Journal of the Korean Data & Information Science Society
2011, **22**(1), 81–88

한국데이터정보과학회지

연관성규칙에서 의미 없는 규칙의 발견에 관한 연구

조광현[1] · 박희창[2]

[1]창원대학교 유아교육학과 · [2]창원대학교 통계학과

접수 2010년 12월 15일, 수정 2011년 01월 06일, 게재확정 2011년 01월 11일

요 약

연관성규칙은 대용량 데이터베이스에서 각 항목들 간의 관련성을 찾아내는 기법으로 둘 또는 그 이상의 품목들 사이의 지지도, 신뢰도, 향상도를 바탕으로 관련성 여부를 측정한다. 연관성규칙에서는 일반적으로 사용하는 연관성규칙 이외에 연관성규칙의 효율성을 개선하기 위하여 여러 가지 제약 기반 연관성규칙의 연구가 활발하게 진행되고 있다. 연관성규칙 생성 시, 종종 많은 규칙들을 발견할 수 있다. 이는 변수들 간에 우연히 관련성이 높게 나타나는 경우가 존재할 수 있고 매개변수에 의하여 직접적인 관련성이 없는 규칙을 발견할 수도 있다. 이에 본 논문에서는 연관성규칙에서 매개변수에 의한 의미 없는 규칙의 발견에 관하여 연구하고자 한다. 본 연구 결과는 연관성 규칙에서 생성된 규칙에 대한 관련성을 보다 정확하게 이해할 수 있도록 함으로써 결과의 해석을 보다 명확하게 할 수 있다.

주요용어: 간접적, 매개검정, 매개변수, 신뢰도, 연관성규칙.

1. 서론

데이터마이닝 (data mining)은 방대한 양의 데이터 속에서 쉽게 드러나지 않는 유용한 정보를 찾아내는 기법으로서 의사결정나무 (decision tree), 연관성규칙 (association rule), 군집분석 (clustering), 신경망 분석 (neural network) 등의 기법이 있다. 데이터마이닝 기법 중 하나인 연관성규칙은 탐색적이며, 비목적성 분석이다. 연관성규칙은 기존의 데이터를 특별한 변형 없이 계산이 용이하게 사용 가능하다는 장점을 가지고 있으나, 계산 과정이 길고, 반복된 계산이 많으며, 적절한 품목의 결정이 어려운 동시에 각 품목의 단위에 따른 표준화가 어렵다는 단점 또한 가지고 있다. 그러나 이러한 단점에도 불구하고 두 품목간의 관계를 지지도 (support), 신뢰도 (confidence), 향상도 (lift)를 이용하여 명확히 수치화함으로써 두 개 이상의 품목간의 관련성을 표시하여 주기 때문에 현업에서 많이 활용되고 있다.

연관성규칙은 Agrawal 등 (1993)에 의해 처음 소개된 이후 연관성규칙의 효율성을 개선하기 위하여 여러 가지 제약 기반 연관성규칙의 연구가 활발하게 진행되고 있다. Agrawal과 Srikant (1994)은 Apriori, AprioriTid 알고리즘을 제안하였고 Park 등 (1995)은 partitioning 알고리즘을 제안하였다. Toivonen (1996)은 sampling 알고리즘을 제안하였으며, Cheung 등 (1996)은 FUP (fast update) 알고리즘에 대한 연구를 하였다.Sergey 등 (1997)에 의해 DIC (dynamic itemset counting)등의 발전된 알고리즘들이 연구되었고 Saygin 등 (2002)은 트랜잭션에 있는 데이터를 별도의 값으로 대치시키고 이때의 최소 지지도와 최대 지지도의 범위를 조정하면서 고려할 항목 집합의 수를 줄이는 방법을 제안하였

[1] (641-773) 경상남도 창원시 사림동 9번지, 유아교육학과 통계학 시간 강사.
[2] 교신저자: (641-773) 경상남도 창원시 사림동 9번지, 통계학과 교수. E-mail : hcpark@changwon.ac.kr

다. 또한 Park과 Cho (2006a, 2006b), Cho와 Park (2006), Kim과 Park (2008), Lee와 Park (2008) 등에 의하여 제약 기반 연관성규칙의 연구가 진행된 바 있다.

