Journal of the Korean Data & Information Science Society
2010, **2**(3), 397–405

한국데이터정보과학회지

향상도 영향 감소화에 의한 연관성 순위결정함수

박희창[1]

[1]창원대학교 통계학과

접수 2010년 3월 15일, 수정 2010년 4월 19일, 게재확정 2010년 4월 28일

요 약

데이터 마이닝은 대규모의 데이터베이스에 내재되어 있는 유용한 정보를 찾아내는 과정이며, 중요한 목표 중의 하나는 여러 변수들 간의 관계를 발견하고 결정하는 것이다. 이를 위해 필요한 기법인 연관성 규칙 마이닝은 각 항목들 간의 관련성을 찾아내는 데 활용되며, 지지도, 신뢰도, 향상도 등의 연관성 측도를 기반으로 두 항목간의 관계를 수치화함으로써 의미 있는 규칙을 찾아낸다. 본 논문에서는 3개의 연관기준값들 중 어느 하나라도 기준 이상이 되는 규칙의 순위를 매겨 필요한 연관성 규칙만을 생성할 수 있는 연관성 순위 결정 함수를 개발하는데 기존의 연구 결과를 개선하기 위해 특정 연관 기준값의 영향을 더 많이 받지 않도록 3개 연관기준값의 범위를 조정한 연관성 순위 결정 함수를 제안하고자 한다. 모의실험을 해본 결과, 대체적으로 본 논문에서 제안한 함수는 연관성 측도들과 최저 연관기준값들간의 차이를 잘 반영하고 있으며, 최저 연관성 기준값들의 범위와는 관계없이 항상 -1과 1 사이의 값을 가지며, 최저 연관기준값을 모두 충족하게 되면 1의 값을 가지며, 3개 모두 충족되지 않으면 -1의 값을 갖게 된다는 사실을 알 수 있었다.

주요용어: 데이터마이닝, 신뢰도, 연관성 규칙, 연관성 의사 결정 함수, 지지도, 향상도.

1. 서론

데이터의 양이 기하급수적으로 증가하고 있는 오늘날 조직의 최적 전략이나 의사결정을 뒷받침해 줄 수 있는 고급정보가 필요하게 되면서 데이터마이닝 (data mining) 기법이 등장하게 되었다. 데이터마이닝은 방대한 양의 데이터 속에서 내재되어 있는 유용한 정보를 찾아내는 과정으로, 대용량의 관측 가능한 데이터를 기반으로 숨겨진 지식, 기대하지 못했던 패턴, 새로운 법칙과 관계를 발견하고 이를 바탕으로 의사결정 등을 위한 정보로 활용하고자 하는 것이다.

데이터마이닝이 적용되는 과정은 탐색 (exploration)을 통해 평균, 이상치, 결측치 등을 발견하고 변형 (modification)으로 자료를 변환하며, 모형화 (modeling)와 모델평가 (assessment)의 단계를 거치게 된다. 대표적인 데이터마이닝 기법으로는 연관성규칙 (association rule), 의사결정나무 (decision tree) 기법, k-평균 군집방법, 신경망모형 (neural network) 등의 분석 기법이 있다.

데이터마이닝 기법 중에서 가장 많이 활용되고 있는 연관성 규칙 (association rule)은 대용량 데이터베이스에서 각 항목들 간의 관련성을 찾아내는 기법으로 여러 가지 연관기준값을 바탕으로 관련성 여부를 측정한다. 이러한 연관성 규칙은 Agrawal 등 (1993)에 의해 처음 소개된 이후, 많은 학자들에 의해 연관성 규칙의 생성에 관한 연구가 수행되었다 (Agrawal과 Srikant, 1994; Park 등, 1995; Srikant와 Agrawal, 1995; Toivonen, 1996; Bayardo, 1998; Cai 등, 1998; Han과 Fu, 1999; Liu 등, 1999; Pasquier

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

등, 1999; Han 등, 2000; Pei 등, 2000; Cho와 Park, 2007; Cho와 Park, 2008; Choi와 Park, 2008; Park, 2008).

