확률비를 리용한 련관규칙의 믿음도에 대한 평가

공혜옥, 배철진

위대한 령도자 **김정일**동지께서는 다음과 같이 교시하시였다.

《과학과 기술이 매우 빨리 발전하고있는 오늘의 현실은 기초과학을 발전시킬것을 더욱 절실하게 요구하고있습니다.》(《**김정일**선집》 증보판 제11권 138페지)

최근 대규모자료에 대한 분석수법인 자료발굴모형에 대한 연구가 활발히 진행되고 있다.

확률론과 통계학은 오래전부터 불확정성을 측정하는 많은 응용령역들에서 광범히 리용된 수법들이다. 론문에서는 이 수법들을 리용하여 련관규칙의 불확정성인자의 하나인 믿음도를 평가하고 그에 기초하여 자료기지에서 흥미있는 련관규칙들을 생성하는 확률적 모형을 제기하였다.

1. 선행연구결과

선행연구[1, 2]에서는 Piatetsky-Shapiro주장을 리용하여 항목모임의 흥미척도를 규정하고 그것에 기초하여 지수적크기의 항목모임공간에서 흥미있는 항목모임들만을 주목하도록 함으로써 탐색공간의 크기를 크게 감소시킬수 있다는것을 고찰하였다.

련관규칙 $X \to Y$ 의 지지도를 $sprt(X \to Y)$ [4], 믿음도를 $conf(X \to Y)$ 로 표시할 때 확률론에 기초하여 다음과 같이 해석할수 있다.

$$sprt(X \to Y) = sprt(X \cup Y) = p(X \cup Y)$$
$$conf(X \to Y) = sprt(Y \mid X) = p(Y \mid X) = \frac{p(X \cup Y)}{p(X)} \tag{1}$$

그러면 Piatetsky-Shapiro주장은 다음과 같이 표시할수 있다.

$$p(X \cup Y) \approx p(X)p(Y)$$

이로부터 흥미척도[3]는 다음과 같다.

$$Interest(X,\ Y) = \frac{sprt(X \cup Y)}{sprt(X)sprt(Y)} \tag{2}$$

식 (2)는 조건부확률 $p(Y \mid X)$ 와 확률 $p(X)$ 사이의 관계 즉 모임 X 와 Y 의 종속성에 대한 통계적정의를 만족시킨다.

론문에서는 식 (1)과 식 (2)에 대한 확률적의미를 분석하여 흥미있는 련관규칙들의 가능한 경우를 평가한다.

2. 확 률 비

식 (2)의 값이 1로부터 멀어질수록 종속성은 더 커진다. 즉 $Interest(X,\ Y) = 1$ 이면 $p(Y \mid X) = p(Y)$ 이며 이것은 Y 와 X 가 독립이라는것을 의미한다. $Interest(X,\ Y) > 1$ 이면

$p(Y|X) > p(Y)$ 이며 이것은 Y 가 X 에 긍정적으로 종속 혹은 Y 가 발생하는 확률이 X 가 발생할 때 증가한다는것을 의미한다. 또한 $Interest(X, Y) < 1$ 이면 $p(Y|X) < p(Y)$ 이며 이것은 Y 가 X 에 부정적으로 종속 혹은 Y 가 발생하는 확률이 X 가 발생할 때 감소한다는것을 의미한다.

조건부확률 $p(Y|X)$ 와 확률 $p(X)$ 사이의 관계를 고찰함으로써 $Interest(X, Y)$ 를 다음과 같은 세가지 경우로 나눌수 있다.

① $Interest(X, Y) = 1$ 혹은 $p(Y|X) = p(Y)$ 이면 Y 와 X 가 독립

② $Interest(X, Y) > 1$ 혹은 $p(Y|X) > p(Y)$ 이면 Y 가 X 에 긍정적으로 종속이고 $p(Y|X) - p(Y)$ 는 다음의 식을 만족시킨다.

$$0 < p(Y|X) - p(Y) \le 1 - p(Y)$$

$$0 < \frac{p(Y|X) - p(Y)}{1 - p(Y)} \le 1 \tag{3}$$

그리고 이 비률이 클수록 긍정적인 종속성이 커진다.

③ $Interest(X, Y) < 1$ 혹은 $p(Y|X) < p(Y)$ 이면 Y 가 X 에 부정적으로 종속(혹은 $-Y$ 가 X 에 긍정적으로 종속)이고 $p(Y|X) - p(Y)$ 는 다음의 식을 만족시킨다.

$$0 > p(Y|X) - p(Y) \ge -p(Y)$$

$$0 < \frac{p(Y|X) - p(Y)}{-p(Y)} \le 1 \tag{4}$$

그리고 이 비률이 클수록 부정적인 종속성이 커진다.

