Journal of the Korean Data & Information Science Society
2014, **25**(1), 97–106

http://dx.doi.org/10.7465/jkdi.2014.25.1.97
한국데이터정보과학회지

연관 규칙 마이닝에서의 코사인 순수 신뢰도의 제안

박희창[1]

[1]창원대학교 통계학과

접수 2013년 12월 11일, 수정 2013년 12월 26일, 게재확정 2014년 1월 2일

요 약

빅 데이터 기술의 발전은 다변화된 현대 사회를 보다 정확하게 예측하고 효율적으로 작동하도록 정보를 제공하는 동시에 과거에는 불가능 했던 기술을 가능케 하였다. 이러한 빅 데이터 분석 기법은 국가 차원에서의 사회, 경제, 정치, 문화, 과학 기술 등 여러 분야에 활용될 수 있다. 빅 데이터 분석을 위해서는 먼저 데이터 마이닝 기술로 방대한 양의 데이터 속에서 가치 있는 정보를 찾는 것이 선행되어야 하는데, 빅 데이터와 관련된 데이터 마이닝 기법으로는 텍스트 마이닝, 평판 분석, 군집 분석, 연관성 규칙 등이 있다. 본 논문에서는 데이터 마이닝 기법 중에서 많이 활용되고 있는 연관성 규칙의 평가 기준으로 코사인 순수 신뢰도를 제안한 후, Piatetsky-Shapiro가 제안한 흥미도 측도의 기준에 대한 충족여부를 점검하는 동시에 여러 가지 특성을 살펴보았다. 또한 예제를 통하여 고찰한 결과, 기존의 신뢰도와 코사인 유사성 측도는 모두 양의 값을 가지므로 연관성의 방향을 알 수 없어서 그 값만으로는 양의 연관성이 있는지 아니면 음의 연관성이 있는지를 알 수 없었다. 그러나 본 논문에서 제안한 코사인 순수 신뢰도는 그 부호에 의해 연관성 규칙의 방향을 알 수 있으므로 신뢰도와 코사인 유사성 측도가 가지고 있는 약점을 보완할 수 있는 측도라는 사실을 확인하였다.

주요용어: 데이터 마이닝, 빅 데이터, 연관성 규칙, 코사인 순수 신뢰도, 코사인 유사성 측도.

1. 서론

오늘날 정부기관이나 기업에서는 대용량 데이터로부터 알려지지 않은 흥미롭고 가치 있는 정보를 얻기를 원함에 따라 빅 데이터 (big data) 분석의 필요성이 대두되었다. 이는 기존의 방식으로는 수집, 저장, 분석이 힘든 데이터를 분석하는 기술을 의미하며, 여기에는 정형화된 데이터뿐만 아니라 비정형화된 데이터를 포함한다. 빅 데이터 분석은 국가 안전 및 위험관리 뿐만 아니라 정치, 사회, 경제, 문화, 과학 기술 등 여러 분야에 활용될 수 있다 (Jung, 2012). 빅 데이터를 분석하기 위해서는 우선 방대한 양의 데이터 속에서 가치 있는 정보를 찾는 것이 선행되어야 하는데, 이를 위한 기법이 데이터 마이닝 (data mining)이다. 이 기법은 빅 데이터로부터 유용한 정보를 추출하는 과정을 의미하며, 마케팅, 소매업, 금융업, 제조업, 의료 분야 등 사회 전반에 걸쳐 활용되고 있다 (Park, 2011a). 특히 빅 데이터와 관련된 데이터마이닝 기법으로는 텍스트 마이닝 (text mining), 평판 분석 (opinion mining), 군집 분석 (cluster analysis), 연관성 규칙 마이닝 (association rule mining) 등이 있다. 이 중에서 연관성 규칙 마이닝은 빅 데이터 속에서 항목들 간의 의미 있는 연관성을 찾아내기 위한 기법으로, 연관성 정도를 지지도 (support), 신뢰도 (confidence), 그리고 향상도 (lift) 등의 연관성 평가 기준에 의해 수치화함으로써 합리적인 의사결정을 위해 활용할 수 있다. 이 기법은 조직의 의사결정 문제, 기업의 교차판매나 고객관

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

리, 그리고 보험회사 및 병원과 관련된 의료분야 등 여러 분야에서 활용되고 있고 Agrawal 등 (1993)에 의해 처음 소개되었으며, 그 이후로 연관성 규칙의 제안 및 효율성 개선과 연관성 측도의 개발 및 기능 향상 등 다양한 연구가 진행되어 왔다 (Liu 등, 1999; Saygin 등, 2002; Cho와 Park, 2011a; Cho와 Park, 2011b; Jin 등, 2011; Park, 2011a; Park, 2011b; Park, 2012a; Park, 2012b).

