Journal of the Korean Data & Information Science Society
2010, **21**(4), 709–717

한국데이터정보과학회지

조건부 확률증분비를 이용한 연관성 순위 결정 함수

박희창[1]

[1]창원대학교 통계학과

접수 2010년 5월 16일, 수정 2010년 6월 30일, 게재확정 2010년 7월 5일

요 약

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 조건부 확률 증분비를 이용한 연관성 순위 결정 함수를 제안하고자 한다. 특히 항목 집합간의 고유한 연관성 정도를 제대로 반영하기 위해 조건부 확률 증분비를 이용하여 연관성 순위 결정 함수를 제안하여 3개의 연관기준값들 중 어느 하나라도 기준 이상이 되는 규칙의 순위를 매겨 필요한 연관성 규칙만을 생성할 수 있도록 한다. 모의실험을 해본 결과, 본 논문에서 제안한 함수는 기존의 함수와는 달리 특정 연관 기준값의 영향을 받지 않으며, 최저 연관성 기준값들의 범위와는 관계없이 항상 -1과 1 사이의 값을 가진다는 사실을 확인할 수 있었다. 또한 조건부 확률 증분비를 이용한 연관순위결정 함수는 대체적으로 연관성 측도들과 최저 연관기준값들간의 차이를 잘 반영하고 있는 것으로 나타났다.

주요용어: 신뢰도, 연관성 의사 결정 함수, 조건부 확률 증분비, 지지도, 향상도.

1. 서론

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 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 최저신뢰도 기준을 만족하고 향상도가 1 이상인 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 1보다 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] (641-773) 경남 창원시 사림동 9번지, 창원대학교 통계학과, 교수. E-mail: hcpark@changwon.ac.kr

1보다 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 (2010a)는 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. 또한 Park (2010b)는 특정 연관 기준값의 영향을 받지 않도록 하기 위해 향상도 영향의 감소에 의한 연관성 순위 결정 함수의 제시한 바 있으나 이 또한 3개의 연관규칙 평가기준의 범위를 일치시키지 못하고 있다. 따라서 본 논문에서는 특정 연관 기준값의 영향을 받지 않도록 조건부 확률증분비 (CPIR; Conditional Probability Increment Ratio)를 이용한 연관순위결정함수를 제안하고자 한다. 본 논문의 2절에서는 CPIR을 이용한 연관순위결정함수를 제시한 후, 3절에서는 구체적인 예제를 통하여 여러 가지 연관순위결정함수들과의 비교를 통해 본 논문에서 제시한 함수의 유용성 여부를 토의한 후, 마지막으로 4절에서 결론을 내리고자 한다.

2. 조건부 확률증분비를 이용한 연관순위결정함수

연관성 규칙을 평가하는 가장 기본적인 기준에는 지지도, 신뢰도, 향상도 등이 있다. 지지도 $supp(x, y)$는 항목 집합 X와 항목 집합 Y가 동시에 발생하는 거래량 (transaction)의 비율을 의미하며, 다음과 같이 정의된다.

$$supp(x, y) = P(X \ and \ Y) \tag{2.1}$$

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

$$conf(x, y) = P(B|A) \tag{2.2}$$

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

$$lift(x, y) = \frac{P(B|A)}{P(B)} = \frac{P(A \ and \ B)}{P(A)P(B)} \tag{2.3}$$

따라서 연관 규칙 마이닝에서는 향상도가 1이상이고, 최저지지도를 만족하는 규칙들 중에서 최저 신뢰도 기준을 초과하는 경우에 일반적으로 연관성 규칙이 생성되는 것으로 간주한다. 만일 이 세 가지 모든 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. 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 (2010a)는 3가지 기준값 모두가 충족되지 않는 경우의 연관성 규칙들을 서열화할 수 있는 식 (2.4)와 같은 연관성 순위 결정 함수를 제안한 바 있다. 이 식에서 Min_s, Min_c, Min_l는 각각 지지도, 신뢰도, 향상도의 최저기준값이다.

$$F_{P1} = \frac{supp(x,y) + conf(x,y) + lift(x,y) - (Min_s + Min_c + Min_l)}{|supp(x,y) - Min_s| + |conf(x,y) - Min_c| + |lift(x,y) - Min_l|} \tag{2.4}$$

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. 이 식에서 Min_i는 관심도의 최저기준값이다.