식 (3), (4)는 $p(Y|X)$ 와 $p(Y)$ 의 관계를 반영하는 식이다. 그러므로 확률론의 확정성인자모형에 귀착시켜 이 식들을 고찰할수 있다. 확정성인자모형은 $p(Y|X)$ 와 $p(Y)$ 의 관계를 반영하는 우수한 모형이다. 론문에서는 그것을 아래에서와 같이 PR (확률비)로 표시한다.

$$PR(Y|X) = \begin{cases} \dfrac{p(Y|X) - p(Y)}{1 - p(Y)}, & p(Y|X) \ge p(Y),\ p(Y) \ne 1 \\ \dfrac{p(Y|X) - p(Y)}{p(Y)}, & p(Y) > p(Y|X),\ p(Y) \ne 0 \end{cases} \tag{5}$$

$p(Y|X) = p(X \cup Y) / p(X)$ 이므로 다음의 식이 성립한다.

$$PR(Y|X) = \begin{cases} \dfrac{p(X \cup Y) - p(X)p(Y)}{p(X)(1 - p(Y))}, & p(X \cup Y) \ge p(X)p(Y),\ p(X)(1 - p(Y)) \ne 0 \\ \dfrac{p(X \cup Y) - p(X)p(Y)}{p(X)p(Y)}, & p(X \cup Y) \le p(X)p(Y),\ p(X)p(Y) \ne 0 \end{cases}$$

웃식에 $sprt(X) = p(X),\ sprt(Y) = p(Y),\ sprt(X \cup Y) = p(X \cup Y)$ 를 대입하면

$$PR(Y|X) = \begin{cases} \dfrac{sprt(X \cup Y) - sprt(X)sprt(Y)}{sprt(X)(1 - sprt(Y))}, \\ \qquad sprt(X \cup Y) \ge sprt(X)sprt(Y),\ sprt(X)(1 - sprt(Y)) \ne 0 \\ \dfrac{sprt(X \cup Y) - sprt(X)sprt(Y)}{sprt(X)sprt(Y)}, \\ \qquad sprt(X \cup Y) < sprt(X)sprt(Y),\ sprt(X)sprt(Y) \ne 0 \end{cases} \tag{6}$$

이 성립한다. 이때 확률비 PR는 다음과 같은 성질을 가진다는것을 알수 있다.

$\Omega_Y=\{Y, \neg Y)\}$를 인식능력의 가설프레임, $X\subseteq\Omega_Y$를 관측에 의한 가능한 증거라고 가정하면 우선권확률에 대한 조건부확률의 증가비 PR는 다음의 식을 만족시킨다.

$$PR(Y\mid X)+PR(\neg Y\mid X)=0$$

우와 같은 고찰로부터 긍정 및 부정 련관규칙들을 둘 다 발견하고 측정하기 위하여 주어진 항목모임 X와 Y에 대한 련관규칙의 믿음도를 $PR(Y|X)$로 취한다.

식 (5), (6)으로부터

$$PR(Y\mid X)=\frac{p(Y\mid X)-p(Y)}{1-p(Y)}$$

혹은

$$PR(Y\mid X)=\frac{sprt(X\bigcup Y)-sprt(X)sprt(Y)}{sprt(X)(1-sprt(Y))}$$

$$(sprt(X\bigcup Y)\geq sprt(X)sprt(Y),\ sprt(X)(1-sprt(Y))\neq 0)$$

이다. 이 확률비를 믿음도로 취하면서 흥미있는 련관규칙에 대한 정의를 다음과 같이 할 수 있다.

정의 I를 자료기지 D안의 항목들의 모임이라고 하자. $S=X\bigcup Y\subseteq I$를 항목모임, $X\bigcap Y=\varnothing$, $sprt(X)\neq 0$, $sprt(Y)\neq 0$이라고 하고 $min.sprt$(최소지지도), $min.conf$(최소믿음도), $min.interest$(최소흥미도)>0이 사용자 혹은 전문가에 의해 주어진다고 하자. 그러면 다음과 같은 경우들에 흥미있는 련관규칙이 추출된다.

① $sprt(X\bigcup Y)\geq min.sprt$, $sprt(X\bigcup Y)-sprt(X)sprt(Y)\geq min.interest$, $PR(Y\mid X)\geq min.conf$이면 $X\rightarrow Y$는 흥미있는 련관규칙이다.

