Journal of the Korean Data & Information Science Society
2013, **24**(6), 1189–1197

http://dx.doi.org/10.7465/jkdi.2013.24.6.1189
한국데이터정보과학회지

인과적 연관성 규칙 평가 기준의 제안

박희창[1]

[1]창원대학교 통계학과

접수 2013년 7월 15일, 수정 2013년 8월 13일, 게재확정 2013년 8월 18일

요 약

연관성 규칙 마이닝은 지지도, 신뢰도, 향상도 등의 흥미도 측도를 기반으로 하여 대용량 데이터베이스를 구성하고 있는 항목들 간의 관련성을 찾아내는 기법이다. 이 기법은 기업의 의사결정 문제, 유통업에서의 교차판매, 고객관리 등 현업에서 많이 활용되고는 있으나, 이러한 기본적인 연관성 평가 기준만으로는 두 항목 간의 인과관계를 설명할 수 없다. 본 논문에서는 이러한 문제를 해결하기 위해 인과적 연관성 규칙을 제안하는 동시에, 고려하는 평가 기준들이 흥미도 측도의 조건을 충족하는지의 여부를 점검하였다. 본 논문에서 제안한 인과적 향상도는 세 가지 조건 모두를 만족하는 것으로 입증되었다. 인과적 지지도와 인과적 신뢰도는 동시 발생 확률의 값에 따라 단조 증가하는 조건과 각 항목의 주변 확률의 값에 따라 단조 감소하는 조건은 만족하였다. 반면에 두 항목이 독립이면 연관성 평가 기준의 값이 1이 되는 조건에 대해서는 기존의 지지도와 신뢰도와 같이 이 조건이 충족되지 않았다. 또한 예제를 통해 기존의 연관성 평가 기준과 인과적 연관성 평가 기준을 비교해 본 결과, 기존의 평가 측도인 지지도와 신뢰도를 기준으로 연관성 규칙 생성 여부를 판단했을 때 탈락되는 규칙도 인과적 평가 기준인 인과적 지지도와 인과적 신뢰도를 이용하여 판단하게 되면 연관성 규칙으로 채택할 수 있다는 사실을 발견하였다.

주요용어: 데이터 마이닝, 연관성 규칙, 인과적 신뢰도, 인과적 지지도, 인과적 향상도.

1. 서론

오늘날 국가, 기업, 그리고 조직 간의 경쟁이 심화되고 정보의 중요성에 대한 인식이 확산됨에 따라 축적된 엄청난 크기의 데이터베이스에 함축적으로 내재되어 있는 지식이나 패턴을 찾아내는 데이터 마이닝 (data mining) 기법이 주목을 받고 있다. 데이터마이닝이란 의미 있는 정보를 발견하기 위해서 통계 및 수학적 기술과 패턴인식 기술 등의 도구를 이용하여 대용량의 데이터를 탐색하고 분석하는 과정을 의미한다. 데이터 마이닝 기법 중에서도 연관성 규칙 (association rule)은 지지도, 신뢰도, 그리고 향상도 등 여러 가지 흥미도 측도 (interestingness measure)의 값을 근거로 하여 고려하는 항목들 간의 관련성 정도를 측정한다. 이 기법은 대용량 데이터베이스로부터 항목 간에 연관성을 탐색하는 것으로 기업의 의사결정 문제, 유통업에서의 교차판매나 고객관리, 보험 및 의료, 생물 정보학 (Bioinformatics) 등 다양한 분야에서 활용되고 있다. 이러한 연관성 규칙은 Agrawal 등 (1993)이 처음 소개하였으며, 그 이후로 연관성 규칙의 제안 및 효율성 개선과 연관성 측도의 개발 및 기능 향상 등 제약 기반 연관성 규칙과 관련된 연구가 진행되어 왔다. 이들 중에서 Agrawal과 Srikant (1994)는 Apriori, AprioriTid 알고리즘을 제안하였고, Park 등 (1995)은 partitioning 알고리즘을 제안한 바 있다. Sergey 등 (1997)에

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

의해 DIC (dynamic itemset counting) 등의 발전된 알고리즘들이 연구되었으며, Saygin 등 (2002)에 의해 트랜잭션에 있는 데이터를 별도의 값으로 대치시키고 이때의 최소 지지도와 최대 지지도의 범위를 조정하면서 고려할 항목 집합의 수를 줄이는 방법이 제안되었다. 국내연구로는 Park (2011, 2012a, 2012b), Cho와 Park (2011a, 2011b) 등이 있다.