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

또한 Park (2010b)는 지지도와 신뢰도에 비해 범위와 크기가 상당히 차이가 나는 향상도의 영향을 감소시키기 위해 향상도의 분모와 분자의 차이를 적용하여 다음과 같은 연관순위결정함수를 제안한 바 있다.

$$F_{P2} = \frac{(supp(x,y) + 2 \cdot conf(x,y)) - (Min_s + Min_c + P(Y) \cdot Min_l)}{|supp(x,y) - Min_s| + |conf(x,y) - Min_c| + |conf(x,y) - P(Y) \cdot Min_l|} \tag{2.6}$$

지지도와 신뢰도가 아무리 큰 값을 갖는다고 해도 향상도가 1 이하가 되면 음의 연관성 규칙이 생성되므로 향상도가 1을 초과하는 것에 대해서만 연관성규칙을 고려하는 것이 바람직하다. 이를 위해 본 논문에서는 CPIR을 이용한 연관순위결정함수를 제안하고자 한다. 먼저 CPIR은 Zhou와 Yau (2007)가 빈발항목과 희귀항목에 대해 동시에 연관성 규칙을 탐색하기 위해 활용한 것으로 양의 연관성 규칙이 생성되는 경우, 즉 $P(Y|X) > P(Y)$인 경우에 다음과 같이 정의된다.

$$CPIR(Y|X) = \frac{P(Y|X) - P(Y)}{1 - P(Y)} \tag{2.7}$$

이 식을 향상도의 관점에서 보면 분모는 향상도의 분모 및 분자의 차이를 의미하며, CPIR은 0과 1사이의 값을 갖게 된다. Park (2010b)와 같이 지지도와 신뢰도에 비해 범위와 크기가 상당히 차이가 나는 향상도의 영향을 감소시키기 위해 향상도의 분모와 분자의 차이만을 고려하게 되면 그 범위는 -1과 1이 되어 지지도와 신뢰도의 범위와 동일하지 않으므로 바람직하지 못한 결론에 다다를 수도 있게 된다. 그러나 CPIR을 이용하면 3개의 연관성 평가기준의 범위가 동일하게 되므로 이들 중 특정 평가기준의 영향을 크게 받는 염려를 하지 않아도 된다. 따라서 식 (2.4)에서 향상도 대신 CPIR과 최저 지지도와의 관계를 고려하여 연관순위결정함수에 적용하면 다음과 같은 식이 얻어진다.

$$F = \frac{(supp(x,y) - Min_s) + (conf(x,y) - Min_c) + [P(Y|X) - P(Y) \cdot Min_l]/[1 - P(Y)]}{|supp(x,y) - Min_s| + |conf(x,y) - Min_c| + |[P(Y|X) - P(Y) \cdot Min_l]/[1 - P(Y)]|} \tag{2.8}$$

3. 예제를 통한 고찰

본 절에서는 예제 데이터를 이용하여 본 논문에서 제안한 CPIR을 고려한 연관 순위 결정 함수를 기존의 3종류의 연관성 순위 결정 함수와 비교해봄으로써 그 유용성을 파악하고자 한다. 이를 위해 항목집합X,Y에 대해 다음과 같이 가정하였다.

먼저 데이터베이스에 있는 총 트랜잭션의 수 (t)를 50명으로 하고, 항목 집합 X는 구매한 냉장고의 금액을 기준으로 100만원 이상 (1) 구매한 사람 수를 20명으로 하고 100만원 미만 (0)을 구매한 사람 수를 30명으로 하였다. 또한 항목 집합 Y를 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. 항

표 3.1 모의실험 데이터 (1)

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

목 집합 X와 Y가 동시에 발생한 빈도 수, 즉 100만원 이상의 냉장고를 구매하고 신용카드로 결제한 빈도수는 $10+r$명으로 하였다. 이를 정리하면 표 3.1과 같다. 이 표에서 e 및 r이 취할 수 있는 정수 값의 범위는 다음과 같다.

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

이로부터 e 및 r의 변화에 따른 지지도, 신뢰도, 지지도와 3개의 함수를 계산한 후, 보다 상세한 설명을 위해 결과를 분할하여 표 3.2 및 표 3.3에 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 $Min_s = 0.25$, $Min_c = 0.65$, 그리고 $Min_l = 1.1$로 가정한다. 또한 $P(X) = 0.4$가 되며, 신뢰도가 $P(Y)$보다 작은 것들은 음의 연관성 규칙을 생성하게 되므로 제외하였다.

