또한 모든 측도들이 음의 값과 양의 값을 가지는 측도이므로 방향성을 나타내고 있다. 그러나 측도 $GO(E,H)$는 비록 연관성 평가 기준의 조건을 충족한다고 할지라도 절대값의 크기가 1을 초과하므로 행태적 해석이 곤란하여 연관성 평가 측도로서는 바람직하지 않다고 할 수 있다. 나머지 측도들 중에서는 각 케이스별 값의 차이가 $KO(E,H)$, $NO(E,H)$, $RI(E,H)$, $MO(E,H)$의 순서로 나타나고 있어서 $KO(E,H)$가 연관성 평가 기준으로서 가장 바람직한 측도라고 할 수 있다. 이들을 좀 더 구체적으로 살펴보기 위해 a, b, c, d의 각 값이 2, 48, 28, 22인 경우와 17, 33, 13, 37인 경우를 비교해보면 전자의 경우에는 모든 측도들이 음의 값을 갖는 반면에 후자의 경우에는 양의 값을 갖는다. 그리고 $GO(E,H)$는 절대값의 크기가 1을 초과하는 경우가 발생하였기 때문에 이들 두 측도는 연관성 측도로서는 바람직하지 않다고 할 수 있다. 나머지 측도 중에서는 각 케이스별로 $KO(E,H)$, $NO(E,H)$, $RI(E,H)$, $MO(E,H)$의 순서로 각각 0.100, 0.600, 0.428, 0.300의 크기 차이를 나타내고 있다. 따라서 측도 $KO(E,H)$가 방향성을 나타냄과 동시에 값의 범위도 [-1, 1] 사이의 값을 가지며, 다른 측도들에 비해 케이별 차이를 가장 크게 나타내고 있으므로 가장 바람직한 연관성 측도라고 할 수 있다.

이번에는 두 항목간의 불일치빈도 b의 값의 변화에 따라 신뢰도와 기준 확인 측도들간의 관계와 이들 측도의 유용성을 고찰하고자 한다.

Table 5.3 Simulation data(2)

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

이를 위해 Table 5.3과 같이 각 셀의 값을 바꾸어 실험하였다. Table 5.3에서 b가 취할 수 있는 정수값의 범위는 $0 \le b \le 30$이다. 이 표로부터 각 셀 값의 변화에 따른 신뢰도와 기준 확인 측도들을 계산하여 그 일부를 나타내면 다음 Table 5.4와 같다.

Table 5.4 Output of confirmation measures by simulation data(2)

a	b	c	d	$supp$	$conf$	$lift$	MO	GO	NO	KO	RI
25	5	25	45	0.250	0.833	2.778	0.333	0.847	0.476	0.400	0.667
24	6	26	44	0.240	0.800	2.667	0.300	0.767	0.429	0.366	0.600
23	7	27	43	0.230	0.767	2.556	0.267	0.687	0.381	0.331	0.533
22	8	28	42	0.220	0.733	2.444	0.233	0.606	0.333	0.294	0.467
21	9	29	41	0.210	0.700	2.333	0.200	0.525	0.286	0.256	0.400
20	10	30	40	0.200	0.667	2.222	0.167	0.442	0.238	0.217	0.333
19	11	31	39	0.190	0.633	2.111	0.133	0.358	0.190	0.177	0.267
18	12	32	38	0.180	0.600	2.000	0.100	0.272	0.143	0.135	0.200
17	13	33	37	0.170	0.567	1.889	0.067	0.184	0.095	0.092	0.133
16	14	34	36	0.160	0.533	1.778	0.033	0.094	0.048	0.047	0.067
15	15	35	35	0.150	0.500	1.667	0.000	0.000	0.000	0.000	0.000
14	16	36	34	0.140	0.467	1.556	-0.033	-0.097	-0.048	-0.049	-0.067
13	17	37	33	0.130	0.433	1.444	-0.067	-0.199	-0.095	-0.099	-0.133
12	18	38	32	0.120	0.400	1.333	-0.100	-0.305	-0.143	-0.152	-0.200
11	19	39	31	0.110	0.367	1.222	-0.133	-0.418	-0.190	-0.206	-0.267
10	20	40	30	0.100	0.333	1.111	-0.167	-0.539	-0.238	-0.263	-0.333
9	21	41	29	0.090	0.300	1.000	-0.200	-0.669	-0.286	-0.323	-0.400