일반적으로 독립변수와 종속변수에 대한 상호 관련성을 분석할 때, 외부 변수에 의하여 그 관련성이 직접적으로 유의미하지 않는 경우가 발생할 수 있다. 특히, 독립변수와 종속변수 사이에 매개변수(intervening variable)가 존재하는 경우, 매개변수에 의한 독립변수와 종속변수의 관계를 명확하게 파악할 필요성이 있다. 독립변수와 종속변수 사이에 매개변수가 있는 것으로 판단되면 독립변수와 종속변수는 직접적 관련성보다는 간접적 관련성을 해석하는 것이 바람직하다. 연관성규칙 적용 시, 종종 많은 규칙들이 생성될 수 있으며, 이들 중 의미 없는 규칙이나 매개변수에 의하여 변수들 간의 직접적인 관련성이 없는 규칙들도 존재 할 수 있다. 만일 생성된 규칙 중에서 전항변수 (antecedent variable)와 후항변수 (consequent variable) 사이에 매개변수가 존재할 경우, 실제로 매개변수에 의하여 전항변수와 후항변수는 간접적 관련성만 있는 것으로 볼 수 있다. 그러므로 전항변수와 후항변수의 관련성의 해석보다는 매개변수와 후항변수의 관련성의 해석이 더 바람직하여, 전항변수와 후항변수의 규칙은 큰 의미가 없다고 볼 수 있다. 이에 본 논문에서는 연관성 규칙에 의해 생성된 규칙들 중 변수들 간의 매개 관계를 파악하여 의미 없는 규칙을 발견할 수 있는 방법에 대하여 연구하고자 한다. 본 연구의 결과는 생성된 규칙에 대한 변수들 간의 관련성을 보다 정확하고 명확하게 이해할 수 있어, 의미 없는 규칙을 제거할 수 있을 뿐만 아니라 연관성규칙의 해석을 보다 명확하게 할 수 있다.

본 논문의 2절에서는 이론적 배경에 대하여 기술하고 3절에서는 모의실험을 통하여 논문에서 제시하는 연구의 효과를 살펴본 후, 4절에서 결론을 맺고자 한다.

2. 이론적 배경

연관성규칙은 항목 집합으로 표현된 트랜잭션에서 각 항목간의 연관성을 반영하는 규칙으로서, 기본적으로 미리 결정된 최소 지지도 이상의 트랜잭션 지지도를 가지는 항목 집합들의 모든 집합들인 빈발항목 집합들을 찾아내어 연관규칙을 생성하는 단계로 이루어진다. 설문조사에서 응답선택지가 “예”와 “아니오”의 이분형으로 구성된 문항 X와 Y가 있다고 할 때, 연관성규칙은 문항 X 를 “예”로 응답하면 문항 Y도 “예”로 응답한다는 것을 의미하는 것으로 $X \Rightarrow Y$와 같이 표현된다.연관성규칙을 평가하는 기준에는 지지도, 신뢰도, 향상도 등이 있으며 다음과 같다.

$$\text{지지도 :}\ S_{(X \Rightarrow Y)} = P(X \cap Y) \tag{2.1}$$

$$\text{신뢰도 :}\ C_{(X \Rightarrow Y)} = P(Y|X) = \frac{P(X \cap Y)}{P(X)} \tag{2.2}$$

$$\text{향상도 :}\ L_{(X \Rightarrow Y)} = \frac{P(Y|X)}{P(Y)} = \frac{P(X \cap Y)}{P(X)P(Y)} \tag{2.3}$$