본 논문에서는 두 항목의 인과적인 관련성 유무를 파악할 수 있는 인과적 연관성 규칙 (casual association rule)의 평가 기준인 인과적 지지도 (causal support), 인과적 신뢰도 (causal confidence), 그리고 인과적 향상도 (causal lift)에 대해 논의하고자 한다. 이 중에서 인과적 지지도와 신뢰도는 Kodratoff (2000)와 Berzal 등 (2005)이 제안한 것을 소개하며, 이들의 아이디어를 확장하여 인과적 향상도를 고안함으로써 기본적인 연관성 평가 기준을 대체할 수 있는 인과적 연관성 규칙의 평가 기준을 제시한다. 본 논문의 2절에서는 인과적 연관성 규칙의 평가 측도를 제시한 후, 이들이 Piatetsky-Shapiro (1991)가 제안한 흥미도 측도의 조건을 충족하는지의 여부를 점검한다. 3절에서는 예제를 통하여 기존의 연관성 평가 기준과 인과적 연관성 평가 기준과의 비교를 통해 인과적 연관성 규칙의 유용성에 대해 알아본 후, 4절에서 결론을 내리고자 한다.

2. 인과적 연관성 규칙의 평가 측도

일반적인 연관성 규칙 마이닝에서는 먼저 사용자가 지정한 최소 지지도를 만족시키는 빈발항목집합을 생성한 후, 이들에 대해 미리 정의한 최저신뢰도 기준을 만족하는 것을 규칙으로 채택하게 된다 (Park, 2012b). 본 절에서는 하나의 트랜잭션에서 항목 X와 Y의 인과적 연관성의 정도를 측정하기 위해 Table 2.1과 같은 2×2분할표를 활용하여 인과적 연관성 규칙 평가 기준에 대해 논의하고자 한다.

Table 2.1 2×2contingency table

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

2.1. 기본적인 연관성 평가 기준

연관성 규칙을 평가하는 기준에는 지지도, 신뢰도, 향상도 등이 있다. 지지도 $supp(X \Rightarrow Y)$는 항목 집합 X와 항목 집합 Y가 동시에 발생하는 거래의 비율을 의미하며, 다음과 같이 정의된다.

$$supp(X \Rightarrow Y) = P(X \text{ and } Y) = \frac{a}{n}.$$

신뢰도 $conf(X \Rightarrow Y)$는 항목 집합 X가 포함된 거래 비율 중 항목 집합 X와 항목 집합 Y가 동시에 포함된 거래의 비율을 의미하며, 다음과 같이 정의된다.

$$conf(X \Rightarrow Y) = P(Y|X) = \frac{a}{a+b}.$$

향상도 $lift(X \Rightarrow Y)$는 항목 집합 X를 구매한 경우 그 거래가 항목 집합 Y를 포함하는 경우와 항목 집합 Y가 임의로 구매되는 경우의 비를 의미하며, 다음과 같이 정의된다.

$$lift(X \Rightarrow Y) = \frac{P(Y|X)}{P(Y)} = \frac{na}{(a+b)(a+c)}.$$

2.2. 인과적 연관성 평가 기준

이 절에서는 두 항목 간의 인과적 연관성 규칙을 생성하기 위한 평가 기준을 제안하고자 한다. 기존의 연관성 평가 기준인 지지도, 신뢰도, 향상도 등은 항목 집합이 발생하는 경우만을 고려한 측도이다. Kodratoff (2000)와 Berzal 등 (2005)은 기본적인 연관성 평가 기준만으로는 전항과 후항의 인과관계를 설명할 수 없으므로 $Y^c \Rightarrow X^c$의 경우를 동시에 고려하여야 한다는 의미에서 다음과 같은 인과적 지지도 (causal support)와 인과적 신뢰도 (causal confidence)를 분류 목적을 위해 제안한 바 있다.

$$supp_c(X \Rightarrow Y) = P(X \text{ and } Y) + P(X^c \text{ and } Y^c) = \frac{a+d}{n}, \tag{2.1}$$

$$conf_c(X \Rightarrow Y) = \frac{1}{2}[P(Y|X) + P(X^c|Y^c)] = \frac{ab + 2ad + bd}{2(a+b)(b+d)}. \tag{2.2}$$