연관성 규칙은 데이터의 형태와 계산방법이 간단하고 분석 결과의 이해가 용이한 반면, 품목의 수가 많을 경우 상당한 계산 과정이 필요하므로 적절한 대상 항목을 선정하여 적용하는 것이 매우 중요하다. 따라서 연관성 규칙의 생성 유무를 판단할 수 있는 연관성 평가 기준이 매우 중요한 역할을 담당한다고 볼 수 있다. 그러나 기존의 신뢰도에 의해서는 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 Ahn과 Kim (2003)은 의학 분야에서 널리 이용되고 있는 기여위험률 (attributable risk)을 순수 신뢰도 (net confidence)라는 이름으로 데이터 마이닝 분야에 적용한 바 있다. 그러나 순수 신뢰도는 순수하게 특정 요인에 의해서만 결과가 얼마인가를 나타내주는 측도이며, 부호에 의해 양의 관련성과 음의 관련성을 판단할 수 있기는 하나, 양의 신뢰도와 음의 신뢰도의 값의 차이가 동일하면 순수 신뢰도의 값도 동일하게 되는 단점을 가지고 있다. 이러한 문제를 보완하기 위해 Park (2011b)는 기여 순수 신뢰도 (attributably pure confidence)를 제안한 바 있으나 동시 비 발생 빈도가 알려져 있지 않은 경우에는 이들을 이용할 수 없는 단점이 있다. 동시 비 발생 빈도가 알려져 있지 않은 경우에는 군집분석이나 다차원 분석에서 활용되는 코사인 계열 (cosine family)의 유사성 측도를 활용할 수 있는데 이들은 모두 양의 값만을 가지고 있기 때문에 연관성의 방향을 가늠할 수 없다. 이에 본 논문에서는 동시 비 발생 빈도가 알려져 있지 않은 경우에 활용할 수 있는 코사인 순수 신뢰도 (cosine net confidence)를 제안하고자 한다. 본 논문의 2절에서는 제안하는 흥미도 측도인 코사인 순수 신뢰도를 정의한 후 여러 가지 특성을 살펴보는 동시에 Piatetsky-Shapiro (1991)가 제안한 흥미도 측도의 기준에 대한 충족여부를 점검한다. 3절에서는 예제를 통하여 기존의 코사인 유사성 측도와의 비교를 통해 그 유용성에 대해 알아본 후, 4절에서 결론을 내리고자 한다.

2. 코사인 순수 신뢰도

이 절에서는 두 항목집합 간의 코사인 순수 신뢰도를 제안하고자 한다. 하나의 트랜잭션에서 항목집합 X와 Y의 연관성의 정도를 측정하기 위해 Park (2012b)에서와 같이 Table 2.1의 2×2 분할표를 활용하여 기본적인 연관성 평가 기준과 코사인 순수 신뢰도에 대해 논의하고자 한다.

Table 2.1 2×2 contingency table

		Y		Total
		1	0	
X	1	a	b	$a+b$
	0	c	d	$c+d$
Total		$a+c$	$b+d$	n

먼저 기존의 연관성 규칙에서 일반적으로 적용하고 있는 지지도 $S(X \Rightarrow Y)$는 항목 집합 X와 항목 집합 Y가 동시에 발생하는 거래의 비율을 의미하며, Table 2.1로부터 a/n으로 계산된다. 신뢰도 $C(X \Rightarrow Y)$는 항목 집합 X가 포함된 거래 비율 중 항목 집합 X와 항목 집합 Y가 동시에 포함된 거래의 비율을 의미하며, $a/(a+b)$이 된다. 향상도 $L(X \Rightarrow Y)$는 항목 집합 X를 구매한 경우 그 거래가 항목 집합 Y를 포함하는 경우와 항목 집합 Y가 임의로 구매되는 경우의 비를 의미하며, $an/[(a+b)(a+$

$c)$]로 계산된다 (Park, 2012a). 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 Ahn과 Kim (2003)은 의학분야에서 널리 이용되고 있는 기여위험률을 순수 신뢰도 ($Nconf$)라는 이름으로 데이터 마이닝 분야에 적용한 바 있다.

$$Nconf(A \Rightarrow B) = P(Y|X) - P(Y|\overline{X})$$

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 (2011b)은 $P(Y|X)$와 $P(Y|\overline{X})$의 차이 크기를 $P(Y|X)$에 대해 상대적으로 나타낸 기여 순수 신뢰도 ($APconf$)를 제안하였다.