② $sprt(X\bigcup\neg Y)\geq min.sprt$, $sprt(X)\geq min.sprt$, $sprt(Y)\geq min.sprt$, $sprt(X\bigcup\neg Y)-sprt(X)$ $sprt(\neg Y)\geq min.interest$, $PR(\neg Y\mid X)\geq min.conf$이면 $X\rightarrow\neg Y$는 흥미있는 련관규칙이다.

③ $sprt(\neg X\bigcup Y)\geq min.sprt$, $sprt(X)\geq min.sprt$, $sprt(Y)\geq min.sprt$, $sprt(\neg X\bigcup Y)-sprt(\neg X)$ $sprt(Y)\geq min.interest$, $PR(Y\mid\neg X)\geq min.conf$이면 $\neg X\rightarrow Y$는 흥미있는 련관규칙이다.

④ $min.sprt$, $sprt(X)\geq min.sprt$, $sprt(Y)\geq min.sprt$, $sprt(\neg X\bigcup\neg Y)-sprt(\neg X)$ $sprt(\neg Y)\geq$ $min.interest$, $PR(\neg Y\mid\neg X)\geq min.conf$이면 $\neg X\rightarrow\neg Y$는 흥미있는 련관규칙이다.

이 정의에서는 네가지 종류의 흥미있는 유효한 련관규칙들을 보여준다. 경우 ①만이 흥미있는 긍정련관규칙을 정의하고 경우 ②－④는 흥미있는 부정련관규칙들을 정의한다. 여기서 $sprt(*)\geq min.sprt$는 련관규칙들이 빈발항목모임들사이의 관계를 묘사하도록 담보하고 $sprt(X\bigcup Y)-sprt(X)sprt(Y)\geq min.interest$는 련관규칙들이 흥미있다는것을 의미하며 $PR(*)\geq min.conf$는 련관규칙들이 유효하고 믿을수 있다는 조건을 의미한다.

이처럼 정의는 확률비를 리용하여 믿음도를 평가하는 관점에서 련관규칙에 대한 정의를 준다.

3. 흥미있는 련관규칙들의 생성

PR모형에서 련관규칙들을 발굴하는 과제는 흥미있는 정 및 부의 련관규칙들을 모두 발견하는것이다. 이 과제는 사실상 다음과 같은 2개의 분과제로 갈라진다.

1) 모든 긍정빈발항목모임들의 족 PS와 모든 부정빈발항목모임들의 족 NS를 생성하는것

2) PS안에서 $A \rightarrow B$ 혹은 $B \rightarrow A$형태의 모든 규칙들과 NS안에서 $\neg A \rightarrow B$ 혹은 $B \rightarrow \neg A$ 혹은 $\neg A \rightarrow \neg B$ 혹은 $\neg B \rightarrow \neg A$형태의 모든 규칙들을 생성하는것

1)에 대한 분과제는 이미 흥미있는 항목모임의 탐색알고리듬(Searching of Interesting Itemsets)을 개발함으로써 수행된다. 2)에 대한 분과제를 수행하기 위하여 다음과 같은 정 및 부의 련관규칙생성알고리듬(Generation of Positive and Negative Association rules)을 개발한다.

