*Journal of the Korean Data & Information Science Society*
*2011,* **22**(3), 495–503

한국데이터정보과학회지

# 이분형 예측 유사성 측도의 연관성 평가 기준 적용 방안

박희창[1]

[1]창원대학교 통계학과

접수 2011년 4월 11일, 수정 2011년 5월 9일, 게재확정 2011년 5월 19일

**요 약**

데이터 마이닝에서의 연관성 규칙은 방대한 양의 데이터베이스에 내재되어 있는 항목들 간의 관련성을 수치화 하는 방법이다. 의미 있는 연관성 규칙을 탐사하기 위한 가장 기본적인 연관성 규칙 평가 기준에는 지지도, 신뢰도, 향상도 등이 있다. 이들 중에서 향상도는 그 값에 의해 양의 연관성이 있는지 아니면 음의 연관성이 있는지, 즉 연관성의 방향을 알 수 있는 반면에 지지도와 신뢰도는 그 방향을 알 수가 없다. 이를 위해 순수 신뢰도와 기여 순수 신뢰도가 제안되었으나 이들 또한 단점을 안고 있다. 본 논문에서는 기존의 여러 형태의 신뢰도가 가지고 있는 문제점을 해결하기 위해 군집분석이나 다차원 분석에서 활용되고 있는 이분형 예측 유사성 측도 중에서 -1과 1 사이의 값을 가지는 Yule의 $Y$ 및 $Q$ 측도를 연관성 평가 기준으로 제안하였다. 또한 기존의 순수 신뢰도 및 기여 순수 신뢰도의 문제점을 파악한 후, 예제를 통하여 이분형 예측 유사성 측도의 유용성에 관해 알아보았다. 그 결과, 본 논문에서 고려한 유사성 측도들은 기존의 측도들이 가지고 있는 문제점을 해결할 수 있어서 본 논문에서 제안한 이분형 예측 유사성 측도가 연관성 평가 기준으로 활용할 수 있다는 사실을 확인하였다.

주요용어: 기여 순수 신뢰도, 순수 신뢰도, 신뢰도, 연관성 평가 기준, 유사성 측도.

## 1. 서론

오늘날 유통업, 제조업, 병원, 보험회사 등 다양한 분야에서 장바구니 분석, 교차 마케팅, 카탈로그 디자인 등에 적용되고 있는 연관규칙 (association rule)은 방대한 양의 데이터에 내재되어 있는 항목들 간의 유용한 관련성을 찾아내는 데 활용되고 있다. 데이터 마이닝에서 연관성 규칙은 장바구니와 같은 거래 데이터 (transaction data)의 분석에 기초를 두고 있으며, 탐사된 규칙에 대하여 연관성 평가 측도를 활용하는 정량적인 의미와 데이터들의 상호 연관성들을 쉽게 파악할 수 있는 정성적인 의미를 동시에 함축하고 있다 (Srikant와 Agrawal, 1995). 이러한 연관성 규칙은 Agrawal 등 (1993)에 의해 처음 제안한 이후 많은 학자들이 다양한 관점에서 연관성 규칙과 관련된 연구를 수행하였다 (Agrawal과 Srikant, 1994; Park 등, 1995; Cai 등, 1998; Han과 Fu, 1999; Liu 등, 1999;, Pasquier 등, 1999; Han 등, 2000; Pei 등, 2000; Park과 Cho, 2005; Cho와 Park, 2008; Choi와 Park, 2008; Park, 2008; Park, 2009 등).

연관성 규칙에서 적용되는 데이터의 형태는 발생시점에서 기록되어진 항목에 관한 정보만으로 구성되어 있다. 의미 있는 연관성 규칙을 탐색하기 위한 가장 기본적인 흥미도 측도에는 지지도 (support), 신뢰도 (confidence), 향상도 (lift) 등이 있으며, 이러한 연관성 규칙 평가 기준을 이용하여 연관성 규칙을 생성하게 된다. 일반적인 연관성 규칙 생성과정은 먼저 사용자가 지정한 최소 지지도를 만족시키는 빈발항목집합을 생성한 후, 이들에 대해 향상도가 1이상이면서 최저 신뢰도 기준을 만족하는 규칙으로