의미 있는 연관성 규칙을 탐색하기 위한 가장 기본적인 연관기준값에는 지지도 (support), 신뢰도 (confidence), 향상도 (lift) 등이 있다. 일반적으로 연관성 규칙 생성과정은 첫 번째 단계에서 사용자가 지정한 최소 지지도를 만족시키는 빈발항목집합을 생성한 후, 두 번째 단계에서 빈발항목집합을 이용하여 최저 신뢰도 기준을 만족하고 향상도가 1이상인 것을 규칙으로 채택하게 된다. 이 때 3가지 기준을 모두 충족하는 경우에는 당연히 이 규칙은 연관성이 있는 것으로 간주할 수 있다. 또한 3가지 기준을 모두 충족하지 못하는 경우에는 당연히 이 규칙은 연관성이 없는 것으로 판단한다. 또한 생성된 규칙들 중에서 지지도의 기준값에는 미치지 못하지만 신뢰도의 값이 상당히 큰 경우나 향상도의 값이 1 보다 많이 큰 경우, 향상도가 1보다 작아서 음의 연관 정도가 양의 연관 정도보다는 강하나 신뢰도 또는 지지도가 높은 경우, 그리고 지지도의 값은 상당히 크지만 신뢰도의 값이 기준에 미치지 못하거나 향상도의 값이 1 보다 작은 경우에도 연관성이 없는 것으로 간주하게 된다. 따라서 연관성 규칙 생성과정에서 3가지 기준을 너무 높게 책정하면 이들 모두 충족하는 경우는 드물게 되어 의미 있는 연관성 규칙들이 발견되지 않을 수 있는 반면에 너무 낮게 잡으면 필요 이상으로 많은 규칙들이 생성될 수도 있다. 특히 발생이 빈번하지 않는 희귀한 사건인 경우에는 3가지 기준을 모두 충족하는 경우가 드물게 되므로 3개의 연관성 측도 중 어느 하나라도 기준 이상이 되는 규칙에 대해 순위를 매겨 필요한 연관성 규칙만을 생성할 수 있는 연관성 순위 결정 함수가 필요하다. Park (2010)은 3가지 기준값 모두가 충족되지 않는 경우의 연관성 규칙들을 서열화할 수 있는 연관성 순위 결정 함수를 개발하여 Wu 등 (2004)이 제안한 함수와 비교한 바 있다. 이 함수는 가장 기본적인 연관성 규칙 평가 기준인 지지도, 신뢰도, 향상도를 조합한 것으로 특정 연관기준값에 크게 영향을 받게 되는 동시에 몇 가지 문제점을 안고 있다.

따라서 본 논문에서는 특정 연관 기준값의 영향을 받지 않도록 3개 연관기준값의 범위를 조정한 연관성 순위 결정 함수를 제안하고자 한다. 본 논문의 2절에서는 향상도 영향의 감소에 의한 연관성 순위 결정 함수의 제시한 후, 3절에서는 구체적인 예제를 통하여 여러 가지 연관성 순위 결정 함수들의 비교를 통해 본 논문에서 제시한 함수의 유용성 여부를 토의한 후, 마지막으로 4절에서 결론을 내리고자 한다.

2. 향상도 영향의 감소에 의한 연관성 순위 결정 함수의 개발

연관성 규칙을 평가하는 가장 기본적인 기준에는 지지도, 신뢰도, 향상도 등이 있다. 지지도 S(A⇒B)는 항목 집합 A와 항목 집합 B가 동시에 발생하는 거래량 (transaction)의 비율을 의미하며, 다음과 같이 정의된다.

$$S(A \Rightarrow B) = \text{A와 B를 동시에 구매하는 거래수/전체 거래수} = P(A \text{ and } B) \tag{2.1}$$

신뢰도 C(A⇒B)는 항목 집합 A가 포함된 거래 비율 중 항목 집합 A와 항목 집합 B가 동시에 포함된 거래의 비율을 의미하며, 다음과 같이 정의된다.

$$C(A \Rightarrow B) = P(B|A) \tag{2.2}$$

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, 다음과 같이 정의된다.

$$L(A \Rightarrow B) = \frac{P(B|A)}{P(B)} = \frac{P(A \text{ and } B)}{P(A)P(B)} \tag{2.3}$$

따라서 연관 규칙 마이닝에서는 향상도가 1이상이고, 최저 지지도를 만족하는 규칙들 중에서 최저 신뢰도 기준을 초과하는 경우에 일반적으로 연관성 규칙이 생성되는 것으로 간주한다. 만일 이 세 가지 모

든 조건을 만족하는 경우에는 연관성 규칙이 생성되는 것으로 간주할 수 있으나, 조건들이 강해지면 이들 조건을 만족하는 의미 있는 연관성 규칙은 기대 이상으로 줄어들게 되는 반면에, 너무 낮게 잡으면 필요 이상으로 많은 규칙이 생성될 수도 있다. 특히 희귀한 사건인 경우에는 3가지 기준을 모두 충족하는 경우가 흔하지 않다. 예를 들어 최저 신뢰도 기준과 최저 향상도 기준은 만족하나 최저 지지도 기준에는 미달되는 경우, 최저 지지도 기준과 최저 향상도 기준은 만족하나 최저 신뢰도 기준에는 미달되는 경우, 그리고 최저 신뢰도 기준은 만족하나 최저 신뢰도 기준과 최저 향상도 기준에는 미달되는 경우 등을 들 수 있다. 이러한 경우에는 3개의 연관성 측도 중 어느 하나라도 기준 이상이 되는 규칙의 순위를 매겨 필요한 연관성 규칙만을 생성할 수 있는 연관성 순위 결정 함수가 필요하다. 이를 위해 Park (2010)은 3가지 기준값 모두가 충족되지 않는 경우의 연관성 규칙들을 서열화할 수 있는 식 (2.4)와 같은 연관성 순위 결정 함수를 제안한 바 있다.