알고리듬: 긍정 및 부정련관규칙생성알고리듬
입력자료: D(자료기지)
min.sprt(최소지지도), *min.conf*(최소믿음도), *min.interest*(최소흥미도)
PS(흥미있는 긍정항목모임족), NS(흥미있는 부정항목모임족)
출력자료: 련관규칙 $X \rightarrow Y$
(1) 알고리듬 1을 호출 // 흥미있는 항목모임족을 얻는다.
// PS안의 모든 긍정련관규칙들을 생성
(2) for do $\forall A \in PS$ do {
for $\forall X \cup Y = A$ and $X \cap Y = \varnothing$ do {
if $sprt(X \cup Y) - sprt(X) \geq min.interest$ then
if $PR(Y|X) \geq min.conf$ then
output "$X \rightarrow Y$", $conf(X \rightarrow Y)$, $sprt(A)$;
if $PR(X|Y) \geq min.conf$ then
output "$Y \rightarrow X$", $conf(Y \rightarrow X)$, $sprt(A)$;
}
}
// NS안의 모든 부정련관규칙들을 생성
(3) for do $\forall A \in NS$ do {
for $\forall X \cup Y = A$ and $X \cap Y = \varnothing$ do {
// $\neg X \rightarrow Y$ 혹은 $Y \rightarrow \neg X$형태의 부정련관규칙들을 생성
if $sprt(X) \geq min.sprt$ and $sprt(Y) \geq min.sprt$ and $sprt(\neg X \cup Y) \geq min.sprt$ then
if $sprt(\neg X \cup Y) - sprt(\neg X) \geq min.interest$ then
if $PR(Y|\neg X) \geq min.conf$ then
output "$\neg X \rightarrow Y$", $conf(\neg X \rightarrow Y)$, $sprt(\neg X \cup Y)$;
if $PR(\neg X|Y) \geq min.conf$ then
output "$Y \rightarrow \neg X$", $conf(Y \rightarrow \neg X)$, $sprt(Y \cup \neg X)$;
// $\neg X \rightarrow \neg Y$ 혹은 $\neg Y \rightarrow \neg X$형태의 부정련관규칙들을 생성
if $sprt(X) \geq min.sprt$ and $sprt(Y) \geq min.sprt$ and $sprt(\neg X \cup \neg Y) \geq min.sprt$ then
if $sprt(\neg X \cup \neg Y) - sprt(\neg X) sprt(\neg Y) \geq min.interest$ then
if $PR(\neg Y|\neg X) \geq min.conf$ then
output "$\neg X \rightarrow \neg Y$", $conf(\neg X \rightarrow \neg Y)$, $sprt(\neg X \cup \neg Y)$;
if $PR(\neg X|\neg Y) \geq min.conf$ then
output "$\neg Y \rightarrow \neg X$", $conf(\neg Y \rightarrow \neg X)$, $sprt(\neg Y \cup \neg X)$;
}

}

긍정 및 부정 련관규칙생성알고리듬(Generation of Positive and Negative Association rules)은 *PS*안의 모든 긍정련관규칙들뿐아니라 *NS*안의 모든 부정련관규칙들도 생성한다.

알고리듬에서는 먼저 자료기지 *D*안에 있는 모든 긍정 및 부정 흥미있는 항목모임들의 족 *PS*와 *NS*를 얻기 위하여 부분틀 Searching of Interesting Itemsets이 호출된다. 다음으로 (2)에서는 모임 *PS*안에서 $X \rightarrow Y$ 혹은 $Y \rightarrow X$형태의 모든 흥미있는 긍정련관규칙들을 생성하여 출력한다. (3)에서는 모임 *NS*안에서 $\neg X \rightarrow Y$ 혹은 $Y \rightarrow \neg X$와 $\neg X \rightarrow \neg Y$ 혹은 $\neg Y \rightarrow \neg X$형태의 흥미있는 모든 부정련관규칙들을 생성하여 출력한다.

4. 결 론

전통적인 련관규칙발굴의 지지도－믿음도모형에서는 흥미있는 련관규칙들뿐아니라 흥미없는 항목모임들사이의 련관규칙도 발굴되였으며 더우기 흥미척도와 믿음도와의 관계속에서 긍정 및 부정 련관규칙의 식별을 할수 없었고 부의 련관규칙들의 종류도 밝히지 못하였다.

론문에서는 선행연구에서 정의한 항목모임의 흥미척도를 확률비와의 호상관계속에서 분석하고 그에 기초하여 흥미있는 련관규칙의 믿음도를 새롭게 규정하였다.

또한 련관규칙의 흥미척도를 두 항목모임사이의 종속성에 대한 통계적의미를 반영할수 있도록 확률비를 통하여 규정함으로써 흥미있는 련관규칙들의 가능한 경우 즉 련관규칙의 종류를 평가하였다. 그리고 이러한 수학적모형에 기초하여 흥미있는 긍정 및 부정 련관규칙의 모든 종류를 생성하는 알고리듬을 제기하였다.

참 고 문 헌

[1] Kong Hye Ok et al.; IJTPC., 12, 12, 13, 2016.

[2] Kong Hye Ok et al.; IJTPC., 10, 12, 1, 2015.

[3] Long Zhao et al.; Journal of Mathematics Science and Technology Education, 13, 8, 5577, 2017.

[4] G. Piatetsky-Shapiro; Discovery, Analysis, and Presentation of Strong Rules, AAAI Press/MIT Press, 229～248, 1991.

주체109(2020)년 3월 15일 원고접수

Estimation about Confidence of Association Rules Using Probability Ratio

Kong Hye Ok, Pae Chol Jin

Probability theory and statistics are widely used in applications for measure of uncertainty for a long time. In this paper we estimate confidence of association rules which is one kind of uncertainties using their methods and provide a probabilistic model to generate interesting association rules based on it.

Keywords: measure of interest, confidence