Journal of the Korean Data & Information Science Society
2010, **21**(2), 263–270

한국데이터정보과학회지

올바른 연관성 규칙 생성을 위한 의사결정과정의 제안

박희창[1]

[1]창원대학교 통계학과

접수 2010년 2월 2일, 수정 2010년 3월 4일, 게재확정 2010년 3월 9일

요 약

데이터마이닝은 방대한 양의 데이터 속에서 쉽게 드러나지 않는 유용한 정보를 체계적이고도 자동적으로 찾아내는 기법이다. 데이터마이닝의 중요한 목표 중의 하나는 여러 변수들 간의 관계를 발견하고 결정하는 것이다. 연관성 규칙은 항목 집합으로 표현된 트랜잭션에서 각 항목간의 연관성을 반영하는 규칙으로서, 항목 집합간의 관계를 지지도, 신뢰도, 순수 신뢰도 등과 같은 흥미도 측도에 의해 명확히 수치화함으로써 두 개 이상의 항목집합간의 관련성을 표시해주기 때문에 현업에서 많이 활용되고 있다. 본 논문에서는 기존에 많이 활용되고 있는 흥미도 측도인 신뢰도와 순수 신뢰도의 문제점을 보완하여 연관성 규칙을 올바르게 생성하기 위한 새로운 의사결정과정을 제안하고자 한다. 본 논문에서 제안하는 의사결정과정은 특히 스트리밍 데이터베이스에서의 연관성 규칙을 탐색하는 데 효율적이다.

주요용어: 데이터마이닝, 순수 신뢰도, 신뢰도, 연관성 규칙, 흥미도 측도.

1. 서론

데이터마이닝 (data mining)은 방대한 양의 데이터 속에서 체계적이고도 자동적으로 쉽게 드러나지 않는 유용한 정보, 통계적 규칙 (rule), 그리고 패턴 (pattern) 등을 찾아내는 기법이다. 이 기법은 점점 과대해지는 조직의 경쟁 상황에서 최적의 전략이나 의사결정을 뒷받침해 줄 수 있는 의미 있는 고급 정보를 필요로 하게 되면서 등장하게 되었다. 데이터마이닝 기법 중 하나인 연관성 규칙 (association rule)은 항목 집합으로 표현된 트랜잭션에서 각 항목간의 연관성을 반영하는 규칙으로서, 항목 집합간의 관계를 지지도 (support), 신뢰도 (confidence), 향상도 (lift) 등의 흥미도 측도에 의해 명확히 수치화함으로써 두 개 이상의 항목집합간의 관련성을 표시해주기 때문에 현업에서 많이 활용되고 있다.

연관성 규칙은 항목 집합간의 지지도와 신뢰도를 계산하여, 미리 분석자에 의해서 정해진 최소지지도 및 최소신뢰도를 모두 만족하는 두 항목 집합의 규칙을 의미 있는 연관 규칙을 가진 것으로 판단한다. 그러나 연관성 규칙을 해석할 때 신뢰도 값이 크면 좋지만 신뢰도가 크다고 모두 최선의 연관성 규칙이라고 할 수는 없다. 두 항목 집합의 기본적인 지지도가 어느 정도 수준 이상이 되어야만 의미가 있는 것이다. 또한 신뢰도와 지지도는 자주 발생되는 항목 집합에 대해서는 연관성 때문이 아니라 우연하게 높게 나올 수도 있으므로 향상도를 잘 관찰할 필요가 있다. 이러한 연관성 규칙은 현장에서 직접 적용이 가능하며, 특히 통신업계에서의 교차 판매 분석, 제조업계에서의 불량판별기준의 설정, 그리고 인터넷 비즈니스 환경에서의 시장바구니 분석, 개인화 추천 서비스, 로그파일의 분석 등 많은 분야에서 적용 가능하다.

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

일반적으로 연관성 규칙 생성과정은 크게 두 단계로 구성된다. 첫 번째 단계는 사용자가 지정한 최소 지지도를 만족시키는 빈발항목집합 생성과정이며, 두 번째 단계에서는 빈발항목집합을 이용하여 항목들 간의 규칙을 생성하고 흥미도 측도를 적용하여 연관성 규칙의 유용성 여부를 판단하게 된다.