여기서 X^c와 Y^c의 의미는 각각 X와 Y가 일어나지 않음을 의미한다. 특히 Berzal 등 (2005)은 $X \Rightarrow Y$과 $Y^c \Rightarrow X^c$는 수학적으로 동일하므로 이들을 동시에 고려하면 잠재적으로 유용한 규칙을 생성할 수 있는 것으로 판단하였다. 또한 Kodratoff (2000)는 흡연과 암발생 유무와의 인과관계를 파악하기 위해서는 흡연(X)하였을 때 암(Y)이 발생하는 경우와 암이 발생하지 않았을 때 흡연하지 않은 경우의 두 가지를 동시에 고려하는 것이 바람직하다고 생각하였다. 본 논문에서는 이들의 아이디어를 확장하여 식 (2.3)과 같은 인과적 향상도 (causal lift)를 고안함으로써 기본적인 연관성 평가 기준을 대체할 수 있는 인과적 연관성 규칙의 기준을 제안하고자 한다.

$$lift_c(X \Rightarrow Y) = \frac{1}{2}\left[\frac{P(Y|X)}{P(Y)} + \frac{P(X^c|Y^c)}{P(X^c)}\right] = \frac{n}{2}\left[\frac{a}{(a+b)(a+c)} + \frac{d}{(b+d)(c+d)}\right]. \tag{2.3}$$

이러한 인과적 연관성 평가 기준은 $X \Rightarrow Y$과 $Y^c \Rightarrow X^c$인 경우를 동시에 고려하기 때문에 희귀한 사건의 발생에 대한 연관성 규칙에도 적용 가능할 것으로 생각된다.

본 논문에서 고려하는 인과적 연관성 평가 기준들이 Piatetsky-Shapiro (1991)가 제안한 흥미도 측도의 조건을 충족하는지의 여부를 조사하면 다음과 같다.

[조건 1] 인과적 연관성 평가 기준은 $P(X)$와 $P(Y)$가 고정되어 있을 때, $P(X \text{ and } Y)$의 값에 따라 단조 증가한다.

(증명) 먼저 인과적 지지도 $supp_c$는 $P(X)$와 $P(Y)$가 고정되어 있을 때, 식 (2.1)이 식 (2.4)와 같이 정리되므로 $P(X \text{ and } Y)$가 증가하면 $supp_c$가 증가하는 것을 쉽게 알 수 있다.

$$supp_c(X \Rightarrow Y) = 1 - P(X) - P(Y) + 2P(X \text{ and } Y). \tag{2.4}$$

따라서 인과적 지지도에 대한 증명은 이것으로 마무리된다.

인과적 신뢰도 $conf_c$의 조건 충족 여부를 증명하기 위해 식 (2.2)의 $conf_c$를 a에 대해 편미분하면 다음과 같이 양의 값을 가지므로 $conf_c$는 $P(X \text{ and } Y)$의 값에 따라 단조 증가한다. 이것으로 인과적 신뢰도에 대한 증명은 마무리된다.

$$\frac{\partial conf_c(X \Rightarrow Y)}{\partial a} = \frac{b}{2(a+b)^2}.$$

인과적 향상도 $lift_c$의 조건 충족 여부를 증명하기 위해 식 (2.3)의 $lift_c$를 a에 대해 편미분하면 다음 식과 같이 0보다 큰 값이 되므로 $P(X \text{ and } Y)$의 값에 따라 $lift_c$는 단조 증가한다.

$$\frac{\partial lift_c(X \Rightarrow Y)}{\partial a} = \frac{bc(2a + b + 2d)}{[(a+b)(a+c)]^2}.$$

따라서 인과적 신뢰도에 대한 증명은 이것으로 마무리된다. □

[조건 2] 인과적 연관성 평가 기준은 $P(X)$의 값에 따라 단조 감소한다.

(증명) 먼저 식 (2.4)부터 인과적 지지도 $supp_c$는 $P(X)$의 값이 증가함에 따라 단조 감소함을 쉽게 알 수 있다.

인과적 신뢰도 $conf_c$의 조건 충족 여부를 증명하기 위해 먼저 식 (2.2)를 정리하면 다음과 같다.

$$conf_c(X \Rightarrow Y) = \frac{1}{2}\left[\frac{P(X \text{ and } Y)}{P(X)} + \frac{1 - P(X) - P(Y) + P(X \text{ and } Y)}{1 - P(Y)}\right].$$

이 식으로부터 $P(X)$의 값이 증가하면 첫 번째 항의 분모가 증가하고, 두 번째 항의 분자가 감소하므로 $conf_c$의 값은 감소하게 된다.