$$F_{Park} = \frac{supp(x,y) + conf(x,y) + lift(x,y) - (Min_s + Min_c + Min_l)}{|supp(x,y) - Min_s| + |conf(x,y) - Min_c| + |lift(x,y) - Min_l|} \tag{2.4}$$

이 함수는 가장 기본적인 연관성 규칙 평가 기준인 지지도, 신뢰도, 향상도를 조합한 것으로 Wu 등 (2004)이 연관성 규칙의 가지치기 전략을 위해 지지도, 신뢰도, 관심도 (interest; int)를 기반으로 제안한 다음의 함수를 변형한 것이다.

$$F_{Wu} = \frac{supp(x,y) + conf(x,y) + int(x,y) - (Min_s + Min_c + Min_i) + 1}{|supp(x,y) - Min_s| + |conf(x,y) - Min_c| + |int(x,y) - Min_i| + 1} \tag{2.5}$$

또한 Park (2010)이 식 (2.4)의 연관성 순위 결정 함수를 제시한 이유는 여러 가지가 있는데, 그 중에서도 먼저 관심도에 비해 향상도가 일반적으로 많이 사용되는 연관성 측도인 동시에 식 (2.5)에서 분모와 분자에 1을 더한 이유가 불분명하다는 것이다. 또한 F_{Park}은 3개의 연관성 측도들의 크기를 잘 반영하는 반면에 F_{Wu}의 값들은 연관성 측도들의 크기에 따라 다양한 값을 취하며, 최저 기준값의 변화에 따라 일관성을 보여주지 못한다는 사실을 예제를 통하여 확인한 바 있다. 그러나 연관성 순위 결정 함수를 개발할 때, 연관 기준값을 변형 없이 그대로 사용하게 되면 특정 연관기준값의 영향을 크게 받을 수 있다. 먼저 지지도와 신뢰도와 향상도의 범위를 비교해보면 지지도와 신뢰도에 비해 지지도는 변화의 폭이 매우 크다.

$$0 \le S(A \Rightarrow B) \le 1, 0 \le C(A \Rightarrow B) \le 1, 0 \le L(A \Rightarrow B) \le \infty$$

또한 정의에 의해 지지도와 신뢰도 및 향상도는 다음과 같은 대소 관계가 성립한다.

$$S(A \Rightarrow B) \le C(A \Rightarrow B) \le L(A \Rightarrow B)$$

신뢰도는 항목 A가 발생할 확률이 작아질수록 지지도와 차이가 더 커진다. 또한 지지도는 항목 A가 발생할 확률과 항목 B가 발생할 확률이 동시에 작아질수록 지지도와 차이가 더 커지며, 항목 B가 발생할 확률이 작아질수록 신뢰도와이 차이가 크게 난다. 이러한 문제점을 해결하기 위해 본 절에서는 지지도와 신뢰도에 비해 범위와 크기가 상당히 차이가 나는 향상도의 영향을 감소시키기 위해 향상도의 분모와 분자의 비가 아닌 차이를 이용한 측도를 고려하고자 한다. 다시 말해서 식 (2.4)의 의사결정함수에서 향상도의 분모와 분자의 차이를 적용하고 정리하면 다음과 같은 연관성 순위결정함수를 얻을 수 있다.

$$F = \frac{(supp(x,y) + 2 \cdot conf(x,y)) - (Min_s + Min_c + P(Y) \cdot Min_l)}{|supp(x,y) - Min_s| + |conf(x,y) - Min_c| + |conf(x,y) - P(Y) \cdot Min_l|} \tag{2.6}$$

3. 예제를 통한 고찰

본 절에서는 연관성 규칙에서의 3가지 기준값 모두가 충족되지 않는 연관성 규칙들을 서열화할 수 있는 3개의 연관성 순위 결정 함수에 대해 그 유용성을 예제에 의해 비교하고자 한다. 항목 집합 X, Y에 대해 다음과 같이 가정하였다. 먼저 데이터베이스에 있는 총 트랜잭션의 수 (t)를 50명으로 하고, 항목 집합 Y는 구매한 냉장고의 금액을 기준으로 100만원 이상 (1) 구매한 사람 수를 30명으로 하고 100만원 미만 (0)을 구매한 사람 수를 20명으로 하였다. 또한 항목 집합 X를 결제 방식을 기준으로 신용 카드로 결제 (1)한 사람 수를 (15+e+r)명으로 하고 신용 카드 이외의 방법으로 결제 (0)한 사람의 수를 (35-e-r)명으로 하였다. 항목 집합 X와 Y가 동시에 발생한 빈도 수, 즉 100만원 이상의 냉장고를 구매하면서 신용카드로 결제한 빈도수는 $(5+e)$명으로 하였다. 이를 정리하면 표 3.1과 같다.

