Journal of Korean
Data & Information Science Society
2005, Vol. 16, No. 2, pp. 207~216

Relation for the Measure of Association and the Criteria of Association Rule in Ordinal Database

Hee Chang Park[1)] · Ho Soon Lee[2)]

Abstract

One of the well-studied problems in data mining is the search for association rules. Association rules are useful for determining correlations between attributes of a relation and have applications in marketing, financial and retail sectors. There are three criteria of association rule; support, confidence, lift. The goal of association rule mining is to find all the rules with support and confidence exceeding some user specified thresholds. We can know there is association between two items by the criteria of association rules. But we can not know the degree of association between two items. In this paper we examine the relation between the measures of association and the criteria of association rule for ordinal data.

Keywords : association rule, confidence, data mining, lift, measure of association, support

1. 서론

연관성 규칙은 각 항목간의 연관성을 반영하는 규칙으로서 둘 또는 그 이상의 항목들 사이의 지지도(support), 신뢰도(confidence), 향상도(lift)를 기반으로 하여 미리 결정된 최소지지도 및 최소 신뢰도 이상의 의미 있는 규칙을 찾아내는 데이터마이닝 기법 중의 하나이다. 연관성 규칙은 Agrawal 등(1993)에 의해 처음 소개되었으며, 그 후로 Agrawal 등(1994)이 후보 항목 집합을 구성하고, 발생 빈도수를 계산하고 난 후에 사용자가 정의한 최소 지지도를 가지고 빈발 항목 집합들을 결정하는 Apriori 및 AprioriTid 알고리즘을 제안하였다. Park 등(1995)은 데이터베이스를 중복되지 않는 크기로 분할하고 한번에 한 개의 분할 영역만을 고려하여 그 안에서 빈발 항목 집합

1) First author : Professor, Department of Statistics, Changwon National University, Changwon, Gyeongnam, 641-773, Korea
E-mail : hcpark@changwon.ac.kr
2) Graduate Student, Department of Statistics, Changwon National University, Changwon, Gyeongnam, 641-773, Korea

을 생성하는 partitioning 알고리즘을 제안하였으며, Toivonen(1996)은 무작위로 선정된 표본을 가지고 빈발 항목 집합들을 찾은 후에 그 결과를 데이터베이스의 나머지 부분을 가지고 증명하는 sampling 알고리즘을 제안하였다. 또한 Cheung 등(1996)은 갱신된 데이터베이스에서 이전에 빈발 항목으로 다루어졌던 항목 집합에 대해서 데이터베이스 스캔 과정을 생략하는 FUP(fast update) 알고리즘에 대한 연구를 하였다. Saygin 등(2002)은 트랜잭션에 있는 데이터를 미지의 값으로 대치시키고 이때의 최소 지지도와 최대 지지도의 범위를 조정하면서 고려할 항목 집합의 수를 줄이는 방법을 제안하였다.

한편, 연관성 규칙에서 카이제곱 통계량의 제안은 Silverstein 등(1998)에 의해 이루어졌으며, 일반적인 연관성 규칙이 두 항목 집합의 동시발생만을 고려하여 지지도와 신뢰도를 계산함으로써, 두 항목 집합의 비발생에 대한 고려를 하지 않아서 발생할 수 있는 문제를 지적하기도 하였다. Park 등(2002)은 기존의 연관성 규칙에서 연관의 근거로 삼고 있는 최소 신뢰도를 통계적 관점에서 접근하여 파악함으로써 보다 객관적인 연관성 규칙의 연관 기준값을 제안하였다. 기존의 연구 결과는 신뢰도가 1 이상인 경우에 한해서 의미 있는 연관성 규칙으로 보며 이를 통해 관련성 여부를 파악했다. 그러나 기존의 연구에서는 관련성 여부는 파악할 수 있었으나 어느 정도의 관련성을 가지고 있는지의 여부는 파악하기 힘들뿐만 아니라 여러 항목들간의 연관성의 비교가 불가능하다는 한계가 있었다.

본 연구에서는 서열 자료에서의 연관측도와 연관성 규칙의 평가 기준과의 관계를 제시함으로써 관련성의 정도가 어느 정도인지를 통계적 관점에서 접근하여 파악함으로써 보다 객관적인 연관성 규칙의 관련성 정도를 제시하고자 한다.

