Proceedings of Joint Conference of Korean Data And Information Science Society
and The Korean Data Analysis Society,
April 30 - May 1, 2004, pp. 25-38

Criteria of Association Rule based on Chi-Square for Nominal Database

Hee Chang Park[1], *Ho Soon Lee*[2]

Abstract

Association rule mining searches for interesting relationships among items in a given database. Association rules are frequently used by retail stores to assist in marketing, advertising, floor placement, and inventory control. There are three primary quality measures for association rule, support and confidence and lift. In this paper we present the relation between the measure of association based on chi square statistic and the criteria of association rule for nominal database and propose the objective criteria for association.

Keywords : data mining, association rule, measure of association, nominal data

1. 서론

데이터 마이닝(data mining)은 대용량 데이터로부터 데이터 내에 존재하는 패턴, 규칙, 관계 등을 탐색하고 찾아내어 의미 있는 지식을 창출해 내는 일련의 과정을 말한다. 데이터마이닝 기법 중의 하나인 연관성 규칙은 각 항목간의 연관성을 반영하는 규칙으로 둘 또는 그 이상의 품목들 사이의 지지도(support), 신뢰도(confidence), 향상도(lift)를 기반으로 하여 미리 결정된 최소지지도 및 최소 신뢰도 이상의 의미 있는 규칙을 찾아낸다. 최근 급격하게 정보량이 증가하고 있는 가운데, 이러한 대용량 데이터들 사이의 흥미 있는 연관성 규칙은 교차판매, 매장 진열, 카탈로그 디자인, 장바구니 분석 등과 같은 많은 비즈니스의 의사 결정 과정에 활용된다.

연관성 규칙은 Agrawal 등(1993)이 처음 소개하였으며, Agrawal 등(1994)은 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 집합을 생성하는 partitioning 알고리즘을 제안하였으며, Tovivonen(1996)은 무작위로 선정된 표본을 가지고 빈발 항목 집합들을 찾은 후, 그 결과를 데이터베이스의 나머지 부분을 가지고 증명하는 샘플링알

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

고리즘을 제안하였다. 또한 Cheung 등(1996)은 갱신된 데이터베이스에서 이전에 빈발 항목으로 다루어 졌던 항목 집합에 대해서 데이터베이스 스캔 과정을 생략하는 FUP(fast update) 알고리즘에 대한 연구를 하였고, Sergey 등(1997)은 데이터가 데이터베이스 전체에 골고루 퍼져있을 경우 적절한 간격을 이용한 알고리즘이 연구되었다. Liu 등(1999)은 후보 항목 집합들을 효율적으로 작게 구하여 이것을 기초로 전체 트랜잭션(transaction)의 크기와 개수를 줄여나가는 DHP(direct hashing and pruning) 알고리즘을 제안하였다. Saygin 등(2002)은 트랜잭션에 있는 데이터를 미지의 값으로 대치시키고 이때의 최소 지지도와 최대 지지도의 범위를 조정하면서 고려할 항목 집합의 수를 줄이는 방법을 제안하였다. 이들 연구들은 주로 대용량 데이터베이스에서 효율적인 연관성을 찾아내고자 하는 처리속도 향상을 위한 알고리즘을 중심으로 연구가 진행되었다.

한편, 연관성 규칙에서 카이제곱 통계량의 제안은 Silverstein 등(1997)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. Park과 Song(2002)은 기존의 연관성 규칙에서 연관의 근거로 삼고 있는 최소 신뢰도를 통계적 관점에서의 카이제곱통계량을 이용하여 연관성 규칙의 연관 기준값을 제안하였다. Park과 Lee(2003)은 순위형 자료에서의 연관성 측도와 연관성 규칙의 평가 기준과의 관계를 제시함으로써 관련성의 정도가 어느 정도인지를 통계적 관점에서 접근하여 파악함으로써 보다 객관적인 연관성 규칙의 관련성 정도를 제시하였다.

본 논문에서는 명목형 자료에서 카이제곱 통계량을 기반으로 한 연관성 측도와 연관성 규칙의 평가 기준과의 관계를 제시함으로써 연관성 규칙의 객관적인 기준을 제안하고자 한다. 본 논문의 2절에서는 연관성 규칙의 평가기준을 간략하게 설명하고, 3절에서 카이제곱 통계량에 근거한 명목형 자료의 연관성 측도를 소개하며, 4절에서는 명목형 자료에서의 연관성 측도와 연관 규칙의 평가기준과의 관계를 규명하고자 한다. 5절에서는 실험을 통해 본 논문에서 제시한 연관성 측도와 연관 규칙의 평가기준과의 관계를 비교 분석하며 마지막으로 제5절에서는 본 논문의 결론 및 향후 연구 과제에 대해서 언급하고자 한다.

2. 연관성 규칙의 평가 기준

연관성 규칙은 X라는 항목집합이 나타나면 Y 라는 항목 집합도 나타난다는 것을 의미하는 것으로 다음과 같이 표현된다.

$$X \Rightarrow Y$$

(if X then Y : 만일 X가 일어나면 Y가 일어난다.)