어떤 변수가 독립변수와 종속변수 사이에서 독립변수의 결과인 동시에 종속변수의 원인이 되는 경우 이 변수는 독립변수와 종속변수를 매개 한다고 한다. 어떤 독립변수가 종속변수에 대해 가지는 전체 효과가 유의미한 반면, 그것의 직접적 효과가 유의미하지 않은 경우에는 변수들 사이의 관계를 파악할 필요성이 있다. 예를 들어 ‘부유층의 자녀가 예체능 대학 진학률이 좋다’고 한다면 부유층과 예체능 대학 진학률과의 직접적인 관련성보다는 예체능 대학일수록 특수 교육 등의 사교육비가 많이 들어가기 때문에 ‘사교육비가 많을수록 예체능 대학 진학률이 높다’라고 하는 것이 바람직할 것이다. 이때 사교육비가 부유층과 예체능 대학 진학률을 매개한다고 할 수 있으며 사교육비가 매개변수가 된다.

어떤 독립변수 X_1이 종속변수 Y에 대해 가지는 전체효과가 유의미한 반면, 그것의 직접적 효과가 유의미하지 않은 경우에는 변수들 사이의 관계를 파악할 필요성이 있다. 예를 들면, 매개효과는 종속변수

Y에 대한 X_1의 효과가 X_2에 의해 매개된다고 가정하는 것으로 X_1의 간접적 효과만을 가정하는 것이다. 변수들 사이의 실제 관계가 이와 같을 때, X_1의 전체효과는 유의미한 반면 직접적 효과는 유의미하지 않으며, 전항변수 X_1이 매개하는 변수 X_2에 대한 원인이라고 할 수 있다. 이때 X_2를 매개변수라고 한다.

Cho와 Park (2006)은 연관성 규칙을 이용한 매개 효과 검정에 대하여 연구한 바 있으며, 연관성 규칙을 이용한 매개 효과 검정의 조건은 다음과 같다. 여기서, Y는 후항변수, X_1은 전항변수, 그리고 X_2는 매개변수라고 한다.

(조건 1) Y와 X_1에 대한 연관성규칙의 결과가 지정된 최소 지지도와 최소 신뢰도보다 커야 한다.
(조건 2) X_1과 X_2에 대한 연관성규칙의 결과가 지정된 최소 지지도와 최소 신뢰도보다 커야 한다.
(조건 3) X_1 및 X_2와 Y와의 연관성규칙의 결과가 지정된 최소 지지도와 최소 신뢰도보다 커야 한다.
(조건 4) X_1 및 X_2와 Y와의 연관성규칙의 신뢰도가 Y와 X_1에 대한 연관성규칙의 신뢰도보다 커야 한다.

연관성규칙 생성 시 종종 많은 규칙들이 도출될 수 있으며, 이중 의미 없는 규칙이나 매개변수에 의하여 직접적인 관련성이 없는 규칙들도 존재할 수 있다. 연관성규칙에 의해 생성된 많은 규칙들 중 의미가 없거나 간접적인 관련성이 있는 규칙들을 찾아 낼 수 있다면 생성된 규칙 해석에 대한 노력과 시간이 단축될 뿐만 아니라 생성된 규칙에 대한 관련성을 보다 정확하게 이해할 수 있으므로 결과의 해석을 보다 명확하게 할 수 있을 것이다. Cho와 Park (2006)은 연관성규칙을 활용하여 매개변수에 의한 매개효과를 검정하였다. 이에 본 논문에서는 매개변수를 찾기 위하여 연관성규칙을 이용하는 것이 아니라 생성된 연관성규칙 중 매개변수에 의해 직접적인 관련성이 없는 규칙을 찾아내는 방법을 연구하였고, 나아가 연관성규칙에 의해 생성된 여러 개의 규칙들에서 매개변수가 여러 개가 존재할 수 있으므로 이를 모두 찾아낼 수 있는 방법을 연구하였다. 연관성규칙에서의 의미 없는 규칙을 발견하기 위한 연구의 적용 단계는 그림 2.1과 같다.