2. 연관성 규칙의 평가 기준

연관성 규칙의 평가기준에는 지지도, 신뢰도, 향상도 등이 있다. 지지도는 항목 집합 X와 항목 집합 Y가 동시에 발생한 거래의 비율을 의미하며, 전체 거래수를 항목 집합 X와 항목 집합 Y가 동시에 발생한 거래의 수를 나누어서 구한다. 지지도는 식 (2.1)과 같이 정의된다.

$$S(X \Rightarrow Y) = \frac{X\text{와 } Y\text{를 동시에 구매한 거래수}}{\text{전체거래수}} = P(X \cap Y) \tag{2.1}$$

신뢰도는 항목집합 X가 포함된 거래 비율 중 항목 집합 X와 Y가 동시에 포함된 거래의 비율을 말한다. 신뢰도는 식 (2.2)와 같이 정의된다.

$$C(X \Rightarrow Y) = P(Y|X) = \frac{P(X \cap Y)}{P(X)} \tag{2.2}$$

향상도는 실제거래발생 확률을 각 항목집합의 거래가 독립적일 경우 그 거래가 동시에 발생할 예상기대확률로 나눈 것을 의미한다. 향상도는 식 (2.3)과 같이 정의된다.

$$L(X \Rightarrow Y) = \frac{P(Y|X)}{P(Y)} = \frac{P(X \cap Y)}{P(X)P(Y)} \tag{2.3}$$

향상도의 값이 1보다 크면 두 아이템이 동시에 발생한 거래확률이 예상확률보다 더 크므로 향상도의 값이 1이상인 경우에 의미 있는 관련성 규칙이라고 할 수 있다. 그 값은 많은 관련성 규칙의 조합 중에서 의미 있는 관련성규칙을 발견하기 위한 기준으로서 많이 사용된다.

3. 연관규칙의 평가 기준과 연관측도와의 관계

3.1 기본 가정

본 절에서는 동시발생 빈도와 연관측도와의 관계, 그리고 연관규칙의 평가기준과 연관측도와의 관계를 규명하기 위해 다음과 같은 2×2 분할표(contingency table)를 가정한다.

<표 1> 기본 가정을 위한 2×2 분할표

		Y		합
		L	T	
X	L	a	$x_1 - a$	x_1
	T	$y_1 - a$	$t-(x_1+y_1)+a$	$x_0 = t - x_1$
합		y_1	$y_0 = t - y_1$	t

각 셀은 다음 조건을 만족해야 한다.

$$\begin{gathered} 0 \le a \le x_1 \\ 0 \le a \le y_1 \\ (x_1+y_1)-t \le a \le x_1+y_1 \end{gathered} \tag{3.1}$$

<표 1>에서 t, x_1, y_1은 단 한번의 데이터베이스 스캔으로 알 수 있으므로, 사전에 이미 알려져 있다고 가정한다. 또한 위에서 t, x_1, y_1이 알려져 있는 경우 항목집합 X와 Y의 동시 발생 빈도인 a에 따라 다른 셀 들은 위의 <표 1>과 같이 계산되어 표현되어 질 수 있다. 여기서 동시 발생 빈도 a에 따른 다른 셀을 계산한 것은 연관성 규칙에서 관심을 가지는 것은 동시에 구매한 경우이므로 동시 발생 빈도를 a라고 정의하였다. 본 연구에서는 t, x_1, y_1을 고정시킨 후, 기존의 연관규칙으로는 제시할 수 없었던 연관성의 정도를 연관측도를 통해 객관적인 연관성 정도의 기준을 제시하고자 한다.