인과적 향상도 $lift_c$의 조건 충족 여부를 증명하기 위해 식 (2.3)를 정리하면 다음과 같다.

$$lift_c(X \Rightarrow Y) = \frac{1}{2}\left[\frac{P(X \text{ and } Y)}{P(X)P(Y)} + \frac{1 - P(X) - P(Y) + P(X \text{ and } Y)}{[1 - P(X)][1 - P(Y)]}\right]. \tag{2.5}$$

이 식에서 보는 바와 같이 $P(X)$의 값이 증가하면 첫 번째 항의 분모가 증가하게 된다. 그리고 두 번째 항을 $P(X)$에 대해 편미분하면 다음과 같은 식이 얻어지며, 음의 값을 갖는다.

$$\frac{\partial lift_c(X \Rightarrow Y)}{\partial P(X)} = \frac{-[P(Y) - P(X \text{ and } Y)][1 - P(Y)]}{[(1 - P(X))(1 - P(Y))]^2}.$$

따라서 $P(X)$의 값이 증가함에 따라 식 (2.5)의 첫 번째 항과 두 번째 모두 감소하게 되므로 $lift_c$는 감소하게 된다. □

[조건 3] $P(X \text{ and } Y) = P(X)P(Y)$이면 인과적 연관성 평가 기준은 1이 된다.

(증명) 원칙적으로는 $P(X \text{ and } Y) = P(X)P(Y)$이면 인과적 연관성 평가 기준이 0이 되어야 하나 기존의 향상도에서도 1이면 두 항목이 독립적이므로 이를 이용하여 조건 충족 여부를 점검하고자 한다. $P(X \text{ and } Y) = P(X)P(Y)$이면 식 (2.3)의 첫 번째 항과 두 번째 항은 모두 1이 되어 인과적 향상도 $lift_c$는 1이 되므로 $lift_c$는 이 조건을 만족하게 된다. □

반면에 인과적 지지도 $supp_c$의 경우 $P(X \text{ and } Y) = P(X)P(Y)$이면 식 (2.1)은 $1 - P(X) - P(Y) + 2P(X)P(Y)$가 되어 이 조건을 충족하지 않으나, 기존의 지지도와 마찬가지로 연관성 규칙 생성을 위한 첫 번째 단계가 최소 지지도를 만족시키는 빈발항목집합을 생성하는 과정이므로 의미 있는 규칙 발견을 위해 필요한 측도라고 볼 수 있다. 또한 인과적 신뢰도의 경우에도 $P(X \text{ and } Y) = P(X)P(Y)$이면 식 (2.2)로부터 $[1 + P(Y)]/2$이 되므로 인과적 신뢰도 또한 이 조건을 만족하지 않는다. 기존의 신뢰도의 경우에도 이 조건은 충족되지 않았으나 연관성 평가 기준 중에서 가장 중심이 되는 측도로써 중요한 역할을 담당한 것과 마찬가지로 $conf_c$도 인과적 연관성 규칙에서 의미 있는 역할을 담당하게 된다.

3. 예제를 통한 고찰

본 절에서는 예제를 통하여 기존의 연관성 규칙의 대표적인 흥미도 측도인 지지도, 신뢰도 및 향상도와 인과적 연관성 규칙을 위한 인과적 지지도, 인과적 신뢰도, 그리고 인과적 향상도에 대해 수치적인 비교를 하고자 한다. 이를 위해 먼저 데이터베이스에 있는 총 트랜잭션의 수 (t)를 50명으로 하고, 항목 집합 X는 흡연 유무를 나타내는 것으로 흡연자 (1)의 수를 20명으로 하고, 비흡연자 (0)의 수를 80명으로 하였다. 또한 항목 집합 Y를 암발생 유무를 나타내는 것으로 암환자 (1)의 수와 정상인 (0)의 수를 각각 50명으로 하였다. 항목 집합 X와 Y가 동시에 발생한 빈도 수, 즉 암환자이면서 흡연자의 빈도수는 a명으로 하였다. 이를 정리하면 Table 3.1과 같다.