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

이 측도는 의학분야에서 노출군과 비노출군을 합한 전체 집단에서 발생한 환자 중에서 요인에 의해서 발생한 환자가 차지하는 비율을 나타내는 기여 분율 (attributable fraction)을 연관성 규칙의 평가기준에 적합하도록 변형한 것이다. 그런데 그동안 개발된 순수 신뢰도와 관련된 흥미도 측도들은 대부분 동시 비 발생 빈도가 알려진 경우에는 활용할 수 있으나 동시 비 발생 빈도가 알려져 있지 않은 경우에는 이들을 이용할 수 없는 것이었다. 이에 본 논문에서는 동시 비 발생 빈도가 알려져 있지 않은 경우에 활용할 수 있는 코사인 계열의 유사성 측도인 코사인 순수 신뢰도를 제안하고자 한다. 먼저 코사인 측도를 제시하면 다음의 식 (2.1)과 같다.

$$S_C = \frac{P(XY)}{\sqrt{P(X)P(Y)}} = \frac{a}{\sqrt{(a+b)(a+c)}} \tag{2.1}$$

이러한 코사인 측도는 연관성의 방향을 나타내주지 못하므로 음의 연관성이 있는 경우에도 양의 값으로 나타나고 있어서 연구자들에게 혼란을 줄 수 있으므로 두 항목의 발생을 나타내는 주변 비율을 고려하여 다음과 같이 변형한 식 (2.2)의 코사인 순수 신뢰도를 제안하고자 한다.

$$S_{CN} = \frac{P(XY) - \frac{1}{2}\max[P(X), P(Y)]}{\sqrt{P(X)P(Y)}} \tag{2.2}$$

본 논문에서 제안한 측도 S_{CN}이 Piatetsky-Shapiro (1991)가 제안한 흥미도 측도의 조건을 충족하는지의 여부를 증명하기 위해 수식을 정리하면 다음과 같이 표현된다.

$$S_{CN} = \begin{cases} \dfrac{2P(XY) - P(X)}{2\sqrt{P(X)P(Y)}}, \text{ if } \max[P(X), P(Y)] = P(X) \\ \\ \dfrac{2P(XY) - P(Y)}{2\sqrt{P(X)P(Y)}}, \text{ if } \max[P(X), P(Y)] = P(Y) \end{cases}$$

위의 두 식으로부터 S_{CN}은 $P(X)$ 또는 $P(Y)$에 따라 감소하고, $P(XY)$의 값에 따라 단조 증가한다는 사실을 알 수 있으므로 흥미도 측도의 첫 번째 조건과 두 번째 조건을 만족하고 있다. 반면에 S_{CN}은

$P(X)$ 또는 $P(Y)$가 0.5인 경우에만 세 번째 조건인 두 항목이 독립이면 인과적 연관성 평가 기준은 0이 된다는 조건을 만족하므로 일반적으로는 이 조건을 충족하지 않는다고 할 수 있다. 그러나 기존의 지지도와 신뢰도가 이 기준을 만족하지 않는다고 하더라도 의미 있는 연관성 규칙 발견을 위해 중요한 역할을 담당한 것과 마찬가지로 S_{CN}도 평가 기준으로서 의미 있는 역할을 담당한다고 볼 수 있다.

다음으로 코사인 순수 신뢰도 S_{CN}이 가지고 있는 성질을 기술하면 다음과 같다.

성질 2.1 $\max[b,c] = b$이면 S_{CN}은 X가 전항이고 Y가 후항인 경우의 순수 신뢰도를 나타내고, $\max[b,c] = c$이면 S_{CN}은 Y가 전항이고 X가 후항인 경우의 순수 신뢰도를 나타내고 있다.

(설명) : 식 (2.2)를 조건부 확률을 이용하여 정리하면 먼저 $\max[b,c] = b$인 경우에는 다음과 같이 나타낼 수 있다.

$$S_{CN} = \begin{cases} \dfrac{\sqrt{P(X)}}{2\sqrt{P(Y)}}[P(Y|X) - P(\overline{Y}|X)], \text{ if } \max[b,c] = b \\ \\ \dfrac{\sqrt{P(Y)}}{2\sqrt{P(X)}}[P(X|Y) - P(\overline{X}|Y)], \text{ if } \max[b,c] = c \end{cases}$$

따라서 S_{CN}은 b와 c의 상대적 크기에 따라 전항과 후항이 바뀌는 형태로 순수 신뢰도가 표현된다고 할 수 있다.