3.2 서열자료의 연관측도

연관측도는 두 변수간의 관련성뿐만 아니라 두 변수간의 연관성의 정도를 측정하기 위한 측도를 말한다. 분할표에서 두 변수가 모두 서열형인 경우에는 다음과 같이 정의되는 일치쌍과 불일치쌍의 수에 의해서 그 연관성의 강도와 방향이 결정된다. 서열자료의 연관측도를 정의하기 위해 일치쌍의 수를 N_c, 불일치쌍의 수를 N_d, X에 대한 동순위 쌍의 수를 N_X, Y에 대한 동순위 쌍의 수를 N_Y, X 및 Y에 대한 동순위 쌍의 수를 N_{XY}, 전체 쌍의 수를 N, 전체 데이터 수를 S_{XY}로 표기한다. <표 1>의 분할표에서 N_c, N_d, N_X, N_Y, N_{XY}, 그리고 N을 정의하면 다음과 같다.

$$\begin{aligned} N_c &= a\times[t-(x_1+y_1)+a] \\ N_d &= (x_1-a)\times(y_1-a) \\ N_X &= a\times(x_1-a)+(y_1-a)\times[t-(x_1+y_1)+a] \\ N_Y &= a\times(y_1-a)+(x_1-a)\times[t-(x_1+y_1)+a] \\ N_{XY} &= N - N_c - N_d - N_X - N_Y \\ N &= S_{XY}(S_{XY}-1)/2 = N_c + N_d + N_X + N_Y + N_{XY} \end{aligned} \tag{3.2}$$

두 변수의 연관성은 연관측도의 값이 -1에 가까울수록 두 변수가 음의 상관을 갖고 +1에 가까울수록 두 변수가 양의 상관을 갖는 것으로 해석된다. 서열 자료에서의 연관측도에는 켄달의 타우 a, 켄달의 타우 b, 켄달의 타우 c, 굿맨-크루스칼의 감마 등이 있다.

1) 켄달의 타우 a

$$\tau_a = \frac{(N_c - N_d)}{N} \tag{3.3}$$

2) 켄달의 타우 b

$$\tau_b = \frac{N_c - N_d}{\sqrt{N_c + N_d + N_X}\sqrt{N_c + N_d + N_Y}} \tag{3.4}$$

3) 켄달의 타우 c

$$\tau_c = \frac{N_c - N_d}{S^2_{XY}[(m-1)/2m]} \tag{3.5}$$

여기서 $m = \min(R,\ C)$이다.

4) 감마

$$\gamma = \frac{(N_c - N_d)}{(N_c + N_d)} \tag{3.6}$$

3.3 연관규칙의 평가기준과 연관측도와의 관계

연관성 규칙은 항목 집합간의 지지도와 신뢰도를 계산하여, 미리 분석자에 의해서 정해진 최소지지도 및 최소 신뢰도를 모두 만족하는 두 항목 집합의 규칙을 강한 연관 규칙을 가진 것으로 판단한다. 그러나 이러한 연관성 규칙은 두 항목 집합간의 관련성의 여부만을 판단할 수 있다. 어느 정도의 관련성이 있는지, 그리고 $X \Rightarrow Y$의 연관성과 $X \Rightarrow Z$의 연관성 중 어느 것이 더 연관성이 높은지에 대해서는 판단하기 힘들다. 따라서 본 논문에서는 서열 자료에서 사용할 수 있는 연관측도와 연관성 규칙의 평가기준과의 관계를 규명하여 연관성 규칙의 관련성의 정도를 통계적 기법을 사용하여 보다 객관적인 연관성 규칙의 연관성 정도를 제시하고자 한다.

먼저, 연관성 규칙을 동시발생빈도와의 관계식으로 나타내면 식 (3.7)과 같이 정의된다.

$$\begin{aligned} S(X \Rightarrow Y) &= \frac{a}{t} \\ C(X \Rightarrow Y) &= \frac{a}{x_1} \\ L(X \Rightarrow Y) &= \frac{ta}{x_1 y_1} \end{aligned} \tag{3.7}$$

연관규칙의 평가기준과 켄달의 타우 a와 관계는 식 (3.8)과 같이 표현된다.

$$\tau_a = \begin{cases} \left(\frac{2t}{t-1}\right) S(X \Rightarrow Y) - \frac{2x_1 y_1}{t(t-1)} \\ \left(\frac{2x_1}{t-1}\right) C(X \Rightarrow Y) - \frac{2x_1 y_1}{t(t-1)} \\ \frac{2x_1 y_1}{t(t-1)} \; L(X \Rightarrow Y) - \frac{2x_1 y_1}{t(t-1)} \end{cases} \tag{3.8}$$

식 (3.8)에서 보는 바와 같이 연관규칙의 평가기준인 지지도, 신뢰도, 향상도 모두 켄달의 타우 a와 선형 관계를 가짐을 알 수 있다.