모의실험결과를 전체적으로 살펴보면, 대체적으로 본 논문에서 제안하는 함수 F와 F_{P2} 및 F_{P1}은 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. 그러나 함수 F_{P1}은 연관성 측도들과 최저연관성 기준값들간의 차이를 잘 반영하지 못하고 향상도의 영향을 크게 받는다. 또한 F_{P2}는 연관성 측도들과 최저 연관성 기준값들간의 차이는 잘 반영하고 있으나 특정 항목 집합간의 고유한 연관성 정도는 제대로 반영해주지 못하고 있다. 한편 F_{Wu}는 Park (2010b)에서 기술한 바와 같이 최저 연관성 기준값이 어떤 값을 취하느냐에 따라 값의 범위가 달라지며, 방향성 없이 0과 1 사이의 값을 갖는다. 따라서 함수 F_{Wu}는 연관 순의 결정 함수로 사용하기에는 다소 무리가 따른다는 사실이 이미 알려져 있으므로 이 절에서는 F_{P1}과 F_{P2}, 그리고 본 논문에서 제안한 함수에 대해 구체적으로 언급하기로 한다. 표 3.2는 연관성 기준값의 변화에 따라 연관성 순위 결정 함수들의 변화하는 정도를 살펴본 자료 중에서 F의 값이 0보다 큰 부분을 제시한 것이다. 여기서 F_{Wu1}은 $Min_i = 0.1$일 때의 F_{Wu}를 의미한다.

이 표로부터 알 수 있는 사실은 F_{P1}과 F_{P2}는 3개의 연관성 기준값의 변화에는 충실히 잘 반영해주고 있으나 신뢰도와 $P(Y)$와의 차이의 크기는 고려하지 않음으로 인하여 항목 집합 Y에 대한 항목 집합 X의 고유한 연관성 정도는 제대로 반영해주지 못하고 있다는 것이다. 특히 F_{P1}은 향상도 값의 변화에도 민감하다는 것도 알 수 있다. 표 3.2를 좀 더 구체적으로 살펴보면 $a = 12$, $b = 8$, $c = 8$, $d = 22$인 경우와$a = 16$, $b = 4$, $c = 21$, $d = 9$인 경우에 지지도는 0.240과 0.320, 신뢰도는 0.600와 0.800, 그리고 향상도는 1.500과 1.081로 계산이 되었으며, F_{P1}은 0.7391와 0.8416, F_{P2}는 0.4545와 0.8803, 그리고 F는 0.6327, 0.6067로 나타났다. 따라서 후자의 경우가 전자에 비해 지지도와 신뢰도는 더 크고, 향상도는 줄어들어 F_{P1}과 F_{P2}는 그 값은 증가하였으나 F값은 줄어드는 것으로 보아 F_{P1}과 F_{P2}는 $P(Y)$의 크기를 고려하지 않고 신뢰도의 값만을 고려한 것을 알 수 있다. CPIR의 값을 보면 각각 0.3333에서 0.2308로 감소하였으므로 F는 항목 집합 Y에 대한 항목 집합 X의 고유한 연관성 정도를 잘 반영하고 있다고 할 수 있다. 또한 $a = 12$, $b = 8$, $c = 11$, $d = 19$인 경우와 $a = 19$, $b = 1$, $c = 25$, $d = 5$인 경우를 비교해보면 지지도는 0.240과 0.380, 신뢰도는 0.600과 0.950, 향상도는 1.304와 1.080으로 계산되었으며, F_{P1}은 0.5461와 0.9092, F_{P2}는 0.2208와 0.9196, 그리고 F는 0.4873, 0.4828로 얻어진 것으로 보

아 F_{P1}과 F_{P2}의 값의 차이가 많은 반면에 F값은 거의 차이가 나지 않는다. 따라서 F_{P1}는 다른 측도에 비해 향상도의 영향을 더 많이 받고, F_{P2}는 신뢰도의 영향을 많이 받는 것으로 나타났다. 반면에 F는 향상도 대신에 CPIR을 이용함으로써 3개의 연관성 측도들을 골고루 반영하는 것으로 나타났다.