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

채택하게 된다. 이 때 사용되는 향상도는 그 값에 의해 양의 연관성이 있는지 아니면 음의 연관성이 있는지, 즉 연관성의 방향을 알 수 있는 반면에 지지도와 신뢰도는 계산된 값만으로는 방향을 알 수가 없다. 이를 위해 안광일과 김성집 (2003)은 의학분야에서 널리 이용되고 있는 기여위험률 (attributable risk)을 순수 신뢰도 (net confidence ; $Nconf$)라는 이름으로 데이터 마이닝 분야에 적용한 바 있다. 그러나 순수 신뢰도는 부호에 의해 양의 관련성과 음의 관련성을 판단할 수 있기는 하나, 양의 신뢰도와 음의 신뢰도의 값의 차이가 동일하면 순수 신뢰도의 값이 동일하게 되는 단점을 가지고 있다. 또한 이 문제를 해결하기 위해 Park (2011)은 기여 순수 신뢰도 (attributably pure confidence ; $APconf$)를 제안하였으나 이 측도의 범위에 대한 하한 값의 제한이 없어서 그 값만으로는 연관성 정도를 파악하기가 쉽지 않다.

본 논문에서는 이러한 기존의 연관성 측도들이 가지고 있는 문제점을 해결하기 위해 군집분석이나 다차원 분석에서 활용되고 있는 이분형 예측 유사성 측도 (predictive similarity measures for binary data) 중에서 Yule의 $Y$ 및 $Q$ 측도를 연관성 평가 기준으로 제안하고자 한다. 본 논문의 2절에서는 본 논문에서 제안하는 측도들을 소개한 후, 기존의 순수 신뢰도와 기여 순수 신뢰도의 문제점을 기술하고자 한다. 그리고 3절에서는 예제를 통하여 이들 측도들과 기존의 순수 신뢰도 및 기여 순수 신뢰도와 비교함으로써 본 논문에서 제시한 측도들의 유용성을 살펴본 후, 4절에서 결론을 맺고자 한다.

## 2. 예측을 위한 이분형 유사성 측도

가장 기본적인 연관성 규칙을 평가하는 기준에는 지지도, 신뢰도, 향상도 등이 있다.

**표 2.1** $2 \times 2$교차표

| | | $B$ | | 합계 |
|---|---|---|---|---|
| | | 1 | 0 | |
| $A$ | 1 | $a$ | $b$ | $a+b$ |
| | 0 | $c$ | $d$ | $c+d$ |
| 합계 | | $a+c$ | $b+d$ | $n$ |

지지도 $S(A \Rightarrow B)$는 항목 집합 $A$와 항목 집합 $B$가 동시에 발생하는 거래의 비율을 의미하며, 향상도 $L(A \Rightarrow B)$는 항목 집합 $A$를 구매한 경우 그 거래가 항목 집합 $B$를 포함하는 경우와 항목 집합 $B$가 임의로 구매되는 경우의 비를 의미한다. 그리고 신뢰도 $C(A \Rightarrow B)$는 항목 집합 $A$가 포함된 거래 비율 중 항목 집합 $A$와 항목 집합 $B$가 동시에 포함된 거래의 비율을 의미한다. 이들에 대해 표 2.1과 같은 분할표를 이용하여 수식으로 나타내면 다음과 같다. 표 2.1에서 각 항목의 값이 0이라는 의미는 그 항목이 발생하지 않았다는 의미이고, 1이라는 의미는 그 항목이 발생했다는 의미이다. 또한 $a$는 동시 발생 빈도로서 $n(A \text{ and } B)$을 의미하며, $b = n(A \text{ and } B^c)$, $c = n(A^c \text{ and } B)$, $d = n(A^c \text{ and } B^c)$을 의미한다.

$$S(A \Rightarrow B) = P(A \text{ and } B) = \frac{a}{n}$$

$$C(A \Rightarrow B) = P(B|A) = \frac{a}{a+b}$$

$$L(A \Rightarrow B) = \frac{P(B|A)}{P(B)} = \frac{a \cdot n}{(a+b) \cdot (a+c)}.$$

안광일과 김성집 (2003)은 의학분야에서 널리 이용되고 있는 기여위험률 (attributable risk)을 순수

신뢰도라는 이름으로 데이터 마이닝 분야에 적용한 바 있다.