연관규칙의 평가기준과 켄달의 타우 b와의 관계는 식 (3.9)와 같이 표현된다.

$$\tau_b = \begin{cases} \frac{t^2 S(X \Rightarrow Y)}{\sqrt{(ty_1 - y_1^2)(tx_1 - x_1^2)}} - \frac{x_1 y_1}{\sqrt{(ty_1 - y_1^2)(tx_1 - x_1^2)}} \\ \frac{tx_1 C(X \Rightarrow Y)}{\sqrt{(ty_1 - y_1^2)(tx_1 - x_1^2)}} - \frac{x_1 y_1}{\sqrt{(ty_1 - y_1^2)(tx_1 - x_1^2)}} \\ \frac{x_1 y_1 L(X \Rightarrow Y)}{\sqrt{(ty_1 - y_1^2)(tx_1 - x_1^2)}} - \frac{x_1 y_1}{\sqrt{(ty_1 - y_1^2)(tx_1 - x_1^2)}} \end{cases} \tag{3.9}$$

식 (3.9)에서 보는 바와 같이 연관규칙의 평가기준인 지지도, 신뢰도, 향상도 모두 켄달의 타우 b와 선형 관계를 가짐을 알 수 있다.

연관규칙의 평가기준과 켄달의 타우 c와의 관계는 식 (3.10)과 같이 표현된다.

$$\tau_c = \begin{cases} 4S(X \Rightarrow Y) - \dfrac{4x_1y_1}{t^2} \\ \dfrac{4x_1}{t} C(X \Rightarrow Y) - \dfrac{4x_1y_1}{t^2} \\ \dfrac{4x_1y_1}{t^2} L(X \Rightarrow Y) - \dfrac{4x_1y_1}{t^2} \end{cases} \tag{3.10}$$

식 (3.10)에서 보는 바와 연관규칙의 평가기준인 지지도, 신뢰도, 향상도 모두 켄달의 타우 c와 선형 관계를 가짐을 알 수 있다.

연관규칙의 평가기준과 굿맨-크루스칼의 감마와의 관계는 식 (3.11)과 같이 표현된다.

$$\gamma = \begin{cases} \dfrac{(t^2S(X \Rightarrow Y) - x_1y_1)}{2t^2S(X \Rightarrow Y)^2 + (t - 2x_1 - 2y_1)tS(X \Rightarrow Y) + x_1y_1} \\ \dfrac{(tx_1C(X \Rightarrow Y) - x_1y_1)}{2x_1^2C(X \Rightarrow Y)^2 + (t - 2x_1 - 2y_1)x_1C(X \Rightarrow Y) + x_1y_1} \\ \dfrac{(y_1x_1L(X \Rightarrow Y) - x_1y_1)}{\dfrac{2x_1y_1^2}{t^2} L(X \Rightarrow Y)^2 + \dfrac{(t - 2x_1 - 2y_1)x_1y_1}{t} L(X \Rightarrow Y) + x_1y_1} \end{cases} \tag{3.11}$$

식 (3.11)에서 보는 바와 같이 연관규칙의 평가기준인 지지도, 신뢰도, 향상도 모두 굿맨-크루스칼의 감마와 비선형 관계를 가짐을 알 수 있다.

기존의 연구 결과는 신뢰도가 1 이상인 경우에 한해서 의미 있는 연관성 규칙으로 보며 이를 통해 관련성 여부를 파악했으나 본 연구 결과를 통해 어느 정도의 관련성을 가지고 있는지의 여부를 파악할 수 있을 뿐만 아니라 $X \Rightarrow Y$의 연관성과 $X \Rightarrow Z$의 연관성의 비교가 가능하며 양과 음의 관련성 여부도 파악할 수 있다.

4. 모의실험

본 절에서는 위에서 논의된 관계식을 기반으로, <표 2>를 이용하여 모의실험을 실시하였다. 모든 실험은 t, x_1, y_1을 $t = 100$, $x_1 = 45$, $y_1 = 30$으로 고정한 후, 카이제곱 검정 결과가 유의한 경우에서의 연관측도와 연관 규칙의 평가기준과의 관계를 살펴보았다.