성질 2.2 $S_{CN} > 0$ 이면 X와 Y가 양의 연관성을 가지고, $S_{CN} < 0$ 이면 X와 Y가 음의 연관성을 가지는 것을 의미한다.(설명) : 식 (2.2)를 Table 2.1의 기호를 이용하여 나타내면 다음과 같이 표현된다.

$$S_{CN} = \begin{cases} \dfrac{a-b}{2\sqrt{(a+b)(a+c)}}, \text{ if } \max[b,c] = b \\ \\ \dfrac{a-c}{2\sqrt{(a+b)(a+c)}}, \text{ if } \max[b,c] = c \end{cases}$$

위의 두 식으로부터 $a > b$ 또는 $a > c$이면 $S_{CN} > 0$이므로 X와 Y가 양의 연관성을 가지게 되고, 이와는 반대로 $a < b$ 또는 $a < c$이면 $S_{CN} < 0$이므로 X와 Y가 음의 연관성을 가지게 된다는 사실을 알 수 있다.

성질 2.3 S_{CN} 값의 범위는 $[-1, +1]$이다.

(설명) : S_{CN} 값이 1이라는 의미는 S_{CN}의 분자와 분모 값이 1이므로 항목집합 X가 발견되는 모든 트랜잭션에서 항목집합 Y가 발견되고 X가 없는 트랜잭션에서는 Y가 전혀 발생하지 않는다는 의미이다. 그리고 S_{CN}의 값이 -1라는 의미는 S_{CN}의 분자 값이 -1이므로 항목집합 X가 발견되지 않는 트랜잭션에서만 거의 모든 항목집합 Y가 발견된다는 것이다.

성질 2.4 $P(XY)$가 $P(X)$ 또는 $P(Y)$의 1/2 보다 크면 양의 연관성이 있다.

(설명) : 식 (2.2)에서 $\max[P(X), P(Y)] = P(X)$인 경우에는 $P(X)$가 $P(XY) + P(X\overline{Y})$이므로 $P(XY)$가 $P(X)$의 1/2 보다 크면 $P(XY)$가 $P(X\overline{Y})$ 보다 커기 때문에 양의 연관성이 있다고 볼 수 있다. 또한 $\max[P(X), P(Y)] = P(Y)$인 경우에는 $P(Y)$가 $P(XY) + P(\overline{X}Y)$이므로 $P(XY)$가 $P(Y)$의 1/2 보다 크면 $P(XY)$가 $P(\overline{X}Y)$ 보다 커기 때문에 양의 연관성이 있다고 볼 수 있다. 본 논문에서 제안하는 측도는 전항과 후항이 바뀌더라도 이들 중에서 주변비율이 큰 것을 고려하는 동시에 방향성이 고려된 순수한 연관성의 정도를 나타내고 있기 때문에 기존의 신뢰도나 코사인 측도에서 나타나는 문제를 해결할 수 있다.

이 외에도 본 논문에서 제시하는 측도 S_{CN}은 b와 c 중에서 큰 값을 취하므로 어느 것이 큰 값을 갖느냐에 따라 전항과 후항의 위치가 결정되고 부호에 의해 순수한 연관성의 방향을 알 수 있다. 따라서 이 측도는 신뢰도의 단점을 보완할 수 있는 측도이며, 동시 비 발생 빈도 d를 알 수 없는 경우에 지지도, 향상도, 순수 신뢰도, 기여 순수 신뢰도는 측정이 불가능하나 이 측도는 d를 고려하지 않으므로 측정이 가능하다. 또한 이 측도는 a에서 b와 c 중에서 큰 값을 빼주므로 즉, $P(XY)$에서 $P(\overline{X}Y)$와 $P(X\overline{Y})$ 중에서 큰 값을 빼주므로 상당히 안전한 순수 흥미도 측도라고 할 수 있다. 이에 대한 아이디어는 Fager와 McGowan (1963)이 제안한 유사성 측도로부터 얻은 것이다.

3. 예제 데이터에 의한 고찰

본 절에서는 예제를 통해 신뢰도와 순수 신뢰도의 문제점을 탐색하고 코사인 순수 신뢰도의 유용성을 고찰하고자 한다. 이를 위해 Park (2012a)에서와 같이 항목 집합 X, Y에 대해 Table 3.1과 같이 가정하였다.

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 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 Table 3.1과 같다. 이 표에서 a가 취할 수 있는 범위는 $0 \le a \le 30$이며, P(X)=0.5이고 P(Y)=0.3이다.