이러한 연관성 규칙은 Agrawal 등 (1993)에 의해 처음 소개된 이후, 많은 학자들이 연관성 측정에 관한 연구를 수행하였다 (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; Park과 Song, 2002; Cho와 Park, 2007; Choi와 Park, 2008; Lee와 Park, 2008).

Silberschatz와 Tuzhilin (1996), Freitas (1999) 등에 의하면 흥미도 측도는 크게 객관적 흥미도 측도 (objective interestingness measure)와 주관적 흥미도 측도 (subjective interestingness measure)로 나누어진다. 객관적 흥미도 측도는 통계적인 또는 논리적인 방법에 의해 제안된 것으로 사용자에게 규칙을 정제할 수 있는 근거를 제시해주며, 주관적 흥미도 측도는 사용자 관점에서 해석이 가능하도록 제안된 것이다. 흥미도 측도에 관한 연구는 많은 학자들에 의해 수행되었는데, 대표적인 연구로는 Hilderman 등 (2000), Bing 등 (2000), 그리고 Tan 등 (2002) 등이 있다.

본 논문에서는 기존에 많이 활용되고 있는 흥미도 측도인 신뢰도 (confidence)와 순수 신뢰도 (net confidence)의 문제점을 보완하여 연관성 규칙을 올바르게 생성하기 위한 새로운 의사결정과정을 제안하고자 한다. 본 논문에서 제안하는 의사결정과정은 특히 스트리밍 (streaming) 데이터베이스에서의 연관성 규칙을 탐색하는 데 효율적이다. 본 논문의 2절에서는 신뢰도와 순수 신뢰도의 정의와 문제점을 고찰해 본 후, 3절에서는 올바른 의사결정과정을 제안하며, 4절에서 예제를 통해 의사결정과정에서 일어날 수 있는 여러 가지 상황을 파악하여 5절에서 결론을 맺고자 한다.

2. 신뢰도와 순수 신뢰도의 문제점 고찰

연관성 규칙을 평가하는 기준에는 지지도, 신뢰도, 향상도 등이 있다. 이 중에서 신뢰도 C (X⇒Y)는 항목 집합 X가 포함된 거래 비율 중 항목 집합 X와 항목 집합 Y가 동시에 포함된 거래의 비율을 의미하며, 다음과 같이 정의된다.

$$C(X \Rightarrow Y) = P(Y|X) \tag{2.1}$$

순수 신뢰도는 신뢰도의 단점을 극복하기 위해 안광일과 김성집 (2003)이 의학 분야에서 널리 이용되고 있는 기여위험률 (attributable risk)을 데이터마이닝 분야에 적용한 것으로 다음과 같이 정의된다.

$$Nconf(X \Rightarrow Y) = P(Y|X) - P(Y|\bar{X}) \tag{2.2}$$

여기서 $\bar{X}$의 의미는 X가 일어나지 않음을 의미한다. 순수 신뢰도는 순수하게 특정 요인에 의해서만 결과가 얼마인가를 나타내주는 측도이며, 부호에 의해 양의 관련성과 음의 관련성을 판단할 수 있다.

본 절에서는 흥미도 측도 중에서 가장 많이 활용되고 있는 신뢰도와 순수 신뢰도의 문제점을 예제 데이터를 이용하여 고찰하고자 한다. 먼저 최저 신뢰도 기준값 (Min_c)을 0.5라고 가정하고 다음과 같은 가상의 분할표를 고려하자.

이 표에서 신뢰도와 순수 신뢰도는 다음과 같이 계산된다.

$$C(X \Rightarrow Y) = 0.6, Nconf(X \Rightarrow Y) = -0.2$$

이러한 경우 최저 신뢰도 기준은 충족하나 $P(Y|X)$에 비해 $P(Y|\bar{X})$의 값이 커서 순수 신뢰도의 값이 음의 값이 나왔으므로 신뢰도의 값만을 이용하여 의사결정을 하게 되면 잘못된 결론에 도달할 수 있다. 따라서 신뢰도는 계산된 값만을 가지고는 양의 연관성을 가지는지 음의 연관성을 가지는지를 알 수 없을

표 2.1 가상의 분할표 (1)

		Y		total
		1	0	
X	1	30	20	50
	0	40	10	50
total		70	30	100

뿐만 아니라 신뢰도만으로는 음의 연관성을 가지는 연관성 규칙을 의미 있는 양의 관계를 가지는 규칙으로 선택하게 되는 오류를 범할 수 있다.