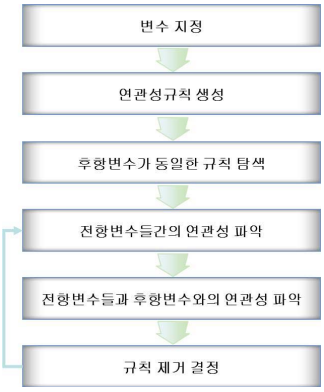

그림 2.1 연구의 적용 단계

적용 단계를 자세하게 살펴보면 다음과 같다.

단계 1) 변수지정 연관성규칙에 사용할 전항변수와 후항변수를 지정한다. 여기서, 매개변수는 생성된

규칙들 중 후항변수가 동일한 전항변수 중에서 선택하도록 한다.

단계 2) 연관성규칙 생성 최소 신뢰도와 최소 지지도를 지정하여 연관성규칙을 생성한다. 이때 최소 신뢰도와 지지도를 낮게 결정하면 연관성규칙이 너무 많이 생성되어 해석의 어려움이 있을 수 있고, 최소 지지도와 신뢰도를 높게 결정하면 의미 있는 규칙을 찾아내지 못하게 되는 경우도 있다.

단계 3) 후항변수가 동일한 규칙 탐색 생성된 규칙들을 바탕으로 실제 매개변수에 의한 불필요한 규칙이 존재하는 지 알아보기 위하여 매개변수를 선정한다. 매개변수는 후항변수가 동일한 규칙들 중에서 전항변수를 하나씩 선정한다.

단계 4) 전항변수들 간의 연관성 파악 후항변수가 동일한 규칙들 중 앞의 단계 3에서 지정한 매개변수와 전항변수와의 연관성 규칙을 생성한다.

단계 5) 전항변수들과 후항변수와의 연관성 파악 단계 4에서 지정한 매개변수와 전항변수를 바탕으로 후항변수와의 연관성 규칙을 생성한다. 이때 최소 신뢰도와 지지도는 단계 2에서와 동일하며, 매개변수와 전항변수 모두를 전항변수로 사용하여 규칙을 생성하므로 본 단계에서는 전항변수의 개수가 2가 된다.

단계 6) 규칙 제거 결정 지정된 매개변수에 의하여 매개 관계가 성립하는 지 알아본다. 실제 매개 관계가 성립하는 경우 그 규칙은 매개변수에 의하여 간접적인 관계만 성립하므로 실제적인 해석은 큰 의미가 없으므로 규칙을 제거하도록 한다. 조건은 다음과 같다.
<조건 1> 단계 4에서 지정한 매개변수와 전항변수와의 연관성이 존재해야 한다.
<조건 2> 단계 5에서 생성한 매개변수 및 전항변수와 후항변수와의 연관성이 존재해야 한다.
<조건 3> 단계 5에서 생성한 연관성 규칙의 신뢰도 값이 단계 4에서 생성한 연관성 규칙의 신뢰도 값보다 커야 한다.

위의 3가지 조건이 만족하면 매개변수에 의한 매개 관계가 성립하여 그 규칙은 매개변수에 의하여 큰 의미가 없는 것으로 판단하며, 조건이 하나라도 맞지 않는 경우에는 매개 관계가 성립하지 않으므로 단계 4로 돌아가 다음 규칙들에 대하여 반복 적용한다. 동일한 후항변수를 가지는 규칙들에 대한 규칙 제거 유무가 결정되면 다른 후항변수를 가지는 규칙들에 대한 단계 4 ∼ 단계 6이 반복 진행된다.

3. 모의실험

본 장에서는 논문에서 제시하는 연구의 효용성을 파악하기 위하여 모의 자료를 바탕으로 분석을 실시하고자 한다. 본 연구에서는 매개변수의 유무를 확인하고 매개변수에 의하여 불필요한 규칙을 찾아낼 수 있는가에 주안점을 두고 있으므로 모의 자료에서는 후항변수 2개와 전항변수 8개의 총 10개의 변수를 생성하였다. 모의 자료에서는 0과 1사이의 소수점 2째 자리까지 랜덤으로 난수를 발생 시켰으며, 생성된 수에 대하여 0에서 0.50까지의 수를 1로 변환하였고, 0.51에서 1까지의 수를 2로 변환하였다. 여기서 자료는 모두 이분형이고 자료의 개수는 1,000개로 하였다. 모의 자료의 설명 및 빈도분석은 표 3.1과 같다.