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

여기서 $\bar{X}$의 의미는 $X$가 일어나지 않음을 의미한다. Park (2011)에서 지적한 바와 같이 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 $P(Y|\bar{X})$의 값이 어떤 값을 가지더라도 두 값의 차이가 동일하면 순수 신뢰도의 값도 동일하게 되는 단점을 가지고 있다. 이러한 문제점을 해결하기 위해 Park (2011)은 다음과 같은 기여 순수 신뢰도를 제안한 바 있다.

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

그러나 이 측도값의 범위는 $[-\infty, 1]$이 되는데, 연관성을 평가하는 기준이 값의 범위에 제한이 없으면 그 값에 의해서는 연관성을 강도를 파악하기가 어렵다.

기존의 순수 신뢰도와 기여 순수 신뢰도가 가지고 있는 단점에 대해 좀 더 구체적으로 알아보기 위해 다음과 같은 가상의 예제를 이용하여 설명하면 다음과 같다.

**표 2.2** 가상의 분할표 (1)

| | | Y | | 합계 |
|---|---|---|---|---|
| | | 1 | 0 | |
| X | 1 | 20 | 80 | 100 |
| | 0 | 60 | 40 | 100 |
| 합계 | | 80 | 120 | 200 |

먼저 표 2.2와 표 2.3으로부터 신뢰도인 $P(Y|X)$을 계산하면 각각 0.2와 0.35이며, $P(Y|\bar{X})$의 값은 각각 0.6과 0.75로 나타났다. 순수 신뢰도는 두 표 모두 -0.4로 동일하므로 순수 신뢰도만을 가지고는 이 두 경우의 차이를 설명할 수 없게 된다. 또한 기여 순수 신뢰도는 각 표에서 -2.0과 -1.14로 계산되어서 두 표 간에 연관성 정도의 차이를 규명할 수는 있으나 범위의 하한에 대한 제한이 없어서 -1 보다 더 작은 값을 가질 수 있으므로 그 값만 가지고는 연관성의 강도를 측정하기가 곤란해진다.

**표 2.3** 가상의 분할표 (2)

| | | Y | | 합계 |
|---|---|---|---|---|
| | | 1 | 0 | |
| X | 1 | 35 | 65 | 100 |
| | 0 | 75 | 25 | 100 |
| 합계 | | 110 | 90 | 200 |

이를 위해 본 논문에서는 예측을 위한 이분형 유사성 측도들을 연관성 평가기준으로 제안한 후 이들의 유용성에 대해 연구하고자 한다. Romesburg (1984)에 의하면 이분형 예측 유사성 측도에는 Goodman과 Kruscal의 $\lambda$ 측도, Anderberg의 $D$ 측도, Yule의 $Y$ 및 $Q$ 측도 등이 있다. 이들 중에서 $\lambda$와 $D$는 범위가 $[0, 1]$이므로 연관성의 방향을 파악하기가 곤란하므로 값의 범위가 $[-1, 1]$인 $Y$와 $Q$를 연관성 평가기준으로 제안하고자 한다. 이들을 표 2.1을 이용하여 나타내면 다음과 같다.

$$Y(A, B) = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}$$

$$Q(A, B) = \frac{ad - bc}{ad + bc}$$

여기서 Yule의 $Y$ 측도는 $2 \times 2$ 교차표에서 교차적비 (cross-product ratio)의 함수이며, 측도 $Q$는 Goodman과 Kruscal의 감마 측도를 $2 \times 2$ 교차표에서 나타낸 것으로 $Y$ 측도와 마찬가지로 교차적비의 함수형태로 표현된다. 위의 표 2.2와 표 2.3으로부터 Yule의 측도 $Y$를 구해보면 각각 -0.420과 -0.405로 나타나고 있어서 연관성 정도의 차이를 규명할 수 있는 동시에 두 값이 -1과 1사이의 값을 가지므로 기존의 두 측도보다 더 바람직한 연관성 평가기준이라고 할 수 있다. 이번에는 Yule의 $Q$ 측도를 계산해보면 각각 -0.714와 -0.695로 나타나고 있어서 이 또한 기존의 두 측도가 가지고 있는 단점을 보완할 수 있어서 기존에 제시된 두 측도보다 더 바람직한 연관성 평가기준이 된다고 할 수 있다.