표 3.1 모의실험 데이터

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

이 표에서 e 및 r 이 취할 수 있는 정수 값의 범위는 다음과 같다.

$$0 \le e \le 25, 0 \le r \le 10$$

이로부터 e 및 r의 변화에 따른 연관성 평가기준인 지지도, 신뢰도, 지지도와 3개의 연관성 순위 결정 함수들을 계산한 후, 보다 상세한 설명을 위해 결과를 분할하여 표 3.2, 표 3.3, 표 3.4에 제시하였다. 여기서 $a = n(X = 1, Y = 1)$, $b = n(X = 1, Y = 0)$, $c = n(X = 0, Y = 1)$, $d = n(X = 0, Y = 0)$을 의미하며, 최저 기준값을 $Min_s = 0.4$, $Min_c = 0.5$, 그리고 $Min_l = 1.1$으로 가정한다.

모의실험결과를 전체적으로 살펴보면, 대체적으로 본 논문에서 제안하는 함수 F는 연관성 측도들과 최저 연관성 기준값들간의 차이를 잘 반영하고 있으며, 최저 연관성 기준값들의 범위와는 관계없이 항상 -1과 1 사이의 값을 가지며, 3개의 연관성 기준값이 모두 충족되면 1의 값을 가지며, 3개 모두 충족되지 않으면 -1의 값을 갖게 된다. 함수 F_{Park}은 함수 F와 마찬가지로 최저 연관성 기준값들의 범위와는 관계없이 항상 -1과 1 사이의 값을 가지며, 3개의 연관성 기준값이 모두 충족되면 1, 그렇지 않으면 -1의 값을 갖게 되는 반면에 연관성 측도들과 최저 연관성 기준값들간의 차이를 잘 반영하지 못하고 있다. 한편 F_{Wu}는 최저 연관성 기준값이 어떤 값을 취하느냐에 따라 값의 범위가 달라지며, 방향성 없이 0과 1 사이의 값을 갖는다. 특히 $Min_i = 0$인 경우의 F_{Wu0}는 3개의 연관성 규칙 기준 중에서 어느 하나가 기준값보다 작더라도 1의 값을 취한다는 사실을 알 수 있다.

표 3.2는 연관성 기준값의 변화에 따라 연관성 순위 결정 함수들의 변화하는 정도를 살펴본 자료 중에서 F의 값이 큰 부분을 제시한 것이다. 여기서 F_{Wu0}, F_{Wu1}, F_{Wu2}는 각각 $Min_i = 0$, $Min_i = 0.1$, $Min_i = 0.2$일 때의 F_{Wu}를 의미한다. 이 표에서 보는 바와 같이 F와 F_{Park}은 3개의 연관성 기준값의 변화에 충실히 잘 반영해주고 있으나 F_{Wu0}는 모두 동일한 값을 나타내고 있다. 또한 F_{Wu1}과 F_{Wu2}는 3개의 연관성 규칙 기준 모두가 기준을 충족하더라도 1보다 작은 값을 취한다는 사실을 알 수 있다. 또한 F와 F_{Park}을 비교해보면 지지도와 신뢰도의 조건을 만족해도 향상도의 조건을 만족하지 못하면 F는 그 영향을 크게 받지 않으나 F_{Park}은 영향을 크게 받는다.

이러한 사실을 좀 더 구체적으로 살펴보기 위해 $a = 24$, $b = 13$, $c = 6$, $d = 7$인 경우와 $a = 27$, $b = 15$, $c = 3$, $d = 5$인 경우를 비교해보면, 각각의 경우에 지지도는 0.480과 0.540, 신뢰도는 0.649와

표 3.2 연관성 기준값의 변화에 따른 연관성 순위 결정 함수의 변화량 (1)