다음과 같은 또 다른 가상의 분할표를 고려하자.

표 2.2 가상의 분할표 (2)

		Y		total
		1	0	
X	1	10	40	50
	0	8	42	50
total		18	82	100

이 표에서는 신뢰도와 순수 신뢰도는 다음과 같이 계산된다.

$$C(X \Rightarrow Y) = 0.2, Nconf(X \Rightarrow Y) = 0.04,$$

이러한 경우에는 최저 신뢰도 기준은 충족하지 않으나 $P(Y|X)$가 $P(Y|\bar{X})$의 값보다 크게 되어 순수 신뢰도의 값이 양의 값이 나왔으므로 순수 신뢰도의 값만을 이용하여 의사결정을 하게 되면 이 또한 잘못된 결론에 도달할 수 있다. 또한 순수 신뢰도는 순수한 연관성 정도와 방향을 알 수는 있으나 Park (2008)이 지적한 바와 같이 동시발생확률 $P(X \text{ and } Y)$와 동시비발생확률 $P(\bar{X} \text{ and } \bar{Y})$의 값이 현저히 차이가 나는 경우에도 순수 신뢰도가 거의 동일한 값으로 나타나는 단점을 가지고 있어서 순수 신뢰도만을 고려하면 잘못된 결론에 이를 수 있다.

3. 의사결정과정의 제안

신뢰도는 위에서 기술한 약점이 있음에도 불구하고 연관성 규칙의 생성을 위해 가장 많이 활용되는 흥미도 측도이며. 순수신뢰도 또한 의학 분야뿐만 아니라 데이터마이닝 분야에서도 많이 활용되고 있다. 따라서 이 절에서는 가장 인기 있는 이들 두 흥미도 측도를 이용하여 다음과 같은 의사결정과정을 제안하고자 한다.

- **[1]** $C(X \Rightarrow Y) > Min_c$이고 $Nconf(X \Rightarrow Y) > 1$이면 연관성 규칙이 생성되는 것으로 간주한다.
- **[2]** $C(X \Rightarrow Y) < Min_c$이고 $Nconf(X \Rightarrow Y) < 1$이면 연관성 규칙이 생성되지 않은 것으로 간주한다.
- **[3]** $C(X \Rightarrow Y) > Min_c$이나 $Nconf(X \Rightarrow Y) < 1$이면 연관성 규칙을 보류한다.
- **[4]** $Nconf(X \Rightarrow Y) > 1$이나 $C(X \Rightarrow Y) < Min_c$이면 연관성 규칙을 보류한다.

4. 적용 예제

본 절에서는 신뢰도와 순수 신뢰도의 문제점을 탐색하고, 연관성 규칙 생성을 위한 올바른 의사결정을 제안하기 위해 항목 집합 X, Y에 대해 다음과 같이 가정하였다. 먼저 데이터베이스에 있는 총 트랜잭션의 수 (t)를 100명으로 하고, 항목 집합 X는 구매한 냉장고의 금액을 기준으로 100만원 이상 (1) 구매한 사람 수를 40명으로 하고 100만원 미만 (0)을 구매한 사람 수를 60명으로 하였다. 또한 항목 집합 Y를 결제 방식을 기준으로 신용 카드로 결제 (1)한 사람 수와 신용 카드 이외의 방법으로 결제 (0)한 사람의 수를 동일하게 50명으로 하였다. 항목 집합 X와 Y가 동시에 발생한 빈도 수, 즉 100만원 이상의 냉장고를 구매하면서 신용카드로 결제한 빈도수는 a명으로 하였다. 이를 정리하면 <표 4.1>과 같다. 이 표에서 a가 취할 정수 값의 범위를 정하면 다음과 같다.

$$0 \leq a \leq 40 \tag{4.1}$$

표 4.1 모의실험 데이터 (1)

		Y		합
		1	0	
X	1	a	$40-a$	40
	0	$50-a$	$a+10$	60
합		50	50	100

이로부터 동시발생빈도 (a)에 따른 신뢰도와 순수 신뢰도를 계산하면 다음 <표 4.2>와 같은 결과를 얻을 수 있다. 여기서 $b = n(X = 1, Y = 0)$, $c = n(X = 0, Y = 1)$, $d = n(X = 0, Y = 0)$을 의미한다.