## 3. 예제를 통한 유용성 고찰

본 절에서는 이분형 예측 유사성 측도들의 유용성을 예제를 통하여 고찰하고자 한다. 이를 위해 항목집합 $X$, $Y$에 대해 다음과 같이 가정하였다. 먼저 총 거래데이터의 수 ($t$)를 100명으로 하고, 항목 집합 $X$는 구매한 물품의 금액을 기준으로 특정금액 이상 (1) 구매한 사람 수와 특정금액 미만 (0)을 구매한 사람 수를 각각 50명으로 하였다. 또한 항목 집합 $Y$를 결제 방식을 기준으로 특정 방법으로 결제 (1) 한 사람 수를 30명으로 하고 그 외의 방법으로 결제 (0) 한 사람의 수를 70명으로 하였다. 항목 집합 $X$와 $Y$가 동시에 발생한 빈도 수, 즉 특정금액 이상의 물품을 구매하면서 특정방법으로 결제한 빈도수는 $a$명으로 하였다. 이를 정리하면 표 3.1과 같다. 이 표에서 $a$가 취할 수 있는 정수 값의 범위는 $0 \le a \le 30$이다.

**표 3.1** 모의실험 데이터 (1)

| | | $Y$ | | 합계 |
|---|---|---|---|---|
| | | 1 | 0 | |
| $X$ | 1 | $a$ | $50-a$ | 50 |
| | 0 | $30-a$ | $a+20$ | 50 |
| 합계 | | 30 | 70 | 100 |

표 3.1을 이용하여 동시발생빈도의 변화에 따라 이분형 예측 유사성 측도들과 지지도 및 신뢰도를 계산한 결과의 일부를 나타내면 표 3.2와 같다. 이 표로부터 알 수 있는 바와 같이 $a$의 값이 커질수록 모든 연관성 규칙 평가 기준들이 증가하고 있다. 그리고 순수 신뢰도와 본 논문에서 제안한 Yule의 $Y$ 및 $Q$ 측도와 비교해보았을 때, 값이 0이 되는 경우에서 많이 벗어나도 순수 신뢰도는 이들 두 측도에 비해 변하는 정도가 크지 않다. 또한 기여 순수 신뢰도는 값의 크기가 다른 측도들에 비해 매우 큰 값으로 나타나는 경우가 있어서 이를 연관성 평가 기준으로 활용할 때에는 상당한 주의가 요구된다. 이를 보다 구체적으로 알아보면, $a$=1, $b$=49, $c$=29, $d$=21인 경우에 지지도는 0.010, 신뢰도는 0.020으로 나타났으며, 이 때 신뢰도를 연관성 평가 기준으로 사용하게 되면 이 규칙은 의미 없는 규칙으로 분류될 수 있다. 그러나 기존의 순수 신뢰도나 기여 순수 신뢰도, 또는 본 논문에서 제안한 $Y$ 측도와 $Q$ 측도를 활용하는 경우에는 이 규칙은 상당히 의미 있는 것으로 판단하게 된다. 따라서 신뢰도 보다는 이들 측도들이 더 바람직하다고 할 수 있다. 또한 순수 신뢰도 -0.560 보다는 측도 $Y$와 $Q$의 값이 각각 -0.783과 -0.971로 계산되어서 -1에 더 가까우므로 $Y$ 및 $Q$ 측도가 더 바람직하다고 할 수 있으며, 기여 순수 신뢰도는 -28.000으로 상당히 큰 값으로 나타나고 있고, 그 값이 연관성 평가 기준으로서의 바람직한 측도가 되기 위한 범위인 $[-1, 1]$을 벗어남으로써 그 값 자체만으로는 연관성 정도를 파악하기가 곤란해진다. 또한 측도 $Y$와 $Q$를 비교해보면 -1의 값에 더 근접해있는 측도 $Q$가 더 바람직한 측도라고 할 수 있다.