Table 3.1 Simulation data(1)

		Y		Total
		1	0	
X	1	a	$20-a$	20
	0	$50+a$	$30-a$	80
Total		50	50	100

이 표에서 a가 취할 수 있는 정수 값의 범위 각각 $0 \le a \le 20$이다. 이로부터 a의 변화에 따른 기존의 평가 기준들과 인과적 평가 기준들을 미니탭 16의 계산기 기능을 이용하여 계산한 후, Table 3.2에 제시하였다. 이 표에서 기술한 기호는 다음과 같다.

$$a = n(X=1, Y=1),\ b = n(X=1, Y=0),$$
$$c = n(X=0, Y=1),\ d = n(X=0, Y=0),$$
$$supp_1 = P(X=1, Y=1),\ upp_2 = P(X=0, Y=0),$$
$$conf_1 = P(Y=1|X=1),\ conf_2 = P(X=0|Y=0),$$
$$lift_1 = P(Y=1|X=1)/P(Y=1),\ lift_2 = P(X=0|Y=0)/P(X=0).$$

이 표에서 보는 바와 같이 $P(X=1)$의 값은 공히 0.200이며, a와 c가 증가하고 b와 d가 감소함에 따라 지지도 $supp_1$은 증가하고 $supp_2$는 감소하며 인과적 신뢰도 $supp_c$는 0.300으로 동일한 값을 갖는 것으로 나타났다. 또한 신뢰도를 살펴보면 $conf_1$, $conf_2$, 그리고 $conf_c$모두 증가하는 것으로 나타났으며, 향상도 $lift_1$, $lift_2$, 그리고 $lift_c$도 모두 증가하는 것으로 나타났다. 특히 기존의 평가 측도인 $supp_1$과 $conf_1$을 기준으로 연관성 규칙 생성 여부를 판단했을 때 탈락되는 규칙도 인과적 평가 기준인 $supp_c$와 $conf_c$를 이용하여 판단하게 되면 연관성 규칙으로 채택할 수 있다. 이를 좀 더 구체적으로 알아보기 위해 $a=9$, $b=11$, $c=59$, $d=21$일 때를 살펴보면 $supp_1 = 0.090$과 $conf_1 = 0.450$으로 나타나서 최저 지지도와 최저 신뢰도를 각각 0.2와 0.5인 경우에는 연관성 규칙으로 생성되지 않는다. 그러나 인과적 평가 기준인 $supp_c conf_c$을 이용한다면 각각의 값이 0.300과 0.553으로 계산되어서 최저 지지도와 최저 신뢰도의 기준을 충족하므로 연관성 규칙으로 채택되는 것을 알 수 있다. 향상도의 경우에는 $lift_1$과 $lift_2$의 값이 모두 1 보다 큰 경우에만 $lift_c$의 값이 1 보다 커짐을 알 수 있으며, 기존의 $lift_1$보다는 $lift_c$이 더 큰 값을 갖는다는 사실을 표를 통해 확인할 수 있다.

Table 3.2 Association thresholds by simulation data(1)

a	b	c	d	$P(Y)$	$supp_1$	$supp_2$	$supp_c$	$conf_1$	$conf_2$	$conf_c$	$lift_1$	$lift_2$	$lift_c$
1	19	51	29	0.520	0.010	0.290	0.300	0.050	0.604	0.327	0.096	0.755	0.426
2	18	52	28	0.540	0.020	0.280	0.300	0.100	0.609	0.354	0.185	0.761	0.473
3	17	53	27	0.560	0.030	0.270	0.300	0.150	0.614	0.382	0.268	0.767	0.517
4	16	54	26	0.580	0.040	0.260	0.300	0.200	0.619	0.410	0.345	0.774	0.559
5	15	55	25	0.600	0.050	0.250	0.300	0.250	0.625	0.438	0.417	0.781	0.599
6	14	56	24	0.620	0.060	0.240	0.300	0.300	0.632	0.466	0.484	0.789	0.637
7	13	57	23	0.640	0.070	0.230	0.300	0.350	0.639	0.494	0.547	0.799	0.673
8	12	58	22	0.660	0.080	0.220	0.300	0.400	0.647	0.524	0.606	0.809	0.707
9	11	59	21	0.680	0.090	0.210	0.300	0.450	0.656	0.553	0.662	0.820	0.741
10	10	60	20	0.700	0.100	0.200	0.300	0.500	0.667	0.583	0.714	0.833	0.774
11	9	61	19	0.720	0.110	0.190	0.300	0.550	0.679	0.614	0.764	0.848	0.806
12	8	62	18	0.740	0.120	0.180	0.300	0.600	0.692	0.646	0.811	0.865	0.838
13	7	63	17	0.760	0.130	0.170	0.300	0.650	0.708	0.679	0.855	0.885	0.870
14	6	64	16	0.780	0.140	0.160	0.300	0.700	0.727	0.714	0.897	0.909	0.903
15	5	65	15	0.800	0.150	0.150	0.300	0.750	0.750	0.750	0.938	0.938	0.938
16	4	66	14	0.820	0.160	0.140	0.300	0.800	0.778	0.789	0.976	0.972	0.974
17	3	67	13	0.840	0.170	0.130	0.300	0.850	0.813	0.831	1.012	1.016	1.014
18	2	68	12	0.860	0.180	0.120	0.300	0.900	0.857	0.879	1.047	1.071	1.059
19	1	69	11	0.880	0.190	0.110	0.300	0.950	0.917	0.933	1.080	1.146	1.113
20	0	70	10	0.900	0.200	0.100	0.300	1.000	1.000	1.000	1.111	1.250	1.181