<표 2> 모의 실험을 위한 2×2 분할표

		Y		합
		T	L	
X	T	a	$45-a$	45
	L	$30-a$	$25+a$	55
합		30	70	100

먼저 동시발생빈도, 연관 규칙의 평가기준, 그리고 연관측도의 값을 계산하면 <표 3>과 같이 얻어진다.

<표 3> 동시발생빈도, 연관 규칙의 평가기준과 연관측도

동시발생빈도	S(X=>Y)	C(X=>Y)	L(X=>Y)	타우 a	타우 b	타우 c	감마
0	0	0.000	0.000	-0.273	-0.592	-0.540	-1.000
1	0.01	0.022	0.074	-0.253	-0.548	-0.500	-0.960
2	0.02	0.044	0.148	-0.232	-0.504	-0.460	-0.914
3	0.03	0.067	0.222	-0.212	-0.461	-0.420	-0.862
4	0.04	0.089	0.296	-0.192	-0.417	-0.380	-0.804
5	0.05	0.111	0.370	-0.172	-0.373	-0.340	-0.739
6	0.06	0.133	0.444	-0.152	-0.329	-0.300	-0.668
7	0.07	0.156	0.519	-0.131	-0.285	-0.260	-0.592
8	0.08	0.178	0.593	-0.111	-0.241	-0.220	-0.510
9	0.09	0.200	0.667	-0.091	-0.197	-0.180	-0.424
10	0.1	0.222	0.741	-0.071	-0.154	-0.140	-0.333
11	0.11	0.244	0.815	-0.051	-0.110	-0.100	-0.240
12	0.12	0.267	0.889	-0.030	-0.066	-0.060	-0.145
13	0.13	0.289	0.963	-0.010	-0.022	-0.020	-0.048
14	0.14	0.311	1.037	0.010	0.022	0.020	0.048
15	0.15	0.333	1.111	0.030	0.066	0.060	0.143
16	0.16	0.356	1.185	0.051	0.110	0.100	0.235
17	0.17	0.378	1.259	0.071	0.154	0.140	0.325
18	0.18	0.400	1.333	0.091	0.197	0.180	0.410
19	0.19	0.422	1.407	0.111	0.241	0.220	0.490
20	0.2	0.444	1.481	0.131	0.285	0.260	0.565
21	0.21	0.467	1.556	0.152	0.329	0.300	0.635
22	0.22	0.489	1.630	0.172	0.373	0.340	0.698
23	0.23	0.511	1.704	0.192	0.417	0.380	0.755
24	0.24	0.533	1.778	0.212	0.461	0.420	0.806
25	0.25	0.556	1.852	0.232	0.504	0.460	0.852
26	0.26	0.578	1.926	0.253	0.548	0.500	0.892
27	0.27	0.600	2.000	0.273	0.592	0.540	0.926
28	0.28	0.622	2.074	0.293	0.636	0.580	0.955
29	0.29	0.644	2.148	0.313	0.680	0.620	0.980
30	0.3	0.667	2.222	0.333	0.724	0.660	1.000

<표 3>에서 보는 바와 같이 $t=100, x_1=45, y_1=30$ 인 경우, a가 취할 수 있는 정수 값의 범위는 식 (4.1)과 같다.

$$0 \le a \le 30 \tag{4.1}$$

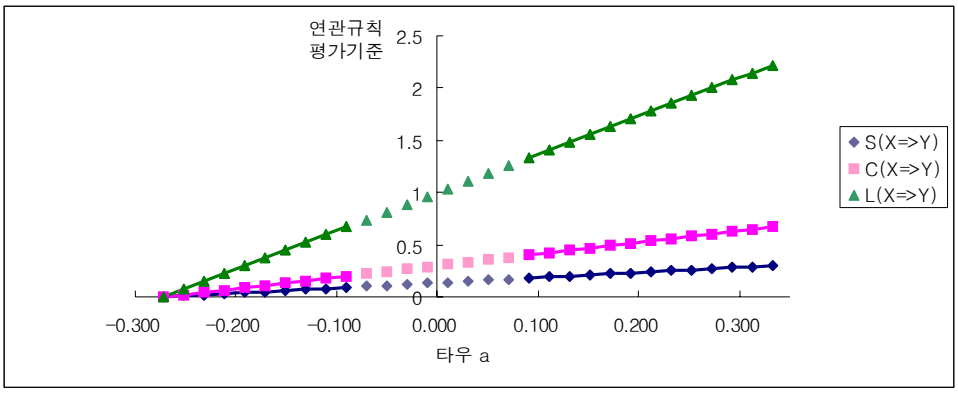
또한 유의수준 $\alpha = 0.05$ 하에서 $\chi^2(1) = 3.84146$이고, 위의 <표 4>에서 카이제곱 검정 결과가 유의한 경우의 a 의 범위는 식 (4.2)와 같다.