기존의 순수 신뢰도 및 기여 순수 신뢰도와 본 논문에서 고려하는 측도들 간의 상관계수를 계산한 결과는 표 3.3과 같다 (** : 유의수준 0.01에서 유의함).

**표 3.2** 모의실험 데이터 (1)에 의한 연관성 평가 기준의 변화 양상

| $a$ | $b$ | $c$ | $d$ | $supp$ | $conf$ | $lift$ | $Nconf$ | $APconf$ | $Y$ | $Q$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 49 | 29 | 21 | 0.010 | 0.020 | 0.041 | -0.560 | -28.000 | -0.783 | -0.971 |
| 2 | 48 | 28 | 22 | 0.020 | 0.040 | 0.083 | -0.520 | -13.000 | -0.694 | -0.937 |
| 3 | 47 | 27 | 23 | 0.030 | 0.060 | 0.128 | -0.480 | -8.000 | -0.622 | -0.897 |
| 4 | 46 | 26 | 24 | 0.040 | 0.080 | 0.174 | -0.440 | -5.500 | -0.558 | -0.851 |
| 5 | 45 | 25 | 25 | 0.050 | 0.100 | 0.222 | -0.400 | -4.000 | -0.500 | -0.800 |
| 6 | 44 | 24 | 26 | 0.060 | 0.120 | 0.273 | -0.360 | -3.000 | -0.445 | -0.743 |
| 7 | 43 | 23 | 27 | 0.070 | 0.140 | 0.326 | -0.320 | -2.286 | -0.392 | -0.679 |
| 8 | 42 | 22 | 28 | 0.080 | 0.160 | 0.381 | -0.280 | -1.750 | -0.340 | -0.610 |
| 9 | 41 | 21 | 29 | 0.090 | 0.180 | 0.439 | -0.240 | -1.333 | -0.290 | -0.535 |
| 10 | 40 | 20 | 30 | 0.100 | 0.200 | 0.500 | -0.200 | -1.000 | -0.240 | -0.455 |
| 11 | 39 | 19 | 31 | 0.110 | 0.220 | 0.564 | -0.160 | -0.727 | -0.192 | -0.370 |
| 12 | 38 | 18 | 32 | 0.120 | 0.240 | 0.632 | -0.120 | -0.500 | -0.143 | -0.281 |
| 13 | 37 | 17 | 33 | 0.130 | 0.260 | 0.703 | -0.080 | -0.308 | -0.095 | -0.189 |
| 14 | 36 | 16 | 34 | 0.140 | 0.280 | 0.778 | -0.040 | -0.143 | -0.048 | -0.095 |
| 15 | 35 | 15 | 35 | 0.150 | 0.300 | 0.857 | 0.000 | 0.000 | 0.000 | 0.000 |
| 16 | 34 | 14 | 36 | 0.160 | 0.320 | 0.941 | 0.040 | 0.125 | 0.048 | 0.095 |
| 17 | 33 | 13 | 37 | 0.170 | 0.340 | 1.030 | 0.080 | 0.235 | 0.095 | 0.189 |

**표 3.3** 모의실험 데이터 (1)에 의한 연관성 평가 기준들 간의 상관계수

| | $Nconf$ | $APconf$ | $Y$ | $Q$ |
|---|---|---|---|---|
| $Nconf$ | 1 | .653** | .999** | .996** |
| $APconf$ | .653** | 1 | .674** | .618** |
| $Y$ | .999** | .674** | 1 | .991** |
| $Q$ | .996** | .618** | .991** | 1 |

이 표에서 보는 바와 같이 순수 신뢰도 및 기여 순수 신뢰도, Yule의 $Y$ 측도와 $Q$ 측도는 모두 상관관계가 매우 유의한 것으로 나타났다. 이로부터 알 수 있는 사실은 본 논문에서 고려한 $Y$ 측도와 $Q$ 측도는 기존의 순수 신뢰도와 기여 순수 신뢰도와 같이 연관성 평가 기준으로 사용할 수 있는 동시에 위에서 기술한 바와 같이 기존의 측도들이 가지고 있는 단점을 모두 해소하였으므로 $Y$ 및 $Q$ 측도가 더 바람직한 연관성 평가 기준이라고 할 수 있다.