이번에는 인과적 연관성 평가 기준들의 유용성을 $P(X)$의 변화 양상과 함께 알아보기 위해 총 트랜잭션의 수 (t)를 50명으로 하고, 흡연자 (1)의 수를 30명으로 하고, 비흡연자 (0)의 수를 20명으로 하였다. 또한 암환자 (1)의 수를 $35-e-r$명으로 하고 정상인 (0)의 수를 $15+e+r$명으로 하였다. 항목 집합 X와 Y가 동시에 발생한 빈도 수, 즉 암환자이면서 흡연자의 빈도수는 $10-r$명으로 하였다. 이를 정리하면 Table 3.3과 같다.

Table 3.3 Simulation data(2)

		Y		Total
		1	0	
X	1	$10-r$	$25-e$	$35-e-r$
	0	$10+r$	$5+e$	$15+e+r$
Total		20	30	50

이 표에서 e및 r이 취할 수 있는 정수 값의 범위 각각 $0 \le e \le 25$와 $0 \le r \le 10$이다. 이로부터 e및 r의 변화에 따른 기존의 평가 기준들과 인과적 평가 기준들을 Table 3.4와 Table 3.5에 제시하였다. Table 3.4는 a를 오름차순으로 정리한 후, a의 값이 같으면 b로, b의 값이 같으면 c로 그리고 c의 오름차순이 같으면 d로 정리하여 그 일부를 나타낸 표이다.

Table 3.4 Association thresholds ascending by a

a	b	c	d	$P(X)$	$supp_1$	$supp_2$	$supp_c$	$conf_1$	$conf_2$	$conf_c$	$lift_1$	$lift_2$	$lift_c$
3	0	17	30	0.060	0.060	0.600	0.660	1.000	1.000	1.000	2.500	1.064	1.782
3	1	17	29	0.080	0.060	0.580	0.640	0.750	0.967	0.858	1.875	1.051	1.463
3	2	17	28	0.100	0.060	0.560	0.620	0.600	0.933	0.767	1.500	1.037	1.269
3	3	17	27	0.120	0.060	0.540	0.600	0.500	0.900	0.700	1.250	1.023	1.136
3	4	17	26	0.140	0.060	0.520	0.580	0.429	0.867	0.648	1.071	1.008	1.040
3	5	17	25	0.160	0.060	0.500	0.560	0.375	0.833	0.604	0.938	0.992	0.965
3	6	17	24	0.180	0.060	0.480	0.540	0.333	0.800	0.567	0.833	0.976	0.904
4	0	16	30	0.080	0.080	0.600	0.680	1.000	1.000	1.000	2.500	1.087	1.793
4	1	16	29	0.100	0.080	0.580	0.660	0.800	0.967	0.883	2.000	1.074	1.537
4	2	16	28	0.120	0.080	0.560	0.640	0.667	0.933	0.800	1.667	1.061	1.364
4	3	16	27	0.140	0.080	0.540	0.620	0.571	0.900	0.736	1.429	1.047	1.238
4	4	16	26	0.160	0.080	0.520	0.600	0.500	0.867	0.683	1.250	1.032	1.141
4	5	16	25	0.180	0.080	0.500	0.580	0.444	0.833	0.639	1.111	1.016	1.064
4	7	16	23	0.220	0.080	0.460	0.540	0.364	0.767	0.565	0.909	0.983	0.946
5	1	15	29	0.120	0.100	0.580	0.680	0.833	0.967	0.900	2.083	1.098	1.591
5	2	15	28	0.140	0.100	0.560	0.660	0.714	0.933	0.824	1.786	1.085	1.435
5	3	15	27	0.160	0.100	0.540	0.640	0.625	0.900	0.763	1.563	1.071	1.317
5	4	15	26	0.180	0.100	0.520	0.620	0.556	0.867	0.711	1.389	1.057	1.223
5	5	15	25	0.200	0.100	0.500	0.600	0.500	0.833	0.667	1.250	1.042	1.146
5	6	15	24	0.220	0.100	0.480	0.580	0.455	0.800	0.627	1.136	1.026	1.081
5	7	15	23	0.240	0.100	0.460	0.560	0.417	0.767	0.592	1.042	1.009	1.025
5	8	15	22	0.260	0.100	0.440	0.540	0.385	0.733	0.559	0.962	0.991	0.976
6	3	14	27	0.180	0.120	0.540	0.660	0.667	0.900	0.783	1.667	1.098	1.382
6	4	14	26	0.200	0.120	0.520	0.640	0.600	0.867	0.733	1.500	1.083	1.292
6	5	14	25	0.220	0.120	0.500	0.620	0.545	0.833	0.689	1.364	1.068	1.216
6	6	14	24	0.240	0.120	0.480	0.600	0.500	0.800	0.650	1.250	1.053	1.151
6	7	14	23	0.260	0.120	0.460	0.580	0.462	0.767	0.614	1.154	1.036	1.095
6	8	14	22	0.280	0.120	0.440	0.560	0.429	0.733	0.581	1.071	1.019	1.045