a	b	c	d	$P(X)$	$P(Y)$	$supp$	$conf$	$lift$	int	F_{Park}	F_{Wu0}	F_{Wu1}	F_{Wu2}	F
28	10	2	10	0.76	0.60	0.560	0.737	1.228	0.104	1.0000	1.0000	1.0000	0.8714	1.0000
29	15	1	5	0.88	0.60	0.580	0.659	1.098	0.052	0.9911	1.0000	0.9308	0.8010	0.9947
27	14	3	6	0.82	0.60	0.540	0.659	1.098	0.048	0.9838	1.0000	0.9230	0.7904	0.9902
25	13	5	7	0.76	0.60	0.500	0.658	1.096	0.044	0.9732	1.0000	0.9148	0.7793	0.9838
23	12	7	8	0.70	0.60	0.460	0.657	1.095	0.040	0.9571	1.0000	0.9060	0.7676	0.9740
21	11	9	9	0.64	0.60	0.420	0.656	1.094	0.036	0.9315	1.0000	0.8968	0.7553	0.9583
30	16	0	4	0.92	0.60	0.600	0.652	1.087	0.048	0.9286	1.0000	0.9259	0.7979	0.9565
28	15	2	5	0.86	0.60	0.560	0.651	1.085	0.044	0.9096	1.0000	0.9181	0.7873	0.9448
26	14	4	6	0.80	0.60	0.520	0.650	1.083	0.040	0.8837	1.0000	0.9098	0.7762	0.9286
29	16	1	4	0.90	0.60	0.580	0.644	1.074	0.040	0.8520	1.0000	0.9133	0.7844	0.9085
24	13	6	7	0.74	0.60	0.480	0.649	1.081	0.036	0.8472	1.0000	0.9010	0.7645	0.9054
27	15	3	5	0.84	0.60	0.540	0.643	1.071	0.036	0.8165	1.0000	0.9050	0.7733	0.8857
30	17	0	3	0.94	0.60	0.600	0.638	1.064	0.036	0.8068	1.0000	0.9087	0.7817	0.8794
22	12	8	8	0.68	0.60	0.440	0.647	1.078	0.032	0.7932	1.0000	0.8916	0.7520	0.8706
25	14	5	6	0.78	0.60	0.500	0.641	1.068	0.032	0.7680	1.0000	0.8961	0.7615	0.8540
28	16	2	4	0.88	0.60	0.560	0.636	1.061	0.032	0.7653	1.0000	0.9003	0.7705	0.8523
29	17	1	3	0.92	0.60	0.580	0.630	1.051	0.028	0.7260	1.0000	0.8958	0.7679	0.8261
26	15	4	5	0.82	0.60	0.520	0.634	1.057	0.028	0.7101	1.0000	0.8914	0.7588	0.8153
20	11	10	9	0.62	0.60	0.400	0.645	1.075	0.028	0.7089	1.0000	0.8817	0.7388	0.8145
23	13	7	7	0.72	0.60	0.460	0.639	1.065	0.028	0.6994	1.0000	0.8867	0.7491	0.8081
30	18	0	2	0.96	0.60	0.600	0.625	1.042	0.024	0.6957	1.0000	0.8915	0.7655	0.8056
27	16	3	4	0.86	0.60	0.540	0.628	1.047	0.024	0.6671	1.0000	0.8869	0.7562	0.7860
28	17	2	3	0.90	0.60	0.560	0.622	1.037	0.020	0.6352	1.0000	0.8825	0.7538	0.7639
24	14	6	6	0.76	0.60	0.480	0.632	1.053	0.024	0.6341	1.0000	0.8819	0.7463	0.7632
29	18	1	2	0.94	0.60	0.580	0.617	1.028	0.016	0.6114	1.0000	0.8784	0.7515	0.7472
21	12	9	8	0.66	0.60	0.420	0.636	1.061	0.024	0.5975	1.0000	0.8767	0.7358	0.7374
30	19	0	1	0.98	0.60	0.600	0.612	1.020	0.012	0.5937	1.0000	0.8743	0.7494	0.7347
25	15	5	5	0.80	0.60	0.500	0.625	1.042	0.020	0.5882	1.0000	0.8774	0.7438	0.7308
19	10	11	10	0.58	0.60	0.380	0.655	1.092	0.032	0.6939	0.9669	0.8584	0.7201	0.7241

0.643, 그리고 향상도는 1.081과 1.071로 계산이 되었으며, F_{Park}는 0.8472와 0.8165로, 그리고 F는 0.9054, 0.8857로 나타났다. 따라서 후자의 경우가 전자에 비해 지지도는 더 크고 신뢰도는 비슷하며, 향상도가 조금 줄어들었으나 F_{Park}값은 F값에 비해 현격하게 감소하는 것으로 보아 F_{Park}는 향상도의 영향을 크게 받는다는 것을 알 수 있다. 또한 $a = 25$, $b = 15$, $c = 5$, $d = 5$인 경우와 $a = 19$, $b = 10$, $c = 11$, $d = 10$인 경우를 비교해보면, 각각의 경우에 지지도는 0.500과 0.380, 신뢰도는 0.625와 0.655, 그리고 향상도는 1.042와 1.092로 계산이 되었으며, F_{Park}는 0.5882와 0.6939로, 그리고 F는 0.7308, 0.7241로 나타났다. 후자의 경우가 전자에 비해 지지도는 줄어들고, 신뢰도는 비슷하고 향상도가 조금 크나 F_{Park}값은 F값에 비해 오히려 증가한 것으로 보아 이 경우에도 F_{Park}는 향상도의 영향을 크게 받는다는 것을 알 수 있다.