이 표로부터 알 수 있는 바와 같이 신뢰도는 모두 양의 값을 가지므로 방향이 없으며, 순수 신뢰도는 그 부호에 의해 연관성 규칙의 방향을 알 수는 있으나 신뢰도에 익숙한 분석가들에게는 해석상의 혼란을 초래할 수 있다. 최저신뢰도의 기준값이 0.4라 가정하고 이 표를 구체적으로 살펴보면 동시발생빈도 a의 값이 15 이하인 경우에는 신뢰도의 값이 모두 0.4보다 작고 순수신뢰도의 값이 음의 값을 가지므로 연관성 규칙이 생성되지 않은 것으로 판단하면 된다. 또한 동시발생빈도 a의 값이 21 이상인 경우에는 신뢰도의 값이 모두 0.4보다 크고 순수신뢰도의 값이 양의 값을 가지므로 연관성 규칙이 생성된 것으로 판단하면 된다. 그러나 동시발생빈도 a의 값이 16과 20 사이의 값을 가지게 되면 순수 신뢰도의 값이 음의 값을 가지게 되어 결론을 내리기가 매우 곤란하다. 다시 말해서 신뢰도를 이용하는 경우에는 순수 신뢰도의 값이 음이므로 잘못된 결론에 다다를 수 있으며, 순수 신뢰도의 값이 음의 값이 나왔다고 이를 버리게 되면 신뢰도의 기준으로 봤을 때는 상당히 아쉬울 수가 있다. 이러한 경우에는 더 많은 트랜잭션의 수를 증가시켜 최저신뢰도의 기준값을 만족하는 동시에 순수 신뢰도의 값이 0보다 클 때까지 결론을 보류하는 것이 바람직하다.

이번에는 <표 4.3>와 같이 냉장고의 구매금액이 100만원 이상 (1) 구매한 사람과 100만원 미만 (0)을 구매한 사람의 수를 동일하게 50명으로 하였다. 또한 항목 집합 Y를 결제 방식을 기준으로 신용 카드로 결제 (1)한 사람 수는 60명으로 하고 신용 카드 이외의 방법으로 결제 (0)한 사람의 수를 40명으로 하였다.

이 표에서 a가 취할 정수 값의 범위를 정하면 다음과 같다.

$$10 \leq a \leq 50 \tag{4.2}$$

이 표로부터 각 셀 값의 변화에 따른 신뢰도와 순수 신뢰도를 계산하면 다음 <표 4.4>과 같은 결과를 얻을 수 있다. 이 표에서도 역시 신뢰도는 모두 양의 값을 가지므로 방향이 없으며, 신뢰도가 0.6의 큰

표 4.2 모의실험 데이터 (1)에 의한 의사결정 결과

a	b	c	d	$P(Y\|X)$	$P(Y\|\bar{X})$	순수 신뢰도	의사결정
5	35	45	15	0.125	0.750	-0.625	버림
6	34	44	16	0.150	0.733	-0.583	”
7	33	43	17	0.175	0.717	-0.542	”
8	32	42	18	0.200	0.700	-0.500	”
9	31	41	19	0.225	0.683	-0.458	”
10	30	40	20	0.250	0.667	-0.417	”
11	29	39	21	0.275	0.650	-0.375	”
12	28	38	22	0.300	0.633	-0.333	”
13	27	37	23	0.325	0.617	-0.292	”
14	26	36	24	0.350	0.600	-0.250	”
15	25	35	25	0.375	0.583	-0.208	”
16	24	34	26	0.400	0.567	-0.167	보류
17	23	33	27	0.425	0.550	-0.125	”
18	22	32	28	0.450	0.533	-0.083	”
19	21	31	29	0.475	0.517	-0.042	”
20	20	30	30	0.500	0.500	0.000	”
21	19	29	31	0.525	0.483	0.042	채택
22	18	28	32	0.550	0.467	0.083	”
23	17	27	33	0.575	0.450	0.125	”
24	16	26	34	0.600	0.433	0.167	”
25	15	25	35	0.625	0.417	0.208	”
26	14	24	36	0.650	0.400	0.250	”
27	13	23	37	0.675	0.383	0.292	”
28	12	22	38	0.700	0.367	0.333	”
29	11	21	39	0.725	0.350	0.375	”
30	10	20	40	0.750	0.333	0.417	”