이 표에서 $P(Y=1)$는 공히 0.600이며, a와 c를 고정한 상태에서 b가 증가하고 d가 감소함에 따라 $supp_1$의 값은 동일한 반면에 그 외의 모든 측도들은 감소하고 있는 것으로 나타났다. 또한 b와 d를 고정한 상태에서 a가 증가하고 c가 감소함에 따라 $supp_2$와 $conf_2$는 값이 고정되어 있는 반면에 그 외의 모든 측도들은 증가하고 있는 것으로 나타났다. 이를 좀 더 구체적으로 살펴보기 위해 $a=5$와 $c=15$인 경우를 살펴보면 $supp_1$은 모두 0.100으로 나타났으며, $b=5$이고 $d=25$인 경우에는 $supp_2=0.500$, $supp_c=0.600$, $conf_1=0.500$, $conf_2=0.833$, $conf_c=0.667$, $lift_1=1.250$, $lift_2=$

1.042, 그리고 $lift_c = 1.146$으로 계산되었다. 또한 $b = 6$이고 $d = 24$인 경우에는 $supp_2 = 0.480$, $supp_c = 0.580$, $conf_1 = 0.455$, $conf_2 = 0.800$, $conf_c = 0.627$, $lift_1 = 1.136$, $lift_2 = 1.026$, 그리고 $lift_c = 1.081$으로 계산되었다. 따라서 a와 c의 값이 동일한 상황에서 b가 증가하거나 d가 감소하면 $supp_1$을 제외한 모든 측도의 값이 감소하였다. $b = 3$과 $d = 27$인 경우를 살펴보면 $supp_2 = 0.540$, $conf_2 = 0.900$으로 동일한 값으로 나타났으며, $a = 3$과 $c = 17$인 경우에는 $supp_1 = 0.060$, $supp_c = 0.600$, $conf_1 = 0.500$, $conf_c = 0.700$, $lift_1 = 1.250$, $lift_2 = 1.023$, 그리고 $lift_c = 1.136$으로 계산되었다. $a = 6$과 $c = 14$인 경우 $supp_1 = 0.120$, $supp_c = 0.660$, $conf_1 = 0.667$, $conf_c = 0.783$, $lift_1 = 1.667$, $lift_2 = 1.098$, 그리고 $lift_c = 1.382$로 계산되었다. 따라서 b와 d의 값이 동일한 상황에서 a가 증가하거나 c가 감소하면 $supp_2$와 $conf_2$을 제외한 모든 측도의 값이 증가하였다. 또한 a의 값이 증가함에 따라 기준의 평가 기준들은 증가하며, a또는 d의 값이 증가하면 인과적 신뢰도가 증가하는 것을 예제를 통해 확인할 수 있었다.

향상도의 경우에도 앞에서와 마찬가지로 $lift_1$과 $lift_2$의 값이 모두 1 보다 큰 경우에만 $lift_c$의 값이 1 보다 커짐을 알 수 있었다. 위에서와 마찬가지로 이 표에서도 위에서와 동일한 최저 지지도와 최저 신뢰도의 기준을 적용할 때 탈락되는 규칙들이 인과적 신뢰도를 이용하게 되면 연관성 규칙이 성립하는 경우가 많이 나타나는 것으로 확인되었다.