표 3.3은 연관성 순위 결정 함수들의 변화하는 정도를 F의 값이 0이 되는 지점을 중심으로 나타낸 것이다. 이 표에서도 F은 3개의 연관성 측도들의 크기를 잘 반영하고 있으나 함수 F_{Wu}와 F_{Park}은 연관성 측도들의 크기에 따라 다양한 값을 취하고 있으며, 최저 기준값의 변화에 따라 일관성을 보여주지 못하고 있다. 또한 F은 0의 값을 기준으로 연관성 측도들의 크기에 따라 양의 값과 음의 값으로 표현되나 최저 기준값의 변화에 따른 F_{Wu}의 값들은 모두 양의 값으로만 나타나고 있어 방향성을 제시하지 못하고 있다. 또한 F_{Park}의 값들은 연관성 규칙 기준값의 변화에 따라 일관성을 보여주지 못하고 있다. 특히 $a = 17$, $b = 10$, $c = 13$, $d = 10$인 경우와 $a = 24$, $b = 19$, $c = 6$, $d = 1$인 경우를 비교해보면, 각각의 경우에 지지도는 0.340과 0.480, 신뢰도는 0.630와 0.558, 그리고 향상도는 1.049과 0.930로 계산이

표 3.3 연관성 기준값의 변화에 따른 연관성 순위 결정 함수의 변화량 (2)

a	b	c	d	$P(X)$	$P(Y)$	$supp$	$conf$	$lift$	int	F_{Park}	F_{Wu0}	F_{Wu1}	F_{Wu2}	F
17	10	13	10	0.54	0.60	0.340	0.630	1.049	0.016	0.0791	0.9005	0.7739	0.6447	0.1785
24	19	6	1	0.86	0.60	0.480	0.558	0.930	-0.036	-0.1027	0.9387	0.7865	0.6565	0.1512
21	15	9	5	0.72	0.60	0.420	0.583	0.972	-0.012	-0.1058	0.9785	0.8157	0.6776	0.1481
20	14	10	6	0.68	0.60	0.400	0.588	0.980	-0.008	-0.1509	0.9854	0.8194	0.6791	0.1029
23	18	7	2	0.82	0.60	0.460	0.561	0.935	-0.032	-0.1541	0.9445	0.7893	0.6571	0.0998
24	20	6	0	0.88	0.60	0.480	0.545	0.909	-0.048	-0.2069	0.9182	0.7676	0.6389	0.0455
19	13	11	7	0.64	0.60	0.380	0.594	0.990	-0.004	-0.1636	0.9571	0.7963	0.6600	0.0417
22	17	8	3	0.78	0.60	0.440	0.564	0.940	-0.028	-0.2111	0.9505	0.7922	0.6577	0.0410
18	12	12	8	0.60	0.60	0.360	0.600	1.000	0.000	-0.1667	0.9298	0.7742	0.6418	0.0000
23	19	7	1	0.84	0.60	0.460	0.548	0.913	-0.044	-0.2702	0.9236	0.7699	0.6390	-0.0216
17	11	13	9	0.56	0.60	0.340	0.607	1.012	0.004	-0.1604	0.8975	0.7530	0.6244	-0.0260
21	16	9	4	0.74	0.60	0.420	0.568	0.946	-0.024	-0.2752	0.9568	0.7953	0.6584	-0.0270
16	10	14	10	0.52	0.60	0.320	0.615	1.026	0.008	-0.1445	0.8670	0.7328	0.6079	-0.0385
22	18	8	2	0.80	0.60	0.440	0.550	0.917	-0.040	-0.3415	0.9292	0.7724	0.6391	-0.1000
20	15	10	5	0.70	0.60	0.400	0.571	0.952	-0.020	-0.3478	0.9634	0.7986	0.6593	-0.1071
23	20	7	0	0.86	0.60	0.460	0.535	0.891	-0.056	-0.3746	0.9027	0.7506	0.6210	-0.1374
19	14	11	6	0.66	0.60	0.380	0.576	0.960	-0.016	-0.3584	0.9352	0.7755	0.6402	-0.1582

되었으며, F_{Park}는 0.0791와 -0.1027, 그리고 F는 0.1785, 0.1512로 나타났다. 따라서 이 경우에도 전자에 비해 후자의 지지도가 더 크고, 신뢰도와 향상도가 조금 줄어들었으나 F_{Park} 값이 F 값에 비해 현격하게 줄어드는 것으로 보아 F_{Park}는 향상도의 영향을 크게 받는다는 것을 알 수 있다.