Table 5.4에서 보는 바와 같이 신뢰도 $conf$의 값이 감소할수록 본 논문에서 고려하는 모든 측도들이 감소하는 양상을 보이고 있으며, 이들 모두 음의 값과 양의 값을 가지는 측도이므로 방향성을 나타내고 있다. 여기서도 측도 $GO(E,H)$는 절대값의 크기가 1을 초과하므로 행태적 해석이 곤란한 것으로 나타났다. 나머지 측도들 중에서는 각 케이스별 값의 차이가 $RI(E,H)$, $NO(E,H)$, $KO(E,H)$, $MO(E,H)$의 순서로 나타나고 있어서 $RI(E,H)$가 연관성 평가 기준으로서 가장 바람직한 측도라고 할 수 있다. 이들을 좀 더 구체적으로 살펴보기 위해 a, b, c, d의 각 값이 18, 12, 32, 38인 경우와 12, 18, 38, 32인 경우를 비교해보면 전자의 경우에는 모든 측도들이 양의 값을 갖는 반면에 후자의 경우에는 음의 값을 가지므로 방향성을 갖는다. 그리고 측도 $GO(E,H)$는 절대값의 크기가 1을 초과하는 경우가 발생하였기 때문에 이는 연관성 측도로서는 바람직하지 않다고 할 수 있다. 나머지 측도 중에서는 각 케이스별로 $RI(E,H)$, $KO(E,H)$, $NO(E,H)$, $MO(E,H)$의 순으로 각각 0.067, 0.048, 0.047, 0.033의 크기 차이를 나타내고 있다. 따라서 측도 $RI(E,H)$가 방향성을 나타냄과 동시에 값의 범위도 [-1, 1] 사이의 값을 가지며, 다른 측도들에 비해 케이별 차이를 가장 크게 나타내고 있으므로 가장 바람직한 연관성 측도라고 할 수 있다.

이러한 결과를 종합해볼 때, 측도 $KO(E,H)$와 $RI(E,H)$가 방향성을 나타냄과 동시에 절대값의 크기도 1을 초과하지 않으므로 연관성 평가 기준으로 바람직한 측도라고 할 수 있다.

6. 결론

연관성 규칙 마이닝은 대용량의 사건 발생 기록 데이터로부터 항목 간의 연관성을 측정하는 기법이다. 지지도, 신뢰도, 향상도 이외에도 의미 있는 연관성 규칙을 탐색하기 위한 흥미도 측도는 객관적 흥미도 측도와 주관적 흥미도 측도, 그리고 의미론적 흥미도 측도로 크게 분류할 수 있다. 이와는 별도로 기준 확인 또는 증거 지원과 관련된 측도들을 개발하기 위해 많은 시도가 있었으나, 기준 확인 측도에 대한 연관성 평가 기준 조건 충족 여부나 기본적인 연관성 평가 측도인 지지도, 신뢰도, 그리고 향상도 등과의 관계는 아직 규명되지 않았다.

본 논문에서는 가장 많이 활용되고 있는 비대칭적 기준 확인 측도에 대해 흥미도 측도의 기준에 대한 조건 충족 여부를 검토하는 동시에 기본적인 연관성 평가 측도들과의 관계를 수식을 통해 유도하였으며, 예제를 통해 연관성 규칙의 관점에서 기준 확인 측도의 유용성을 살펴보았다. 그 결과, 본 논문에서 고려한 모든 기준 확인 측도들이 흥미도 측도의 기준에 대한 조건들을 모두 만족하였다. 또한 측도 $MO(A,B)$는 지지도와 두 항목이 독립일 때의 확률값의 차이를 전항 발생 확률에 대해 나타내고 있는 동시에, 신뢰도와 후항 발생 확률의 차이를 나타내고 있었으며, 향상도와 1의 차이를 후항 발생 확률의 크기에 따라 나타내주는 측도가 되었다. 측도 $NO(A,B)$는 전항 발생 확률과 전항 비발생 확률에 대해 지지도와 두 항목이 독립일 때의 확률 차이를 나타내고 있는 동시에, 전항의 비발생 확률에 대해 신뢰도와 후항 발생 확률의 차이를 나타내고 있으며, 향상도와는 두 항목이 독립이면 지지도가 1이므로 지지도와 1의 차이를 $P(B)$와 $P(A^c)$의 비에 따라 나타내주는 측도가 된다. 따라서 $NO(A,B)$는 $MO(A,B)$에 비해 분자에 전항 비발생 확률을 추가로 고려한 측도로 생각할 수 있다. 측도 $RI(A,B)$는 전항 발생 확률과 후항 비발생 확률에 대해 지지도와 두 항목이 독립일 때의 확률 차이를 나타내고 있는 동시에, 후항의 비발생 확률에 대해 신뢰도와 후항 발생 확률의 차이를 나타내고 있으며, 향상도와 1의 차이를 후항 발생 확률과 후항 비발생 확률의 비에 따라 나타내는 측도가 된다. 따라서 측도 $RI(A,B)$는 $NO(A,B)$에서 전항 비발생 확률 대신 후항 비발생 확률을 고려한 측도로 생각할 수 있다. 측도 $KO(A,B)$는 위에서 기술한 측도들과 같이 지지도와 두 항목이 독립일 때의 확률 차이, 신뢰도와 후항 발생 확률의 차이, 그리고 향상도와 1의 차이의 크기를 나타내주는 측도이나 분모의 값이 다른 측도들에 비해 좀 더 복잡한 수식으로 나타났다. 마지막으로 측도 $GO(A,B)$는 $NO(A,B)$를 구성