표 3.2 연관 순위 결정 함수 비교 (1)

a	b	c	d	$P(Y)$	$supp$	$conf$	$lift$	$CPIR$	F_{Wu1}	F_{P1}	F_{P2}	F
13	7	5	25	0.36	0.260	0.650	1.806	0.4531	1.0000	1.0000	1.0000	1.0000
18	2	23	7	0.82	0.360	0.900	1.098	0.4444	0.9710	0.9865	0.9890	0.9401
17	3	22	8	0.78	0.340	0.850	1.090	0.3182	0.9545	0.9317	0.9463	0.7772
14	6	18	12	0.64	0.280	0.700	1.094	0.1667	0.9286	0.8551	0.9048	0.7561
12	8	5	25	0.34	0.240	0.600	1.765	0.3939	0.9016	0.8344	0.5804	0.7018
12	8	6	24	0.36	0.240	0.600	1.667	0.3750	0.9000	0.8085	0.5455	0.6832
12	8	7	23	0.38	0.240	0.600	1.579	0.3548	0.8983	0.7773	0.5041	0.6606
12	8	8	22	0.40	0.240	0.600	1.500	0.3333	0.8966	0.7391	0.4545	0.6327
16	4	21	9	0.74	0.320	0.800	1.081	0.2308	0.9365	0.8416	0.8803	0.6067
12	8	9	21	0.42	0.240	0.600	1.429	0.3103	0.8947	0.6912	0.3939	0.5972
12	8	10	20	0.44	0.240	0.600	1.364	0.2857	0.8929	0.6292	0.3182	0.5508
20	0	26	4	0.92	0.400	1.000	1.087	1.0000	0.9737	0.9492	0.9531	0.5385
12	8	11	19	0.46	0.240	0.600	1.304	0.2593	0.8909	0.5461	0.2208	0.4873
19	1	25	5	0.88	0.380	0.950	1.080	0.5833	0.9589	0.9092	0.9196	0.4828
18	2	24	6	0.84	0.360	0.900	1.071	0.3750	0.9429	0.8529	0.8750	0.4118
12	8	12	18	0.48	0.240	0.600	1.250	0.2308	0.8889	0.4286	0.0909	0.3953
15	5	20	10	0.70	0.300	0.750	1.071	0.1667	0.9167	0.6800	0.7647	0.3846
11	9	5	25	0.32	0.220	0.550	1.719	0.3382	0.7937	0.6528	0.2073	0.3827
11	9	6	24	0.34	0.220	0.550	1.618	0.3182	0.7903	0.5985	0.1503	0.3445
17	3	23	7	0.80	0.340	0.850	1.063	0.2500	0.9254	0.7710	0.8125	0.3182
11	9	7	23	0.36	0.220	0.550	1.528	0.2969	0.7869	0.5339	0.0845	0.2985
12	8	13	17	0.50	0.240	0.600	1.200	0.2000	0.8868	0.2500	-0.0909	0.2500
11	9	8	22	0.38	0.220	0.550	1.447	0.2742	0.7833	0.4553	0.0076	0.2418
16	4	22	8	0.76	0.320	0.800	1.053	0.1667	0.9063	0.6457	0.7188	0.1892
11	9	9	21	0.40	0.220	0.550	1.375	0.2500	0.7797	0.3580	-0.0833	0.1702
10	10	5	25	0.30	0.200	0.500	1.667	0.2857	0.6923	0.4783	-0.0811	0.0968
11	9	10	20	0.42	0.220	0.550	1.310	0.2241	0.7759	0.2342	-0.1927	0.0771
18	2	25	5	0.86	0.360	0.900	1.047	0.2857	0.9155	0.7413	0.7734	0.0456

표 3.3은 연관성 순위 결정 함수들의 변화하는 정도를 살펴본 자료 중에서 F가 음의 값을 가지는 부분을 제시한 것이다. 이 표에서도 F_{P1}과 F_{P2}는 3개의 연관성 기준값의 변화를 충실히 잘 반영해주고 있으나 신뢰도와 $P(Y)$와의 차이의 크기는 제대로 반영해주지 못하고 있다는 사실을 알 수 있다.