표 3.4 연관성 기준값의 변화에 따른 연관성 순위 결정 함수의 변화량 (3)

a	b	c	d	$P(X)$	$P(Y)$	$supp$	$conf$	$lift$	int	F_{Park}	F_{Wu0}	F_{Wu1}	F_{Wu2}	F
17	14	13	6	0.62	0.60	0.340	0.548	0.914	-0.032	-0.6713	0.8387	0.6904	0.5643	-0.5601
13	10	17	10	0.46	0.60	0.260	0.565	0.942	-0.016	-0.6409	0.7445	0.6125	0.4990	-0.5652
16	13	14	7	0.58	0.60	0.320	0.552	0.920	-0.028	-0.6686	0.8137	0.6698	0.5470	-0.5690
14	11	16	9	0.50	0.60	0.280	0.560	0.933	-0.020	-0.6538	0.7667	0.6308	0.5143	-0.5714
15	12	15	8	0.54	0.60	0.300	0.556	0.926	-0.024	-0.6629	0.7898	0.6499	0.5303	-0.5726
21	20	9	0	0.82	0.60	0.420	0.512	0.854	-0.072	-0.7688	0.8696	0.7143	0.5829	-0.6423
20	18	10	2	0.76	0.60	0.400	0.526	0.877	-0.056	-0.7887	0.8965	0.7361	0.6007	-0.6711
19	17	11	3	0.72	0.60	0.380	0.528	0.880	-0.052	-0.7928	0.8691	0.7133	0.5815	-0.6914
18	16	12	4	0.68	0.60	0.360	0.529	0.882	-0.048	-0.7951	0.8425	0.6911	0.5628	-0.7059
17	15	13	5	0.64	0.60	0.340	0.531	0.885	-0.044	-0.7956	0.8168	0.6697	0.5447	-0.7159
12	10	18	10	0.44	0.60	0.240	0.545	0.909	-0.024	-0.7706	0.7007	0.5728	0.4627	-0.7159
16	14	14	6	0.60	0.60	0.320	0.533	0.889	-0.040	-0.7945	0.7919	0.6489	0.5271	-0.7222
13	11	17	9	0.48	0.60	0.260	0.542	0.903	-0.028	-0.7801	0.7222	0.5907	0.4779	-0.7222
15	13	15	7	0.56	0.60	0.300	0.536	0.893	-0.036	-0.7917	0.7679	0.6288	0.5101	-0.7253
14	12	16	8	0.52	0.60	0.280	0.538	0.897	-0.032	-0.7869	0.7446	0.6094	0.4937	-0.7253
20	19	10	1	0.78	0.60	0.400	0.513	0.855	-0.068	-0.9007	0.8742	0.7155	0.5815	-0.8397
19	18	11	2	0.74	0.60	0.380	0.514	0.856	-0.064	-0.9027	0.8469	0.6927	0.5622	-0.8498
18	17	12	3	0.70	0.60	0.360	0.514	0.857	-0.060	-0.9038	0.8205	0.6706	0.5435	-0.8571
11	10	19	10	0.42	0.60	0.220	0.524	0.873	-0.032	-0.8895	0.6569	0.5329	0.4261	-0.8599
17	16	13	4	0.66	0.60	0.340	0.515	0.859	-0.056	-0.9043	0.7949	0.6491	0.5252	-0.8623
12	11	18	9	0.46	0.60	0.240	0.522	0.870	-0.036	-0.8945	0.6781	0.5507	0.4414	-0.8641
16	15	14	5	0.62	0.60	0.320	0.516	0.860	-0.052	-0.9040	0.7701	0.6282	0.5075	-0.8656
13	12	17	8	0.50	0.60	0.260	0.520	0.867	-0.040	-0.8983	0.7000	0.5692	0.4571	-0.8667
15	14	15	6	0.58	0.60	0.300	0.517	0.862	-0.048	-0.9029	0.7460	0.6080	0.4902	-0.8674
14	13	16	7	0.54	0.60	0.280	0.519	0.864	-0.044	-0.9011	0.7226	0.5883	0.4734	-0.8677
18	19	12	1	0.74	0.60	0.360	0.486	0.811	-0.084	-1.0000	0.7582	0.6161	0.4953	-1.0000

표 3.4는 연관성 순위 결정 함수들의 변화하는 정도를 F_1의 값이 가장 작은 부분을 나타낸 것이다. 이 표에서도 F_{Wu}의 값들은 연관성 측도들의 크기에 따라 다양한 값을 취하고 있으며, 최저 기준값의 변화에 따라 일관성을 보여주지 못하고 있는 동시에, 모두 양의 값으로만 나타나고 있어 방향성을 제시하지 못하고 있다. 또한 F_{Park}값도 향상도의 영향을 크게 받고 있는 반면에 함수 F는 3가지 연관성 측도들의 크기를 잘 반영하고 있다. 특히 $a = 17$, $b = 14$, $c = 13$, $d = 6$인 경우와 $a = 13$, $b = 10$, $c = 17$, $d = 10$인 경우에 이러한 사실을 잘 알 수 있다. 이들을 비교해보면, 지지도는 0.340과 0.260, 신뢰도는 0.548와 0.565, 그리고 향상도는 0.914과 0.942로 F_{Park}는 -0.06713과 -0.6409, 그리고 F는 -0.5601, -0.5652로 나타났다. 따라서 전자에 비해 후자의 지지도가 더 작고, 신뢰도와 향상도가 조금 늘어났으나 F_{Park}값이 더 증가하는 것으로 보아 F_{Park}는 향상도의 영향을 크게 받는다는 사실을 확인할 수 있다. 이러한 현상은 $a = 17$, $b = 15$, $c = 13$, $d = 5$인 경우와 $a = 12$, $b = 10$, $c = 18$, $d = 10$인 경우에도 잘 나타나고 있다.