Table 3.1로부터 동시발생빈도 (a)에 따른 지지도 $P(XY)$, 신뢰도인 $P(Y|X)$ 및 $P(X|Y)$, 음의 신뢰도인 $P(\overline{Y}|X)$ 및 $P(X|\overline{Y})$, 코사인 유사성 측도 S_C, 그리고 코사인 순수 신뢰도 S_{CN}을 계산하여 그 일부를 나타내면 Table 3.2와 같다. 이 표에서 $b = P(X = 1, Y = 0)$, $c = P(X = 0, Y = 1)$, $d = P(X = 0, Y = 0)$을 의미한다. 이 표로부터 알 수 있는 바와 같이 동시발생빈도 a의 값이 커질수록 $P(XY)$, $P(Y|X)$, $P(X|Y)$, S_C, 그리고 S_{CN}은 증가하고 있는 반면에 음의 신뢰도인 $P(\overline{Y}|X)$와 $P(X|\overline{Y})$은 감소하고 있다. 또한 신뢰도 $P(Y|X)$와 $P(X|Y)$, 그리고 코사인 유사성 측도 S_C는 모두 양의 값을 가지므로 연관성의 방향을 알 수 없어서 그 값만으로는 양의 연관성이 있는지 아니면 음의 연관성이 있는지를 알 수 없다. 그러나 본 논문에서 제안하는 S_{CN}을 연관성 측도로 활용하면 그 부호에 의해 연관성 규칙의 방향을 알 수 있으므로 양의 연관성이 더 강한지, 아니면 음의 연관성이 더 강한지를 파악할 수 있다. 이에 대해 좀 더 구체적으로 알아보기 위해 a=23, b=27, c=7, d=43인 경우와 a=27, b=23, c=3, d=47인 경우를 비교해보면, 신뢰도 $P(Y|X)$와 $P(X|Y)$는 각각 0.460과 0.767, 0.540과 0.900, 음의 신뢰도인 $P(\overline{Y}|X)$와 $P(X|\overline{Y})$는 각각 0.540과 0.233, 0.460과 0.100, S_C와 S_{CN}은 각각 0.594와 -0.052, 0.697과 0.052로 나타나서 a가 증가하면 신뢰도와 코사인 측도, 그리고 코사인 순수 신뢰도는 모두 증가하며, 음의 신뢰도는 감소하고 있다. 또한 신뢰도와 코사인 측도는 각각 0.460과 0.540, 0.767과 0.900, 그리고 0.594와 0.697로 두 경우 모두 양의 값으로 나타나나 코사인 순수 신뢰도는 양의 신뢰도와 음의 신뢰도를 동시에 고려함으로써 각각 -0.052와 0.052로 나타나게 되어 연관성의 방향을 파악할 수 있는 측도가 되는 동시에 두 경우의 절대값의 크기는 동일하므로 대칭형의 측도라고도 할 수 있다.

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

a	b	c	d	$P(XY)$	$P(Y\|X)$	$P(X\|Y)$	$P(Y\|X)$	$P(X\|Y)$	S_C	S_{CN}
11	39	19	31	0.110	0.220	0.367	0.780	0.633	0.284	-0.361
12	38	18	32	0.120	0.240	0.400	0.760	0.600	0.310	-0.336
13	37	17	33	0.130	0.260	0.433	0.740	0.567	0.336	-0.310
14	36	16	34	0.140	0.280	0.467	0.720	0.533	0.361	-0.284
15	35	15	35	0.150	0.300	0.500	0.700	0.500	0.387	-0.258
16	34	14	36	0.160	0.320	0.533	0.680	0.467	0.413	-0.232
17	33	13	37	0.170	0.340	0.567	0.660	0.433	0.439	-0.207
18	32	12	38	0.180	0.360	0.600	0.640	0.400	0.465	-0.181
19	31	11	39	0.190	0.380	0.633	0.620	0.367	0.491	-0.155
20	30	10	40	0.200	0.400	0.667	0.600	0.333	0.516	-0.129
21	29	9	41	0.210	0.420	0.700	0.580	0.300	0.542	-0.103
22	28	8	42	0.220	0.440	0.733	0.560	0.267	0.568	-0.077
23	27	7	43	0.230	0.460	0.767	0.540	0.233	0.594	-0.052
24	26	6	44	0.240	0.480	0.800	0.520	0.200	0.620	-0.026
25	25	5	45	0.250	0.500	0.833	0.500	0.167	0.645	0.000
26	24	4	46	0.260	0.520	0.867	0.480	0.133	0.671	0.026
27	23	3	47	0.270	0.540	0.900	0.460	0.100	0.697	0.052
28	22	2	48	0.280	0.560	0.933	0.440	0.067	0.723	0.077
29	21	1	49	0.290	0.580	0.967	0.420	0.033	0.749	0.103
30	20	0	50	0.300	0.600	1.000	0.400	0.000	0.775	0.129

이번에는 불일치빈도 b의 값의 변화에 따른 지지도 $P(XY)$, 신뢰도인 $P(Y|X)$ 및 $P(X|Y)$, 음의 신뢰도인 $P(\overline{Y}|X)$ 및 $P(X|\overline{Y})$, 코사인 유사성 측도 S_C, 그리고 코사인 순수 신뢰도 S_{CN}의 값을 비교하기 위해 다음과 같이 각 셀의 값을 바꾸어 실험하였다. Table 3.3에서 b가 취할 수 있는 정수 값의 범위는 $0 \leq b \leq 30$이며, $P(X)$=0.5이고 $P(Y)$=0.7이다.