표 3.1의 자료를 바탕으로 연관성규칙을 생성하였다. 연관성규칙 생성 시 최소 신뢰도와 최소 지지도는 임의로 60과 10으로 지정하였다. 연관성규칙 생성 결과 총 7개의 규칙이 생성되었으며, 생성된 규칙

표 3.1 모의 자료 설명 및 빈도분석

변수	구분	형태	빈도분석
X_1	후항변수	이분형 (범주 :1, 2)	범주 1 : 469개범주 2 : 531개
X_2	후항변수	이분형 (범주 :1, 2)	범주 1 : 493개범주 2 : 507개
X_3	전항변수	이분형 (범주 :1, 2)	범주 1 : 553개범주 2 : 447개
X_4	전항변수	이분형 (범주 :1, 2)	범주 1 : 502개범주 2 : 498개
X_5	전항변수	이분형 (범주 :1, 2)	범주 1 : 449개범주 2 : 551개
X_6	전항변수	이분형 (범주 :1, 2)	범주 1 : 474개범주 2 : 526개
X_7	전항변수	이분형 (범주 :1, 2)	범주 1 : 518개범주 2 : 482개
X_8	전항변수	이분형 (범주 :1, 2)	범주 1 : 435개범주 2 : 565개
X_9	전항변수	이분형 (범주 :1, 2)	범주 1 : 466개범주 2 : 534개
X_{10}	전항변수	이분형 (범주 :1, 2)	범주 1 : 529개범주 2 : 471개

및 신뢰도 값은 표 3.2와 같다.

표 3.2 연관성규칙 생성 결과

규칙	전항변수	후항변수	신뢰도
1	X_4	X_1	62.4
2	X_6	X_1	64.1
3	X_9	X_1	60.3
4	X_{10}	X_1	65.8
5	X_4	X_2	61.5
6	X_6	X_2	63.9
7	X_8	X_2	65.7

표 3.2에서 생성된 연관성규칙의 결과를 바탕으로 그림 2.1의 적용 단계를 적용하였으며, 이에 대한 내용을 간략하게 표 3.3에 기술하였다.

표 3.3 의미 없는 규칙 발견에 대한 최종 연관성 규칙 결과

No	해당 규칙	매개변수	전항변수	후항변수	신뢰도	매개조건 성립 여부		
						조건1	조건2	조건3
1	2	X_4	X_6		61.5	O	-	-
2	2	X_4	X_6	X_1	68.6	-	O	O
3	3	X_4	X_9		62.7	O	-	-
4	3	X_4	X_9	X_1	70.1	-	O	O
5	4	X_4	X_{10}		56.3	X	-	-
6	1	X_6	X_4		61.2	O	-	-
7	1	X_6	X_4	X_1	60.4	-	X	X
8	3	X_6	X_9		52.8	X	-	-
9	4	X_6	X_{10}		50.4	X	-	-
10	1	X_9	X_4		56.8	X	-	-
11	2	X_9	X_6		57.2	X	-	-
12	4	X_9	X_{10}		55.9	X	-	-
13	1	X_{10}	X_4		49.5	X	-	-
14	2	X_{10}	X_6		52.4	X	-	-
15	3	X_{10}	X_9		54.1	X	-	-
16	6	X_4	X_6		61.5	O	-	-
17	6	X_4	X_6	X_2	65.6	-	O	O
18	7	X_4	X_8		52.1	X	-	-
19	5	X_6	X_4		61.2	O	-	-