본 논문에서 제안한 Yule의 $Y$ 및 $Q$ 측도의 유용성을 좀 더 살펴보기 위해 이번에는 표 3.4와 같이 $c$의 값의 변화함에 따라 각각의 측도들을 계산하여 그 결과를 표 3.5에 나타내었다. 이 표로부터 알 수 있는 바와 같이 $c$의 값이 커질수록 모든 연관성 측도들은 감소하고 있다. 또한 본 논문에서 제안한 측도인 $Y$ 및 $Q$가 취할 수 있는 값의 범위가 $[-1, 1]$로 나타난 반면에 기여 순수 신뢰도는 이 범위를 초과하는 경우가 많이 나타나고 있다.

**표 3.4** 모의실험 데이터 (2)

| | | $Y$ | | 합계 |
|---|---|---|---|---|
| | | 1 | 0 | |
| $X$ | 1 | $30-c$ | $50+c$ | 80 |
| | 0 | $c$ | $20-c$ | 20 |
| 합계 | | 30 | 70 | 100 |

위의 표 3.2에서와 마찬가지로 순수 신뢰도는 $Y$ 및 $Q$ 측도와 비교해보았을 때, 중간지점으로부터 벗어날수록 순수 신뢰도는 이들 두 측도에 비해 변하는 폭이 크지 않는 동시에 그 범위가 두 측도에 비해 좁게 나타나고 있다. 기여 순수 신뢰도는 표 3.2에 비해서는 작게 나타나고는 있으나 이 또한 값의 범위

가 $[-1, 1]$을 초과하여 나타나고 있어서 기여 순수 신뢰도를 연관성 평가 기준으로 활용할 때에는 주의하여야 한다.

**표 3.5** 모의실험 데이터 (2)에 의한 연관성 평가 기준의 변화 양상

| $a$ | $b$ | $c$ | $d$ | $supp$ | $conf$ | $lift$ | $Nconf$ | $APconf$ | $Y$ | $Q$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 30 | 50 | 0 | 20 | 0.300 | 0.375 | 0.750 | 0.375 | 1.000 | 1.000 | 1.000 |
| 29 | 51 | 1 | 19 | 0.290 | 0.363 | 0.711 | 0.313 | 0.862 | 0.533 | 0.831 |
| 28 | 52 | 2 | 18 | 0.280 | 0.350 | 0.673 | 0.250 | 0.714 | 0.375 | 0.658 |
| 27 | 53 | 3 | 17 | 0.270 | 0.338 | 0.637 | 0.188 | 0.556 | 0.259 | 0.485 |
| 26 | 54 | 4 | 16 | 0.260 | 0.325 | 0.602 | 0.125 | 0.385 | 0.162 | 0.316 |
| 25 | 55 | 5 | 15 | 0.250 | 0.313 | 0.568 | 0.063 | 0.200 | 0.077 | 0.154 |
| 24 | 56 | 6 | 14 | 0.240 | 0.300 | 0.536 | 0.000 | 0.000 | 0.000 | 0.000 |
| 23 | 57 | 7 | 13 | 0.230 | 0.288 | 0.504 | -0.063 | -0.217 | -0.072 | -0.143 |
| 22 | 58 | 8 | 12 | 0.220 | 0.275 | 0.474 | -0.125 | -0.455 | -0.140 | -0.275 |
| 21 | 59 | 9 | 11 | 0.210 | 0.263 | 0.445 | -0.188 | -0.714 | -0.205 | -0.394 |
| 20 | 60 | 10 | 10 | 0.200 | 0.250 | 0.417 | -0.250 | -1.000 | -0.268 | -0.500 |
| 19 | 61 | 11 | 9 | 0.190 | 0.238 | 0.389 | -0.313 | -1.316 | -0.329 | -0.594 |
| 18 | 62 | 12 | 8 | 0.180 | 0.225 | 0.363 | -0.375 | -1.667 | -0.389 | -0.676 |
| 17 | 62 | 13 | 7 | 0.170 | 0.215 | 0.347 | -0.435 | -2.021 | -0.445 | -0.743 |
| 16 | 64 | 14 | 6 | 0.160 | 0.200 | 0.313 | -0.500 | -2.500 | -0.507 | -0.806 |
| 15 | 65 | 15 | 5 | 0.150 | 0.188 | 0.288 | -0.563 | -3.000 | -0.566 | -0.857 |
| 14 | 66 | 16 | 4 | 0.140 | 0.175 | 0.265 | -0.625 | -3.571 | -0.626 | -0.899 |
| 13 | 67 | 17 | 3 | 0.130 | 0.163 | 0.243 | -0.688 | -4.231 | -0.688 | -0.934 |
| 12 | 68 | 18 | 2 | 0.120 | 0.150 | 0.221 | -0.750 | -5.000 | -0.754 | -0.962 |
| 11 | 69 | 19 | 1 | 0.110 | 0.138 | 0.199 | -0.813 | -5.909 | -0.832 | -0.983 |
| 10 | 70 | 20 | 0 | 0.100 | 0.125 | 0.179 | -0.875 | -7.000 | -1.000 | -1.000 |