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

이 표로부터 각 셀 값의 변화에 따른 $P(XY)$, $P(Y|X)$, $P(X|Y)$, $P(\overline{Y}|X)$, $P(X|\overline{Y})$, S_C, 그리고 본 논문에서 제안하는 코사인 순수 신뢰도 S_{CN}을 계산하여 그 일부를 나타내면 다음의 Table 3.4와 같다.

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

a	b	c	d	$P(XY)$	$P(Y\|X)$	$P(X\|Y)$	$P(Y\|X)$	$P(X\|Y)$	S_C	S_{CN}
40	10	30	20	0.400	0.800	0.571	0.200	0.429	0.676	0.085
39	11	31	19	0.390	0.780	0.557	0.220	0.443	0.659	0.068
38	12	32	18	0.380	0.760	0.543	0.240	0.457	0.642	0.051
37	13	33	17	0.370	0.740	0.529	0.260	0.471	0.625	0.034
36	14	34	16	0.360	0.720	0.514	0.280	0.486	0.609	0.017
35	15	35	15	0.350	0.700	0.500	0.300	0.500	0.592	0.000
34	16	36	14	0.340	0.680	0.486	0.320	0.514	0.575	-0.017
33	17	37	13	0.330	0.660	0.471	0.340	0.529	0.558	-0.034
32	18	38	12	0.320	0.640	0.457	0.360	0.543	0.541	-0.051
31	19	39	11	0.310	0.620	0.443	0.380	0.557	0.524	-0.068
30	20	40	10	0.300	0.600	0.429	0.400	0.571	0.507	-0.085
29	21	41	9	0.290	0.580	0.414	0.420	0.586	0.490	-0.101
28	22	42	8	0.280	0.560	0.400	0.440	0.600	0.473	-0.118
27	23	43	7	0.270	0.540	0.386	0.460	0.614	0.456	-0.135
26	24	44	6	0.260	0.520	0.371	0.480	0.629	0.439	-0.152
25	25	45	5	0.250	0.500	0.357	0.500	0.643	0.423	-0.169
24	26	46	4	0.240	0.480	0.343	0.520	0.657	0.406	-0.186
23	27	47	3	0.230	0.460	0.329	0.540	0.671	0.389	-0.203
22	28	48	2	0.220	0.440	0.314	0.560	0.686	0.372	-0.220
21	29	49	1	0.210	0.420	0.300	0.580	0.700	0.355	-0.237
20	30	50	0	0.200	0.400	0.286	0.600	0.714	0.338	-0.254

이 표로부터 알 수 있는 바와 같이 불일치빈도 b의 값이 커질수록 음의 신뢰도인 $P(\overline{Y}|X)$와 $P(X|\overline{Y})$은 증가하고 있는 반면에, 다른 측도들 $P(XY)$, $P(Y|X)$, $P(X|Y)$, S_C, 그리고 S_{CN}은 감소하고 있다. 또한 이 표에서 보는 바와 같이 b의 값이 커짐에도 불구하고 신뢰도 $P(Y|X)$와 $P(X|Y)$, 그리고 코사인 유사성 측도 S_C는 모두 양의 값을 가지므로 음의 연관성의 정도를 나타내주지 못한다. 그러나 본 논문에서 제안하는 S_{CN}을 연관성 측도로 활용하면 그 부호에 의해 연관성 규칙의 방향을 알 수 있으므로 연관성의 강도와 방향을 파악할 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, 신뢰도 $P(Y|X)$와 $P(X|Y)$는 각각 0.760과 0.543, 0.640과 0.457, 음의 신뢰도인 $P(\overline{Y}|X)$와 $P(X|\overline{Y})$는 각각 0.240과 0.457, 0.360과 0.543, S_C와 S_{CN}은 각각 0.642와 0.051, 0.541과 -0.051로 나타나서 b가 증가하면 음의 신뢰도는 증가하는 반면에 신뢰도와 코사인 측도, 그리고 코사인 순수 신뢰도는 모두 감소하며, 또한 신뢰도와 코사인 측도는 각각 0.760과 0.543, 0.642와 0.640, 그리고 0.457와 0.541로 두 경우 모두 양의 값으로 나타나나 코사인 순수 신뢰도는 양의 신뢰도와 음의 신뢰도를 동시에 고려함으로써 각각 0.051과 -0.051로 나타나게 되어 연관성의 방향을 가늠할 수 있으며, 이 또한 두 경우의 절대값의 크기는 동일하므로 대칭형의 측도가 된다.