4. 결론

연관성 규칙을 생성하는 일반적인 기준은 향상도가 1이상이고, 최저 지지도를 만족하는 규칙들 중에서 최저 신뢰도 기준을 초과하는 경우이다. 만약 이 세 가지 모든 조건을 만족하는 경우에는 연관성 규칙이 생성되는 것으로 간주할 수 있으나, 이둘 중 하나라도 만족하지 못하면 버리게 되는데, 최저 신뢰도 기준과 최저 향상도 기준은 만족하나 최저 지지도 기준에는 미달되는 경우, 최저 지지도 기준과 최저 향상도 기준은 만족하나 최저 신뢰도 기준에는 미달되는 경우, 그리고 최저 신뢰도 기준은 만족하나 최저 신뢰도 기준과 최저 향상도 기준에는 미달되는 경우 중에서 연관성 규칙으로 인정해야 할 경우도 종종 발생할 수 있다. 이러한 경우에는 3개의 연관성 측도 중 어느 하나라도 기준 이상이 되는 규칙의 순위를 매겨 필요한 연관성 규칙만을 생성할 수 있는 연관성 순위 결정 함수가 필요하다.

본 논문에서는 기존의 연관성 의사결정 함수에 대해 향상도의 영향을 감소시킬 수 있는 연관성 의사결정 함수를 제안하였다. 이 함수에 대한 유용성을 알아보기 위해 모의실험한 결과, 대체적으로 본 논문에서 제안하는 연관성 의사결정함수는 연관성 측도들과 최저 연관성 기준값들간의 차이를 잘 반영하고 있으며, 최저 연관성 기준값들의 범위와는 관계없이 항상 -1과 1 사이의 값을 가지며, 3개의 연관성 기준값이 모두 충족되면 1의 값을 가지며, 3개 모두 충족되지 않으면 -1의 값을 갖게 된다. 반면에 기존의 함수들은 연관성 측도들과 최저 연관성 기준값들간의 차이를 잘 반영하지 못하거나, 최저 연관성 기준값이 어떤 값을 취하느냐에 따라 값의 범위가 달라지며, 방향성 없이 0과 1 사이의 값을 갖는다는 사실을 확인할 수 있었다.

참고문헌

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.

Bayardo, R. J. (1998). Efficiently mining long patterns from databases. *Proceedings of ACM SIGMOD Conference on Management of Data*, 85-93.

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. (2007). Association rule mining by environmental data fusion. *Journal of the Korean Data & Information Science Society*, **18**, 279-287.

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. (2010). Development of associative rank decision function using basic association rule thresholds. *Journal of the Korean Data Analysis Society*, **12**, unpublished.

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.

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

Toivonen, H. (1996). Sampling large database for association rules. *Proceedings of the 22nd VLDB Conference*, 134-145.

Wu, X., Zhang, C. and Zhang, S. (2004). Eficient mining of both positive and negative association rules. *ACM Transactions on Information Systems*, **22**, 381-405.

Journal of the Korean Data & Information Science Society
2010, **2**(3), 397–405

한국데이터정보과학회지

Association rule ranking function by decreased lift influence

Hee Chang Park[1]

[1]Department of Statistics, Changwon National University

Received 15 March 2010, revised 19 April 2010, accepted 28 April 2010

Abstract

Data mining is the method to find useful information for large amounts of data in database, and one of the important goals is to search and decide the association for several variables. The task of association rule mining is to find certain association relationships among a set of data items in a database. There are three primary measures for association rule, support and confidence and lift. In this paper we developed a association rule ranking function by decreased lift influence to generate association rule for items satisfying at least one of three criteria. We compared our function with the functions suggested by Park (2010), and Wu *et al.* (2004) using some numerical examples. As the result, we knew that our decision function was better than the function of Park's and Wu's functions because our function had a value between -1 and 1 regardless of the range for three association thresholds. Our function had the value of 1 if all of three association measures were greater than their thresholds and had the value of -1 if all of three measures were smaller than the thresholds.

Keywords: Association rule ranking function, confidence, data mining, interest, lift, support.

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