하고 있는 각 항의 차이를 비로 나타낸 후, 로그함수를 적용한 것으로 지지도와 전항 발생 확률 및 후항 발생 확률의 차이, 신뢰도와 후항 발생 확률의 차이, 그리고 향상도와 1의 차이의 크기에 따라 그 값이 결정되는 사실을 발견할 수 있었다.

또한 예제를 통해 살펴본 결과, 본 논문에서 고려하는 모든 측도들은 신뢰도의 값이 증가할수록 증가하는 양상을 보였으며, 신뢰도의 값이 감소하면 모든 측도들은 감소하는 양상을 보였다. 또한 모든 측도들이 음의 값과 양의 값을 가지는 측도이므로 방향성을 나타내고 있는 동시에, 측도 $GO(E,H)$를 제외한 모든 측도들은 절대값의 크기가 1을 초과하지 않으므로 행태적 해석이 가능하여 연관성 평가 측도로 바람직하다고 할 수 있다. 이러한 결과를 종합해볼 때, 측도 $KO(E,H)$와 $RI(E,H)$가 방향성을 나타냄과 동시에 절대값의 크기가 1을 초과하지 않으므로 연관성 평가 기준으로 가장 바람직한 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.

Cho, K. H. and Park, H. C. (2011). Discovery of insignificant association rules using external variable. *Journal of the Korean Data Analysis Society*, **13**, 1343-1352.

Crupi, V., Tentori, K. and Gonzalez, M. (2007). On Bayesian measures of evidential support: Theoretical and empirical issues. *Philosophy of Science*, 74, 229-252.

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

Geng, L. and Hamilton, H. J. (2006). Interestingness measures for data mining: A survey. *ACM Computing Surveys*, **38**, 1-32.

Glass, D. H. (2013). Confirmation measures of association rule interestingness. *Knowledge-Based Systems*, 44, 65-77.

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.

Kemeny, J. G. and Oppenheim, P. (1952). Degree of factual support. *Philosophy of Science*, **19**, 307-324.

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.

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

Mortimer, H. (1988), *The logic of induction*, Prentice Hall, Paramus.

Nozick, R. (1981), *Philosophical explanations*, Clarendon Press, Oxford.

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. (2012a). Negatively attributable and pure confidence for generation of negative association rules. *Journal of the Korean Data & Information Science Society*, **14**, 707-716.

Park, H. C. (2012b). Exploration of PIM based similarity measures as association rule thresholds. *Journal of the Korean Data & Information Science Society*, **23**, 1127-1135.

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. *Proceedings of the 9th National Conference on Artificial Intelligence: Knowledge Discovery in Databases*, 229-248.

Rips, L. J. (2001). Two kinds of reasoning. *Psychological Science*, 12, 129-134.

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.

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

Srikant, R. and Agrawal, R. (1995). Mining generalized association rules. *Proceedings of the 21st VLDB Conference*, 407-419.

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
2013, **24**(4), 835–845

http://dx.doi.org/10.7465/jkdi.2013.24.4.835
한국데이터정보과학회지

Exploration of relationship between confirmation measures and association thresholds

Hee Chang Park[1]

[1]Department of Statistics, Changwon National University

Received 18 June 2013, revised 8 July 2013, accepted 13 July 2013

Abstract

Association rule of data mining techniques is the method to quantify the relevance between a set of items in a big database, andhas been applied in various fields like manufacturing industry, shopping mall, healthcare, insurance, and education. Philosophers of science have proposed interestingness measures for various kinds of patterns, analyzed their theoretical properties, evaluated them empirically, and suggested strategies to select appropriate measures for particular domains and requirements. Such interestingness measures are divided into objective, subjective, and semantic measures. Objective measures are based on data used in the discovery process and are typically motivated by statistical considerations. Subjective measures take into account not only the data but also the knowledge and interests of users who examine the pattern, while semantic measures additionally take into account utility and actionability.In a very different context, researchers have devoted a lot of attention to measures of confirmation or evidential support. The focus in this paper was on asymmetric confirmation measures, and we compared confirmation measures with basic association thresholds using some simulation data. As the result, we could distinguish the direction of association rule by confirmation measures, and interpret degree of association operationally by them. Futhermore, the result showed that the measure by Rips and that by Kemeny and Oppenheim were better than other confirmation measures.

Keywords: Association rule, confidence, confirmation measure, interestingness measure, lift, support.

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