표 4.3 모의실험 데이터 (1)

		Y		합
		1	0	
X	1	a	$50-a$	50
	0	$60-a$	$a-10$	50
합		60	40	100

값을 가짐에도 불구하고 순수 신뢰도는 0의 값을 가지게 되어 연관성 규칙을 생성하는 데 혼란을 초래하게 된다. 이 표에서도 최저신뢰도의 기준값을 0.4로 가정하면 동시발생빈도 a의 값이 19 이하인 경우에는 신뢰도의 값이 모두 0.4보다 작고 순수 신뢰도의 값이 음의 값을 가지므로 연관성 규칙이 생성되지 않은 것으로 판단하면 된다. 또한 동시발생빈도 a의 값이 31이상인 경우에는 신뢰도의 값이 모두 0.4보다 크고 순수 신뢰도의 값이 양의 값을 가지므로 연관성 규칙이 생성된 것으로 판단하면 된다. 그러나 동시발생빈도 a의 값이 20과 30 사이의 값을 가지게 되면 최저신뢰도의 기준은 충족하나 순수 신뢰도의 값이 음의 값을 가지게 되어 결론을 내리기가 매우 곤란하다. 이러한 경우에도 앞의 모의실험결과에서와 같이 더 많은 트랜잭션의 수를 증가시켜 최저신뢰도의 기준값을 만족하는 동시에 순수 신뢰도의 값이 0보다 클 때까지 결론을 보류하는 것이 바람직하다.

5. 결론

본 논문에서는 올바른 연관성 규칙을 생성하기 위해 흥미도 측도로서 가장 많이 활용되고 있는 흥미도

표 4.4 모의실험 데이터 (2)에 의한 신뢰도의 계산 결과

a	b	c	d	$P(Y\|X)$	$P(Y\|\bar{X})$	순수 신뢰도	결정
15	35	45	5	0.300	0.900	-0.600	버림
16	34	44	6	0.320	0.880	-0.560	”
17	33	43	7	0.340	0.860	-0.520	”
18	32	42	8	0.360	0.840	-0.480	”
19	31	41	9	0.380	0.820	-0.440	”
20	30	40	10	0.400	0.800	-0.400	보류
21	29	39	11	0.420	0.780	-0.360	”
22	28	38	12	0.440	0.760	-0.320	”
23	27	37	13	0.460	0.740	-0.280	”
24	26	36	14	0.480	0.720	-0.240	”
25	25	35	15	0.500	0.700	-0.200	”
26	24	34	16	0.520	0.680	-0.160	”
27	23	33	17	0.540	0.660	-0.120	”
28	22	32	18	0.560	0.640	-0.080	”
29	21	31	19	0.580	0.620	-0.040	”
30	20	30	20	0.600	0.600	0.000	”
31	19	29	21	0.620	0.580	0.040	채택
32	18	28	22	0.640	0.560	0.080	”
33	17	27	23	0.660	0.540	0.120	”
34	16	26	24	0.680	0.520	0.160	”
35	15	25	25	0.700	0.500	0.200	”
36	14	24	26	0.720	0.480	0.240	”
37	13	23	27	0.740	0.460	0.280	”
38	12	22	28	0.760	0.440	0.320	”
39	11	21	29	0.780	0.420	0.360	”
40	10	20	30	0.800	0.400	0.400	”

측도인 신뢰도와 의학 분야뿐만 아니라 데이터마이닝 분야에서도 많이 활용되고 있는 순수 신뢰도의 문제점을 보완하여 새로운 의사결정과정을 제안하였다.