마지막으로 불일치빈도 c의 값의 변화에 따른 지지도 $P(XY)$, 신뢰도인 $P(Y|X)$ 및 $P(X|Y)$, 음의 신뢰도인 $P(\overline{Y}|X)$ 및 $P(X|\overline{Y})$, 코사인 유사성 측도 S_C, 그리고 코사인 순수 신뢰도 S_{CN}의 값을 비교하기 위해 다음과 같이 각 셀의 값을 바꾸어 실험하였다. Table 3.5에서 c가 취할 수 있는 정수 값의 범위는 $0 \le c \le 20$이며, $P(X)$=0.8이고 $P(Y)$=0.3이다.

Table 3.5 Simulation data(3)

		Y		Total
		1	0	
X	1	$30-c$	$50+c$	80
	0	c	$20-b$	20
Total		30	70	100

이 표로부터 각 셀 값의 변화에 따른 $P(XY)$, $P(Y|X)$, $P(X|Y)$, $P(\overline{Y}|X)$, $P(X|\overline{Y})$, S_C, 그리고 본 논문에서 제안하는 코사인 순수 신뢰도 S_{CN}을 계산하면 다음의 Table 3.6과 같은 결과를 얻을 수 있다.

Table 3.6 Output of some association thresholds by simulation data(3)

a	b	c	d	$P(XY)$	$P(Y\|X)$	$P(X\|Y)$	$P(\overline{Y}\|X)$	$P(X\|\overline{Y})$	S_C	S_{CN}
21	59	9	11	0.210	0.263	0.700	0.738	0.300	0.429	-0.388
20	60	10	10	0.200	0.250	0.667	0.750	0.333	0.408	-0.408
19	61	11	9	0.190	0.238	0.633	0.763	0.367	0.388	-0.429
18	62	12	8	0.180	0.225	0.600	0.775	0.400	0.367	-0.449
17	63	13	7	0.170	0.213	0.567	0.788	0.433	0.347	-0.469
16	64	14	6	0.160	0.200	0.533	0.800	0.467	0.327	-0.490
15	65	15	5	0.150	0.188	0.500	0.813	0.500	0.306	-0.510
14	66	16	4	0.140	0.175	0.467	0.825	0.533	0.286	-0.531
13	67	17	3	0.130	0.163	0.433	0.838	0.567	0.265	-0.551
12	68	18	2	0.120	0.150	0.400	0.850	0.600	0.245	-0.572
11	69	19	1	0.110	0.138	0.367	0.863	0.633	0.225	-0.592
10	70	20	0	0.100	0.125	0.333	0.875	0.667	0.204	-0.612

이 표로부터 알 수 있는 바와 같이 불일치빈도 c의 값이 커질수록 음의 신뢰도인 $P(\overline{Y}|X)$와 $P(X|\overline{Y})$은 증가하고 있는 반면에, 다른 측도들 $P(XY)$, $P(Y|X)$, $P(X|Y)$, S_C, 그리고 S_{CN}은 감소하고 있다. 또한 이 표에서 보는 바와 같이 c의 값이 커짐에도 불구하고 신뢰도 $P(Y|X)$와 $P(X|Y)$, 그리고 코사인 측도 S_C는 모두 양의 값을 가지므로 음의 연관성의 정도를 나타내주지 못한다. 그러나 본 논문에서 제

안하는 코사인 순수 신뢰도 S_{CN}을 연관성 측도로 활용하면 그 부호에 의해 연관성 규칙의 방향을 알 수 있으므로 연관성의 강도와 방향을 파악할 수 있다. 이러한 결과는 불일치빈도 b의 변화에 따른 양상과 일치하고 있다.

4. 결론

위키백과사전에 의하면 빅 데이터 기술의 발전은 다변화된 현대 사회를 보다 정확하게 예측하고 효율적으로 작동하도록 정보를 제공하며 개인화된 현대 사회 구성원들에게 있어서 맞춤형 정보를 제공, 관리, 분석 가능케 하며 과거에는 불가능 했던 기술을 진일보 시킨다고 한다. 따라서 빅 데이터 분석에 활용 가능한 데이터 마이닝 기법은 정치, 사회, 경제, 문화, 과학기술과 같은 전 영역에 걸쳐 그 중요성이 부각되고 있다. 본 논문에서는 데이터 마이닝 기법 중에서 많이 활용되고 있는 연관성 규칙의 평가 기준으로 코사인 순수 신뢰도를 제안한 후, Piatetsky-Shapiro가 제안한 흥미도 측도의 기준에 대한 충족여부를 점검하는 동시에 여러 가지 특성을 살펴보았다. 또한 예제를 통하여 고찰한 결과, 기존의 신뢰도, 순수 신뢰도, 그리고 기여 순수 신뢰도가 가지고 있는 약점을 보완할 수 있는 측도라는 사실을 확인하였다. 이를 좀 더 구체적으로 기술하면 다음과 같다.