표 3.3을 좀 더 구체적으로 살펴보면 $a = 11$, $b = 9$, $c = 11$, $d = 19$인 경우와 $a = 20$, $b = 0$, $c = 27$, $d = 3$인 경우에 지지도는 0.220과 0.400, 신뢰도는 0.550와 1.000, 그리고 향상도는 1.250과 1.064로 계산이 되었으며, F_{P1}은 0.0714와 0.8651, F_{P2}는 -0.3265와 0.8727, 그리고 F는 -0.0490, -0.0625로 나타났다. 이로부터 후자가 전자에 비해 지지도와 신뢰도가 약 2배 정도로 커지고, 향상도는 줄어들게 되어 F_{P1}과 F_{P2}의 값은 상당히 증가하였으나 F값은 오히려 줄어드는 것으로 보아 F는 신뢰도와 $P(Y)$의 차이를 고려한 반면에 F_{P1}과 F_{P2}는 $P(Y)$의 크기를 고려하지 않고 신뢰도의 값만을 고려한 것임을 알 수 있다.

표 3.3 연관 순위 결정 함수 비교 (2)

a	b	c	d	$P(Y)$	$supp$	$conf$	$lift$	$CPIR$	F_{Wu1}	F_{P1}	F_{P2}	F
12	8	14	16	0.52	0.240	0.600	1.154	0.1667	0.8519	-0.0541	-0.3636	-0.0141
10	10	7	23	0.34	0.200	0.500	1.471	0.2424	0.6825	0.2990	-0.2270	-0.0233
11	9	11	19	0.44	0.220	0.550	1.250	0.1964	0.7719	0.0714	-0.3265	-0.0490
20	0	27	3	0.94	0.400	1.000	1.064	1.0000	0.9481	0.8651	0.8727	-0.0625
16	4	23	7	0.78	0.320	0.800	1.026	0.0909	0.8769	0.4948	0.5827	-0.0902
10	10	8	22	0.36	0.200	0.500	1.389	0.2188	0.6774	0.1818	-0.3158	-0.1034
10	10	9	21	0.38	0.200	0.500	1.316	0.1935	0.6721	0.0380	-0.4184	-0.2039
18	2	26	4	0.88	0.360	0.900	1.023	0.1667	0.8889	0.6466	0.6822	-0.2230
17	3	25	5	0.84	0.340	0.850	1.012	0.0625	0.8696	0.5340	0.5934	-0.2292
11	9	12	18	0.46	0.220	0.550	1.196	0.1667	0.7544	-0.1522	-0.4943	-0.2294
15	5	22	8	0.74	0.300	0.750	1.014	0.0385	0.8548	0.2686	0.4019	-0.2427
19	1	27	3	0.92	0.380	0.950	1.033	0.3750	0.9067	0.7290	0.7480	-0.2863
14	6	20	10	0.68	0.280	0.700	1.029	0.0625	0.8621	0.0625	0.2500	-0.3043
10	10	10	20	0.40	0.200	0.500	1.250	0.1667	0.6667	-0.1429	-0.5385	-0.3333
17	3	26	4	0.86	0.340	0.850	0.988	-0.0714	0.8429	0.4441	0.5026	-0.4056
13	7	17	13	0.60	0.260	0.650	1.083	0.1250	0.9057	-0.2500	0.0000	-0.4286
20	0	28	2	0.96	0.400	1.000	1.042	1.0000	0.9231	0.7910	0.7986	-0.4737
10	10	11	19	0.42	0.200	0.500	1.190	0.1379	0.6393	-0.3770	-0.6807	-0.5065
11	9	13	17	0.48	0.220	0.550	1.146	0.1346	0.7241	-0.4787	-0.7105	-0.5089
19	1	28	2	0.94	0.380	0.950	1.011	0.1667	0.8816	0.6559	0.6732	-0.5301
12	8	15	15	0.54	0.240	0.600	1.111	0.1304	0.8182	-0.6875	-0.8182	-0.6429
10	10	12	18	0.44	0.200	0.500	1.136	0.1071	0.6129	-0.6923	-0.8519	-0.7500
20	0	29	1	0.98	0.400	1.000	1.020	1.0000	0.8987	0.7254	0.7301	-0.7727
13	7	18	12	0.62	0.260	0.650	1.048	0.0789	0.8704	-0.6754	-0.5238	-0.7877
17	3	30	0	0.94	0.340	0.850	0.904	-1.5000	0.7432	0.1940	0.2236	-0.8272
13	7	19	11	0.64	0.260	0.650	1.016	0.0278	0.8364	-0.7881	-0.6875	-0.8750
12	8	17	13	0.58	0.240	0.600	1.034	0.0476	0.7544	-1.0000	-1.0000	-1.0000