4. 결론

데이터 마이닝 기법들 중에서 연관성 규칙은 방대한 데이터베이스에서 항목 간의 관계를 지지도, 신뢰도, 향상도 등의 연관성 평가 기준을 기반으로 명확히 수치화함으로써 관련성을 표시하여 주기 때문에 현장에서 직접 적용이 가능하다. 그러나 기본적인 연관성 평가 기준만으로는 전항과 후항의 인과관계를 설명할 수 없다.

본 논문에서는 이러한 문제를 해결하기 위해 인과적 연관성 규칙에 대한 평가 기준인 인과적 지지도, 인과적 신뢰도, 그리고 인과적 향상도를 제안하였다. 또한 이들이 흥미도 측도의 조건을 충족하는지의 여부를 점검하였다. 세 가지 측도 모두 두 항목이 동시 발생 확률의 값에 따라 단조 증가하는 조건과 각 항목의 주변 확률의 값에 따라 단조 감소하는 조건은 만족하였다. 반면에 두 항목이 독립이면 연관성 평가 기준의 값이 1이 되는 조건에 대해서는 인과적 지지도와 신뢰도는 기존의 지지도와 신뢰도와 같이 이 조건을 충족하지 않는다. 그러나 인과적 지지도는 기존의 지지도와 마찬가지로 인과적 연관성 규칙 생성을 위한 첫 번째 단계라고 볼 수 있으므로 의미 있는 규칙 발견을 위해 필요한 측도이다. 또한 인과적 신뢰도의 경우에도 기존의 신뢰도와 마찬가지로 인과적 연관성 평가 기준 중에서 가장 중심이 되는 측도이므로 중요한 역할을 담당하는 측도로 볼 수 있다. 다음으로는 모의실험을 통하여 기존의 연관성 평가 기준과 인과적 연관성 평가 기준과의 비교를 통해 인과적 연관성 규칙의 유용성에 대해 알아보았다. 그 결과, 기존의 평가 측도인 지지도와 신뢰도를 기준으로 연관성 규칙 생성 여부를 판단했을 때 탈락되는 규칙도 인과적 평가 기준인 인과적 지지도와 인과적 신뢰도를 이용하여 판단하게 되면 연관성 규칙으로 채택할 수 있다는 사실을 발견하였다. 따라서 이러한 인과적 연관성 평가 기준은 희귀한 사건의 발생에 대한 연관성 규칙에도 적용할 수 있는 것으로 판단된다.

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.

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. *Proceedings of the 20th VLDB Conference*, 487-499.

Berzal, F., Cubero, J., Marin, N., Sanchez, D., Serrano, J. and Vila, A. (2005). Association rule evaluation for classification purposes. *Actas del III Taller Nacional de Minería de Datos y Aprendizaje, TAMIDA2005*, 135-144.

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.

Kodratoff, Y. (2000). Comparing machine learning and knowledge discovery in databases: An application to knowledge discovery in texts. *Proceeding of Machine Learning and its Applications: Advanced Lectures*, 1-21.

Park, H. C. (2011). Association rule ranking function by decreased lift influence. *Journal of the Korean Data & Information Science Society*, **22**, 179-188.

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.

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*, 104-123.

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.

Sergey, B., Rajeev M., Jeffrey D.U. and Shalom T. (1997). Dynamic itemset counting and implication rules for market data. *Proceedings of ACM SIGMOD Conference on Management of Data*, 255-264.

Journal of the Korean Data & Information Science Society
2013, **24**(6), 1189–1197

http://dx.doi.org/10.7465/jkdi.2013.24.6.1189
한국데이터정보과학회지

Proposition of causal association rule thresholds

Hee Chang Park[1]

[1]Department of Statistics, Changwon National University

Received 15 July 2013, revised 13 August 2013, accepted 18 August 2013

Abstract

Data mining is the process of analyzing a huge database from different perspectives and summarizing it into useful information. One of the well-studied problems in data mining is association rule generation. Association rule mining finds the relationship among several items in massive volume database using the interestingness measures such as support, confidence, lift, etc. Typical applications for this technique include retail market basket analysis, item recommendation systems, cross-selling, customer relationship management, etc. But these interestingness measures cannot be used to establish a causality relationship between antecedent and consequent item sets.This paper propose causal association thresholds to compensate for this problem, and then check the three conditions of interestingness measures. The comparative studies with basic and causal association thresholds are shown by numerical example. The results show that causal association thresholds are better than basic association thresholds.

Keywords: Association rule, causal confidence, causal lift, causal support, data mining.

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