표 3.3 (계속)의미 없는 규칙 발견에 대한 최종 연관성 규칙 결과

No	해당 규칙	매개변수	전항변수	후항변수	신뢰도	매개조건 성립 여부		
						조건1	조건2	조건3
20	5	X_6	X_4	X_2	60.1	-	X	X
20	7	X_6	X_8		47.6	X	-	-
21	5	X_8	X_4		50.5	X	-	-
22	6	X_8	X_6		45.1	X	-	-

표 3.3의 결과를 자세하게 살펴보면 다음과 같다. 매개변수에 의한 매개 관계가 성립하는 경우에는 회색 (아미)처리를 하였다. 예를 들어, 표의 1번과 2번을 살펴보면 매개 조건 1, 2, 3이 성립하므로 매개 관계가 성립됨을 알 수 있고, 표의 5번을 살펴보면 매개 조건 1이 성립되지 않아 매개 관계가 성립되지 않음을 알 수 있으며, 표의 6번과 7번을 살펴보면 매개 조건 1은 성립되나 2, 3이 성립되지 않으므로 매개 관계가 성립되지 않음을 알 수 있다.

(1) 규칙 2번 대하여 X_4를 매개변수로 지정하였을 경우, 매개변수인 X_4와 전항변수인 X_6의 연관성 규칙 (조건 1)이 성립하고, 매개변수 X_4 및 전항변수 X_6와 후항변수 X_1의 연관성 규칙 (조건 2)이 성립되며, 전항변수 X_4와 후항변수 X_1의 신뢰도 62.4 보다 매개변수 X_4 및 전항변수 X_6와 후항변수 X_1의 신뢰도 68.6이 더 크므로 (조건 3) 매개변수에 의한 매개 관계가 성립된다.

(2) 규칙 3번 대하여 X_4를 매개변수로 지정하였을 경우, 매개변수인 X_4와 전항변수인 X_9의 연관성 규칙 (조건 1)이 성립하고, 매개변수 X_4 및 전항변수 X_9와 후항변수 X_1과의 연관성 규칙 (조건 2)이 성립되며, 전항변수 X_4와 후항변수 X_1의 신뢰도 62.4 보다 매개변수 X_4 및 전항변수 X_9와 후항변수 X_1의 신뢰도 70.1이 더 크므로 (조건 3) 매개변수에 의한 매개 관계가 성립된다.

(3) 규칙 6번 대하여 X_4를 매개변수로 지정하였을 경우, 매개변수인 X_4와 전항변수인 X_6의 연관성 규칙 (조건 1)이 성립하고, 매개변수 X_4 및 전항변수 X_6와 후항변수 X_2과의 연관성 규칙 (조건 2)이 성립되며, 전항변수 X_4와 후항변수 X_2의 신뢰도 61.5 보다 매개변수 X_4 및 전항변수 X_6와 후항변수 X_2의 신뢰도 65.6이 더 크므로 (조건 3) 매개변수에 의한 매개 관계가 성립된다.

결과적으로 표 3.2에서 생성된 7개의 규칙 중 매개변수 X_4에 의하여 규칙 2, 규칙 3, 규칙 6은 직접적인 관련성이 없는 규칙이라고 할 수 있으므로 해석할 필요성이 없다. 즉, 7개의 규칙들 중 의미 없는 규칙이 3개 발견되었다고 볼 수 있다. 이와 같이 본 논문에서 제시하는 방법으로 생성된 규칙들 중 직접적인 관련성이 없는 규칙들을 찾아낼 수 있으므로 생성된 규칙에 대한 변수들 간의 관련성을 보다 정확하고 명확하게 이해할 수 있고, 의미 없는 규칙을 제거할 수 있어 연관성규칙의 해석을 보다 명확하게 할 수 있었다.