또한 $a = 17$, $b = 3$, $c = 30$, $d = 0$인 경우와 $a = 13$, $b = 7$, $c = 19$, $d = 11$인 경우를 비교해보면 지지도는 0.340과 0.260, 신뢰도는 0.850과 0.650, 향상도는 0.094와 1.016으로 계산되었으며, F_{P1}은 0.1940와 -0.7881, F_{P2}는 0.2236와 -0.6875, 그리고 F는 -0.8272와 -0.8750으로 얻어졌다. 3개의 함수 공히 전자에 비해 후자가 더 작은 값을 가지기는 하나 F_{P1}과 F_{P2}는 각각 향상도와 신뢰도의 영향을 크게 받음으로써 각각의 값이 급격하게 줄어든 반면에 F는 신뢰도와 $P(Y)$의 차이와 $P(Y)$의 크기를 동시에 고려하였으므로 전자의 경우와 후자의 값의 차이는 미미하게 줄어든 것을 알 수 있다. 또한 표 3.1의 행과 열을 바꾸어서 계산하여도 이와 유사한 결과를 얻을 수 있었다.

4. 결론

연관성 규칙을 생성하는 일반적인 기준의 모든 조건을 만족하는 경우에는 연관성 규칙이 생성되는 것으로 간주할 수 있으나, 그렇지 못할 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 본 논문에서는 특정 연관 기준값의 영향을 받지 않도록 조건부 확률증분비를 이용한 연관순위결정함수를 제안하였다. 이 함수에 대한 유용성을 알아보기 위해 모의 실험한 결과, 대체적으로 본 논문에서 제안하는 연관성 의사결정함수는 연관성 측도들과 최저연관성 기준값들간의 차이를 잘 반영하고 있으며, 최저 연관성 기준값들의 범위와는 관계없이 항상 -1과 1 사이의 값을 가지는 것을 확인할 수 있었다. 또한 3개의 연관성 기준값이 모두 충족되면 1의 값을 가지며, 3개 모두 충족되지 않으면 -1의 값을 갖게 된다. 반면에 기존의 함수들은 연관성 측도들과 최저연관성 기준값들간의 차이를 잘 반영하지 못하거나, 최저 연관성 기준값이 어떤 값을 취하느냐에 따라 값의 범위가 달라지며, 방향성 없이 0과 1 사이의 값을 갖는다는 것을 알 수 있었다. 따라서 조건부 확률 증분비를 이용한 연관순위결정함수는 대체적으로 연관성 측도들과 최저 연관기준값들간의 차이를 잘 반영하고 있는 것으로 나타났다. 또한 본 논문에서 제안하는 연관성 의사결정함수는 조건부 확률 증분비를 사용함으로써 대체적으로 항목 집합간의 고유한 연관성 정도를 제대로 반영해준다는 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. (2010a). Development of associative rank decision function using basic association rule thresholds. *Journal of the Korean Data Analysis Society*, **12**, to appear.

Park, H. C. (2010b). Association rule ranking function by decreased lift influence. *Journal of the Korean Data & Information Science Society*, **21**, 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). Efficient mining of both positive and negative association rules. *ACM Transactions on Information Systems*, **22**, 381-405.

Zhou, L. and Yau, S. (2007). Efficient association rule mining among both frequent and infrequent items. *Computers and Mathematics with Applications*, **54**, 737-749.

Journal of the Korean Data & Information Science Society
2010, **21**(4), 709–717

한국데이터정보과학회지

Association rule ranking function using conditional probability increment ratio

Hee Chang Park[1]

[1]Department of Statistics, Changwon National University

Received 16 May 2010, revised 30 June 2010, accepted 5 July 2010

Abstract

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 using conditional probability increment ratio.We compared our function with several association rule ranking functions by some numerical examples. As the result, we knew that our decision function was better than the existing functions. The reasons were that the proposed function of the reference value is not affected by a particular association threshold, and our function had a value between -1 and 1 regardless of the range for three association thresholds. And we knew that the ranking function using conditional probability increment ratio was very well reflected in the difference between association rule measures and the minimum association rule thresholds, respectively.

Keywords: Association rule ranking function, conditional probability increment ratio, confidence, lift, support.

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