예제를 통하여 확인해본 결과, 신뢰도만으로는 음의 연관성을 가지는 연관성 규칙을 의미 있는 양의 관계를 가지는 규칙으로 선택하게 되는 오류를 범할 수 있으며, 최저 신뢰도 기준은 충족하나 순수 신뢰도의 값이 음의 값이 나오는 경우에는 신뢰도의 값만을 이용하여 의사결정을 하게 되면 잘못된 결론에 도달할 수 있다는 사실을 확인하였다. 또한 최저 신뢰도 기준은 충족하지 않으나 순수 신뢰도의 값이 양의 값이 나오는 경우에 순수 신뢰도의 값만을 이용하여 의사결정을 하게 되면 이 또한 잘못된 결론에 도달할 수 있다는 사실을 확인하였다. 이뿐만 아니라 순수 신뢰도는 동시발생확률과 동시비발생확률의 값이 현저히 차이가 나는 경우에는 순수 신뢰도의 값이 비슷한 경향을 보이는 단점을 가지고 있어서 잘못된 결론에 다다를 수 있다는 것을 알 수 있었다.

따라서 최저신뢰도 기준은 충족하나 순수 신뢰도의 값이 음인 경우와 최저신뢰도 기준은 충족하지 않으나 순수 신뢰도의 값이 양인 경우에는 트랜잭션의 수를 증가시켜 최저신뢰도의 기준값을 만족하는 동시에 순수 신뢰도의 값이 0보다 클 때까지 결론을 보류하는 것이 바람직한 것으로 생각된다.

향후에는 신뢰도와 순수 신뢰도에 의해 연관성 규칙에 대한 의사결정을 유보할 경우 트랜잭션의 크기에 대한 연구가 수행되어야 할 것이다.

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.

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

Bing, Liu, B., Hsu, W., Chen, S. and Ma, Y. (2000). Analyzing the subjective interestingness of association rules. *IEEE Intelligent Systems*, **15**, 47-55.

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.

Freitas, A. (1999). On rule interestingness measures. *Knowledge-based System*, **12**, 309-315.

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.

Hilderman, R. J. and Hamilton H. J. (2000). Applying objective interestingness measures in data mining systems. *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, 432-439.

Lee, K. W. and Park, H. C. (2008). Application of k-means clustering for association rule using measure of association. *Journal of the Korean Data & Information Science Society*, **19**, 925-935.

Liu, B., Hsu, W. and Ma, Y. (1999). Mining association rules with multiple minimum supports. *Proceedings of the 5th Int. 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. and Song, K. M. (2002). Statistical decision making of association threshold in association rule data mining. *Journal of the Korean Data & Information Science Society*, **13**, 115-128.

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.

Silberschatz, A. and Tuzhilin, A. (1996). What makes patterns interesting in knowledge discovery systems. *IEEE transactions on Knowledge Data Engineering*, **8**, 970-974.

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

Tan, P. N., Kumar, V. and Srivastava, J. (2002). Selecting the right interestingness measure for association patterns. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 32-41.

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
2010, **21**(2), 263–270

한국데이터정보과학회지

Decision process for right association rule generation

Hee Chang Park[1]

[1]Department of Statistics, Changwon National University

Received 2 February 2010, revised 4 March 2010, accepted 9 March 2010

Abstract

Data mining is the process of sorting through large amounts of data and picking out useful information. An important goal of data mining is to discover, define and determine the relationship between several variables. Association rule mining is an important research topic in data mining. An association rule technique finds the relation among each items in massive volume database. Association rule technique consists of two steps: finding frequent itemsets and then extracting interesting rules from the frequent itemsets. Some interestingness measures have been developed in association rule mining. Interestingness measures are useful in that it shows the causes for pruning uninteresting rules statistically or logically. This paper explores some problems for two interestingness measures, confidence and net confidence, and then propose a decision process for right association rule generation using these interestingness measures.

Keywords: Association rule, confidence, data mining, interestingness measure, net confidence.

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