이를 좀 더 구체적으로 알아보면, $a$=12, $b$=68, $c$=18, $d$=2인 경우와 $a$=11, $b$=69, $c$=19, $d$=1인 경우에 신뢰도가 각각 0.150과 0.138로 계산되어서 연관성 평가 기준으로 신뢰도를 이용하게 되면 두 규칙 모두 의미 없는 것으로 판단하게 된다. 반면에 $Y$ 측도와 $Q$ 측도를 활용하는 경우에는 이 두 규칙은 상당히 의미 있는 것으로 판단하게 된다. 또한 이 두 경우에도 기여 순수 신뢰도는 각각 -5.000과 -5.909로 나타나서 바람직한 측도의 범위인 $[-1, 1]$을 벗어나게 되어 연관성 정도를 파악하기가 곤란해진다. 순수 신뢰도는 각각 -0.750과 -0.813으로 계산되어서 측도 $Y$와 $Q$에 비해 -1에 덜 가까우므로 $Y$ 및 $Q$ 측도가 더 바람직하다고 할 수 있다. 측도 $Y$와 $Q$ 중에서는 -1의 값에 더 근접해있고 각 케이스별로 값이 더 확연하게 차이가 나는 측도 $Q$가 더 바람직한 측도라고 할 수 있다.

기존의 순수 신뢰도 및 기여 순수 신뢰도와 본 논문에서 고려하는 측도들 간의 상관계수를 계산한 결과는 표 3.6과 같다 (** : 유의수준 0.01에서 유의함). 위의 표 3.3과 마찬가지로 이 표에서도 순수 신뢰도 및 기여 순수 신뢰도, Yule의 $Y$ 측도와 $Q$ 측도는 모두 상관관계가 매우 유의한 것으로 나타났다.

**표 3.6** 모의실험 데이터 (2)에 의한 연관성 평가 기준들 간의 상관계수

| | $Nconf$ | $APconf$ | $Y$ | $Q$ |
|---|---|---|---|---|
| $Nconf$ | 1 | .957** | .972** | .965** |
| $APconf$ | .957** | 1 | .900** | .850** |
| $Y$ | .972** | .900** | 1 | .974** |
| $Q$ | .965** | .850** | .974** | 1 |

이 외에도 $b$ 및 $d$의 값이 변화함에 따른 순수 신뢰도 및 기여 순수 신뢰도와 본 논문에서 고려한 유사성 측도들에 대해서도 살펴보았는데 이와 유사한 결과를 얻을 수 있었다.

## 4. 결론

방대한 양의 데이터에 내재되어 있는 항목들 간의 유용한 관련성을 찾아내는 데 활용되고 있는 연관성 규칙에서 사용되는 데이터 형태는 거래 발생시점에서 기록된 정보만으로 구성되어 있다. 일반적으로 의미 있는 연관성 규칙을 생성하기 위한 평가 기준으로 신뢰도가 가장 많이 활용되고 있으나 이 값만으로는 연관성의 방향을 알 수가 없다. 이를 위해 순수 신뢰도와 기여 순수 신뢰도가 제안된 바 있다. 그러나 순수 신뢰도는 양의 신뢰도와 음의 신뢰도의 값의 차이가 동일하면 순수 신뢰도의 값이 동일하게 되는 단점을 가지고 있으며, 기여 순수 신뢰도는 범위가 제한되어 있지 않아서 그 값만으로는 연관성 정도를 파악하기가 곤란하다는 단점이 있다.