4. 결론

일반적으로 독립변수와 종속변수에 대한 상호 관련성을 분석할 때, 매개변수가 존재하는 경우 독립변수와 종속변수는 직접적 관련성보다는 간접적 관련성을 해석하는 것이 바람직하다. 연관성규칙 적용시, 종종 많은 규칙들이 생성될 수 있으며, 이중 의미 없는 규칙이나 매개변수에 의하여 변수들 간의 직접적인 관련성이 없는 규칙들도 존재 할 수 있다. 만일 생성된 규칙 중에서 전항변수와 후항변수 사이에 매개변수가 존재할 경우, 실제로 매개변수에 의하여 전항변수가 후항변수의 간접적 관련성만 있는 것으로 볼 수 있다. 그러므로 전항변수와 후항변수의 관련성의 해석보다는 매개변수와 후항변수의 관련성의 해석이 더 바람직하여, 생성된 규칙은 큰 의미가 없다고 볼 수 있다.

이에 본 논문에서는 연관성 규칙에서 변수들 간의 매개 관계를 규명하여 의미 없는 규칙을 발견할 수

있는 방법을 제안하였고, 모의실험을 통하여 이에 대한 효율성을 파악하였다. 연구 결과, 본 논문에서 제시하는 방법을 통하여 생성된 규칙들 중 직접적인 관련성이 없는 규칙 (의미 없는 규칙)들을 발견할 수 있었다. 이는 생성된 규칙에 대한 변수들 간의 관련성을 보다 정확하고 명확하게 이해할 수 있게 해 주고, 의미 없는 규칙을 제거할 수 있으므로 연관성규칙의 해석을 보다 간결하고 명확하게 해 줄 수 있다. 향후 과제로 본 논문에서 제안하는 방법을 국가 통계, 기업체 및 연구 자료 등에 적용하여 변수들 간의 관계를 보다 명확하게 규명할 필요성이 있다.

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.

Cheung, D. W., Han, J., Ng, V., Fu, A. W. and Fu Y. (1996). A fast distribution algorithm for mining association rules. *Proceedings of International Conference on Parallel and Distributes Information System*, 31-43.

Cho, K. H. and Park, H. C, (2006).A study for intervening effect verification using association rules. *Journal of the Korean Data Analysis Society*, **8,** 1905-1914.

Kim, M. H. and Park, H. C. (2008).Development of componenta ssociation rule s and macro algorithm. *Journal of the Korean Data & Information Science Society*, **19**, 197-207.

Lee. K. W. and Park, H. C. (2008).A study for statistical criterion in negative association rules using boolean analyzer. *Journal of the Korean Data & Information Science Society*, **19** , 569-576.

Park, H. C. and Cho, K. H. (2006a).Discovery of association rules using latent variables. *Journal of the Korean Data & Information Science Society*, **17**, 149-160.

Park, H. C. and Cho, K. H. (2006b).A study for antecedent association rules. *Journal of the Korean Data & Information Science Society*, **17**, 1077-1083.

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.

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.

Toivonen, H. (1996). Sampling Large Database for Association Rules. *Proceedings of the 22nd VLDB Conference*,134-145.

Journal of the Korean Data & Information Science Society
2011, **22**(1), 81–88

한국데이터정보과학회지

A study on insignificant rules discovery in association rule mining

Kwang-Hyun Cho[1] · Hee-Chang Park[2]

[1]Department of Early Childhood Education, Changwon National University
[2]Department of Statistics, Changwon National University

Received 15 December 2010, revised 06 January 2011, accepted 11 January 2011

Abstract

Association rule mining searches for interesting relationships among items in a given database. There are three primary quality measures for association rule, support and confidence and lift. In order to improve the efficiency of existing mining algorithms, constraints were applied during the mining process to generate only those association rules that are interesting to users instead of all the association rules. When we create relation rule, we can often find a lot of rules. This can find rule that direct relativity by intervening variable does not exist. In this study we try to discovery an insignificant rule in association rules by intervening variable. Result of this study can understand relativity about rule that is created in relation rule more exactly.

Keywords: Association rule, indirect, intervening variable, intervening verification, support.

[1] A part-time lecturer, Department of Childhood Education, Changwon National University, Changwon, Gyeongnam, 641-773, Korea.
[2] Corresponding author: Professor, Department of Statistics, Changwon National University, Changwon, Gyeongnam, 641-773, Korea. E-mail: hcpark@changwon.ac.kr