$$a \le 9.0317, \quad a \ge 17.9683 \tag{4.2}$$

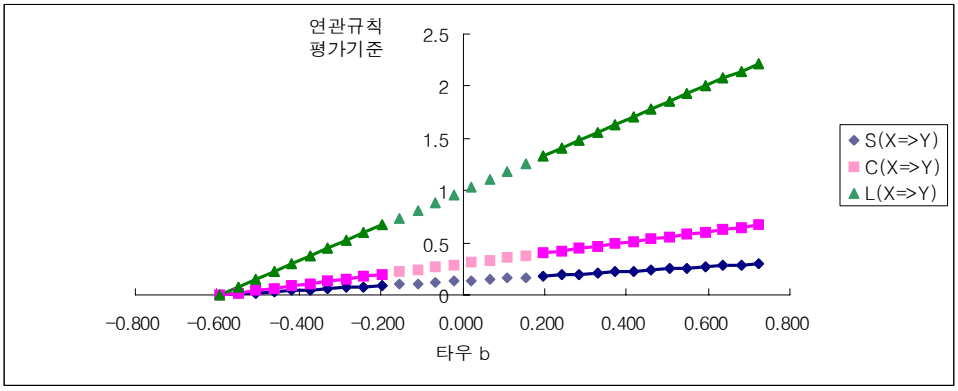
따라서 조건 (4.1)과 조건 (4.2)에서 a는 정수값만 허용되므로 a가 취할 수 있는 값은 식 (4.3)과 같다.

$$0 \le a \le 9, \quad 18 \le a \le 30 \tag{4.3}$$

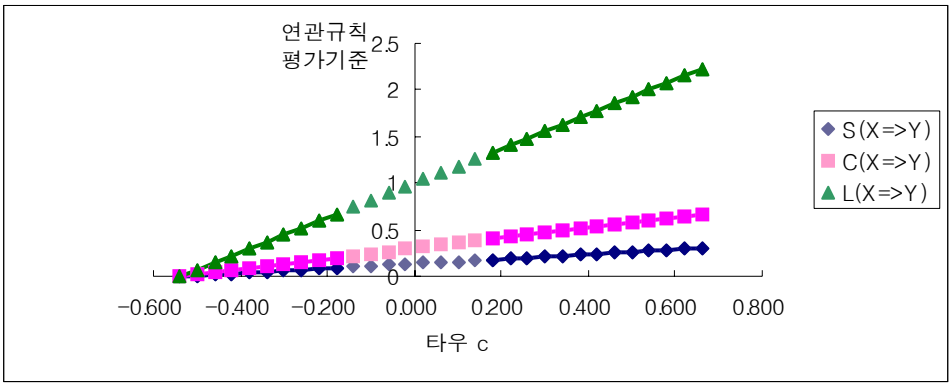
<표 3>에 대해 연관규칙의 평가기준과 연관측도와의 관계를 그림으로 표현하면 다음과 같다. 여기서 연한 선으로 되어 있는 부분이 유의한 경우이다.