본 논문에서는 기존의 순수 신뢰도와 기여 순수 신뢰도의 단점을 보완하기 위해 이분형 예측 유사성 측도 중에서 Yule의 $Y$ 및 $Q$ 측도를 연관성 평가 기준으로 제안하였다. 교차표를 구성하고 있는 각 항의 여러 가지 값을 대입한 예제를 통해 고찰한 결과, 양의 신뢰도와 음의 신뢰도가 같은 값을 가지더라도 Yule의 측도 $Y$ 및 $Q$는 값이 달라지므로 순수 신뢰도의 단점을 보완한 측도라고 할 수 있으며, 기여 순수 신뢰도는 하한의 범위에는 제한이 없으나 $Y$ 및 $Q$ 측도는 항상 -1과 1 사이의 값을 가지므로 기여 순수 신뢰도보다 더 바람직한 측도라고 할 수 있다. 그리고 기존의 순수 신뢰도 및 기여 순수 신뢰도와 본 논문에서 고려하는 측도들 간의 상관계수가 매우 유의한 것으로 나타났으므로 본 논문에서 고려한 $Y$ 측도와 $Q$ 측도는 기존의 순수 신뢰도와 기여 순수 신뢰도와 같이 연관성 평가 기준으로 사용할 수 있다. 동시에 기존의 측도들이 가지고 있는 단점을 모두 보완하였으므로 $Y$ 및 $Q$ 측도가 더 바람직한 연관성 평가 기준이 된다. 또한 측도 $Y$와 $Q$ 중에서는 측도 $Q$가 더 바람직한 측도인 것을 모의실험의 결과에서 확인할 수 있었다. 그럼에도 불구하고 측도 $Y$는 측도 $Q$ 보다 더 보수적인 연관성 측도인 동시에 실제적 의미는 거의 찾기 어려운 이론적인 측도이므로 향후에는 보다 실제적인 유사성 측도에 대해 연관성 평가 기준을 적용하는 방안에 대해 연구가 필요할 것으로 사료된다.

## 참고문헌


안광일, 김성집 (2003). 연관규칙 탐색에서의 새로운 흥미도 척도의 제안. <대한산업공학회지>, **29**, 41-48.

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.

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. (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., Pei, J. and Yin, Y. (2000). Mining frequent patterns without candidate generation. *Proceedings of ACM SIGMOD Conference on Management of Data*, 1-12.

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, H. C. (2011). The proposition of attributably pure confidence in association rule mining. *Journal of the Korean Data & Information Science Society*, **22**, 235-243.

Park, H. C. and Cho, K. H. (2005). Waste database analysis joined with local information using association rules. *Journal of the Korean Data Analysis Society*, **7**, 763-772.

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.

Romesburg, H. C. (1984). *Cluster analysis for researchers*, Lifetime Learning Publications, Belmont, California.

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

*Journal of the Korean Data & Information Science Society*
*2011,* **22**(3), 495–503

한국데이터정보과학회지

# The application for predictive similarity measures of binary data in association rule mining

Hee Chang Park[1]

[1]Department of Statistics, Changwon National University

Received 11 April 2011, revised 9 May 2011, accepted 19 May 2011

**Abstract**

The most widely used data mining technique is to find association rules. Association rule mining is the method to quantify the relationship between each set of items in very huge database based on the association thresholds. There are some basic association thresholds to explore meaningful association rules ; support, confidence, lift, etc. Among them, confidence is the most frequently used, but it has the drawback that it can not determine the direction of the association. The net confidence and the attributably pure confidence were developed to compensate for this drawback, but they have other drawbacks.In this paper we consider some predictive similarity measures for binary data in cluster analysis and multi-dimensional analysis as association threshold to compensate for these drawbacks. The comparative studies with net confidence, attributably pure confidence, and some predictive similarity measures are shown by numerical example.

*Keywords*: Association thresholds, attributably pure confidence, confidence, net confidence, similarity measure.

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