첫째, 동시발생빈도의 값이 커질수록 $P(XY)$, $P(Y|X)$, $P(X|Y)$, S_C, 그리고 S_{CN}은 증가하고 있는 반면에 음의 신뢰도인 $P(\overline{Y}|X)$와 $P(X|\overline{Y})$은 감소하였다. 또한 신뢰도 $P(Y|X)$와 $P(X|Y)$, 그리고 코사인 유사성 측도 S_C는 모두 양의 값을 가지므로 연관성의 방향을 알 수 없어서 그 값만으로는 양의 연관성이 있는지 아니면 음의 연관성이 있는지를 알 수 없었다. 그러나 본 논문에서 제안하는 S_{CN}을 연관성 측도로 활용하면 그 부호에 의해 연관성 규칙의 방향을 알 수 있으므로 양의 연관성이 더 강한지, 아니면 음의 연관성이 더 강한지를 파악할 수 있었다.

둘째, 불일치빈도 b와 c값이 커질수록 음의 신뢰도인 $P(\overline{Y}|X)$와 $P(X|\overline{Y})$은 증가하고 있는 반면에, 다른 측도들 $P(XY)$, $P(Y|X)$, $P(X|Y)$, S_C, 그리고 S_{CN}은 감소하고 있다. 또한 불일치빈도의 값이 커짐에도 불구하고 신뢰도 $P(Y|X)$와 $P(X|Y)$, 그리고 코사인 유사성 측도 S_C는 모두 양의 값을 가지므로 음의 연관성의 정도를 나타내주지 못한다. 그러나 본 논문에서 제안하는 S_{CN}을 연관성 측도로 활용하면 그 부호에 의해 연관성 규칙의 방향을 알 수 있으므로 연관성의 강도와 방향을 파악할 수 있다.

이 논문에서는 단지 두 가지 항목에 대한 결과만을 살펴보았으나, 향후에는 본 논문에서 제시한 측도를 보다 현실적인 상황에 적용하기 여러 가지 항목을 동시에 고려한 측도로 확장하는 연구가 필요할 것으로 사료된다.

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

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). A study on insignificant rules discovery in association rule mining. *Journal of the Korean Data & Information Science Society*, **22**, 81-88.

Fager, E. W. and McGowan, J. A. (1963). Zooplankton species groups in the North Pacific. *Science*, **140**, 453-460.

Jin, D. S., Kang, C., Kim, K. K., Choi, S. B. (2011). CRM on travel agency using association rules. *Journal of the Korean Data Analysis Society*, **13**, 2945-2952.

Jung, Y. C. (2012). *Big data*, Communicationbooks Press, Seoul.

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. (2011a). Association rule ranking function by decreased lift influence. *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. (2012a). Negatively attributable and pure confidence for generation of negative association rules. *Journal of the Korean Data & Information Science Society*, **23**, 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.

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.

Journal of the Korean Data & Information Science Society
2014, **25**(1), 97–106

http://dx.doi.org/10.7465/jkdi.2014.25.1.97
한국데이터정보과학회지

The proposition of cosine net confidence in association rule mining

Hee Chang Park[1]

[1]Department of Statistics, Changwon National University

Received 11 December 2013, revised 26 December 2013, accepted 2 January 2014

Abstract

The development of big data technology was to more accurately predict diversified contemporary society and to more efficiently operate it, and to enable impossible technique in the past. This technology can be utilized in various fields such as the social science, economics, politics, cultural sector, and science technology at the national level. It is a prerequisite to find valuable information by data mining techniques in order to analyze big data. Data mining techniques associated with big data involve text mining, opinion mining, cluster analysis, association rule mining, and so on. 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, similarity measures, etc.This paper proposed cosine net confidence as association thresholds, and checked the conditions of interestingness measure proposed by Piatetsky-Shapiro, and examined various characteristics. The comparative studies with basic confidence and cosine similarity, and cosine net confidence were shown by numerical example. The results showed that cosine net confidence are better than basic confidence and cosine similarity because of the relevant direction.

Keywords: Association rule, big data, cosine net confidence, cosine similarity measure, data mining.

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