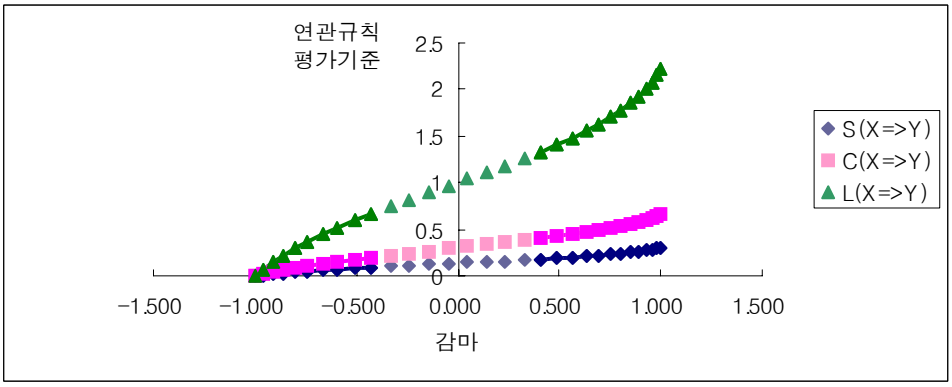
<그림 1> 타우 a와 평가기준과의 관계



<그림 2> 타우 b와 평가기준과의 관계



<그림 3> 타우 c와 평가기준과의 관계



<그림 4> 감마와 평가기준과의 관계

위의 그림에서 보는 바와 같이 감마와 연관규칙의 평가기준들과의 관계를 제외하고는 타우 a, 타우 b, 타우 c와 연관규칙의 평가기준들과의 관계는 선형관계를 가짐을 알 수 있다. 즉, 연관규칙의 평가기준인 지지도, 신뢰도, 향상도가 증가할수록 연관측도들의 값이 감마를 제외하고는 모두 선형적으로 증가함을 알 수 있다. 그러나 감마와 연관규칙의 평가기준들과의 관계에서도 선형 관계를 가지는 것은 아니지만 지지도, 신뢰도, 향상도가 증가할수록 감마값도 증가함을 알 수 있다.

5. 결론 및 향후 과제

연관성 규칙은 둘 또는 그 이상의 항목들 사이의 관련성을 발견하고 분석하는 방법이다. 그러나 기존의 연구에서는 관련성의 여부는 파악할 수 있었으나 어느 정도의 관련성이 있는지의 정도는 파악할 수 없었으며, 항목 집합간의 연관성을 비교 분석할 수 없었다. 그러나 본 연구에서는 기존의 연관성 규칙에서 사용하는 세 가지 평가기준과 서열 자료에서의 연관측도를 관련시킴으로써 연관성 규칙에 대한 관련성 정도를

객관적으로 제시해주는 동시에, 항목집합간의 연관성을 비교 분석할 수 있도록 하였다. 본 연구를 통해 얻을 수 있었던 사실은 연관규칙의 평가기준인 지지도, 신뢰도, 향상도가 증가할수록 연관측도들의 값이 감마를 제외하고는 모두 선형적으로 증가한다는 것이다. 또한 감마와 연관규칙의 평가기준들과의 관계에서도 선형 관계를 가지는 것은 아니지만 지지도, 신뢰도, 향상도가 증가할수록 감마값도 증가함을 알 수 있다. 향후 연구 과제로는 N×N 분할표에서의 연관측도와 연관규칙의 평가기준과의 관계에 대한 연구가 필요하다.

참고문헌

1. Agrawal, R., Imielinski, R., Swami, A. (1993). Mining association rules between sets of items in large databases, *Proceedings of the ACM SIGMOD Conference on Management of Data*, Washington, D.C., 207-216.
2. Agrawal, R., Srikant, R. (1994). Fast algorithms for mining association rules, *Proceedings of the 20th VLDB Conference*, Santiago, Chile, 487-499.
3. Cheung, D.W., Han, J., Ng, V., Fu, A.W., and Fu, Y. (1996). A Fast distribution algorithm for mining association rules, *Int's Conference on Parallel and Distributes Information System*, Miami Beach, Florida, 106-114.
4. Park, H.C., and Song, G.M (2002). Statistical Decision making of Association Threshold in Association Rule Data Mining, *Journal of Korean Data & Information Science Society 2002*, 13(2), 115-128.
5. 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.
6. Saygin, Y., Vassilios, S.V., and Clifton, C. (2002). Using Unknowns to Prevent Discovery of Association Rules, *2002 Conference on Research Issues in Data Engineering*, 45-54.
7. Silverstein, C., Brin, S., and Motwani, R. (1998). Beyond Market Baskets: Generalizing Association Rules to Dependence Rules, *Data Mining and Knowledge Discovery*, 2, 39-68.
8. Toivonen, H. (1996). Sampling Large Database for Association Rules, *Proceedings of the 22nd VLDB Conference,* Mumbai(Bombay), India, 134-145.

[2005년 1월 접수, 2005년 4월 채택]