여기서 X와 Y는 항목집합으로 X는 전제이고 Y는 결론이라 할 수 있다. 연관성 규칙의 평가기준에는 지지도, 신뢰도, 향상도 등이 있다. 지지도는 두 항목 집합 X와 Y가 동시에 발생한 거래의

비율을 나타내는 것으로 식 (2.1)과 같이 정의된다.

$$S_{(X \Rightarrow Y)} = \frac{X와\ Y를\ 동시에\ 구매한\ 거래수}{전체거래수} = P(X \cap Y) \tag{2.1}$$

신뢰도는 항목집합 X가 포함된 거래 비율 중 항목 집합 X와 Y가 동시에 포함된 거래의 비율을 의미하며, 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 것으로 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.

3. 명목형 자료에서의 연관성 측도

연관성 측도(Measure of Association)는 두 변수간의 관련성 뿐만 아니라 두 변수간의 연관성의 정도를 측정하기 위한 측도를 말한다. 본 절에서는 X와 Y가 명목형 변수인 다음과 같은 2×2 분할표(contingency table)를 고려한다.

<표 1> 2×2 분할표

		Y		계
		1	0	
X	1	n_{11}	n_{12}	$n_{1.}$
	0	n_{21}	n_{22}	$n_{2.}$
계		$n_{.1}$	$n_{.2}$	n

3.2 연관성 측도의 종류

명목형 자료에 대해 카이제곱 통계량을 기반으로 한 연관성 측도에는 파이제곱 계수(phi squared coefficient), Cramer의 V(Cramer's V), Tschuprow의 T, 그리고 분할 계수(contingency coefficent) 등이 있다.

3.2.1 파이제곱 계수

파이제곱 계수는 분할표의 크기가 커질수록, 즉 범주의 수가 많아질수록 통계량의 값이 커지는 성질이 있으며, 식 (3.1)과 같이 정의되는 파이 계수를 제곱한 통계량이다.

$$\phi=\begin{cases}\dfrac{n_{11}n_{22}-n_{12}n_{21}}{\sqrt{(n_{1.}n_{2.}n_{.1}n_{.2})}} & \text{, for } 2\times 2 \textit{ table}\\ \sqrt{\dfrac{\chi^2}{n}} & \text{, for } r\times c \textit{ table}\end{cases} \tag{3.1}$$

여기서 χ^2은 $r\times c$ 분할표에서의 카이제곱 통계량 값이다.

파이 계수가 가질 수 있는 범위는 식 (3.2)와 같다.

$$\begin{cases}-1\le \phi \le 1 & \text{, for } 2\times 2 \textit{ table}\\ 0\le \phi \le \min(\sqrt{r-1},\sqrt{c-1}) & \text{, for } r\times c \textit{ table}\end{cases} \tag{3.2}$$

여기서 r은 행범주의 수이고, c는 열범주의 수이다.

3.2.2 Coefficent of Contingency

분할계수 P는 식 (3.3)과 같이 정의된다. 분할계수는 이론적으로는 0과 1 사이의 값을 갖지만 두 변수가 완전한 연관을 갖는다 해도 항상 1의 값을 갖지는 않는다.

$$P=\sqrt{\frac{\chi^2}{\chi^2+n}} \tag{3.3}$$

여기서 χ^2은 $r\times c$ 분할표에서의 카이제곱 통계량 값이며, r은 행범주의 수이며 c는 열범주의 수를 의미한다. 분할 계수가 가질 수 있는 범위는 식 (3.4)와 같다.

$$0\le P\le\sqrt{\frac{(m-1)}{m}} \tag{3.4}$$

여기서 $m=\min(r,\ c)$이다.

3.2.3 Tschuprow의 T

파이제곱 계수의 또 다른 형태인 Tschuprow의 T는 식 (3.5)와 같이 정의되며, 0과 1 사이의 값을 가지면서 변한다.

$$T=\sqrt{\frac{\chi^2/n}{\sqrt{(R-1)(C-1)}}} \tag{3.5}$$

행범주의 수 r과 열범주의 수 c의 값이 동일한 장방형의 교차표에서는 T는 최대값인 1을 취할

수 있으나, 두 값이 다르면 즉, 비대칭교차표에서 T는 1보다 작은 값을 갖는다.

3.2.4 Cramer's V

Cramer의 V는 파이제곱 계수, 분할계수 P, 그리고 Tschuprow의 T의 결점을 보완한 측도로 비대칭형 교차표에서도 최대값이 1이 된다. Cramer의 V는 식 (3.3)과 같이 정의된다.

$$V=\begin{cases}\phi & , \text{for } 2\times 2 \textit{ table} \\ \sqrt{\dfrac{\chi^2/n}{\min(r-1,c-1)}} & , \text{for } r\times c \textit{ table}\end{cases} \tag{3.3}$$

Cramer's V가 가질 수 있는 범위는 식 (3.4)와 같다.

$$\begin{cases} -1\le V\le 1 & , \text{for } 2\times 2 \textit{ table} \\ 0\le V\le 1 & , \text{for } r\times c \textit{ table}\end{cases} \tag{3.4}$$

4. 연관규칙의 평가 기준과 연관성 측도와의 관계

이 절에서는 명목형 자료에서의 연관성 측도와 연관성 규칙의 관련성에 대한 관계를 규명하고자 한다. 2×2 분할표에서는 Tschuprow의 T가 Cramer의 V와 동일하므로 본 절에서는 파이제곱 계수, Cramer의 V, 그리고 분할 계수(contingency coefficent) 등에 대해 동시발생빈도와 연관성 규칙의 평가기준과의 관계를 알아보고자 한다.

4.1 기본 가정

본 절에서는 연관규칙의 평가 기준과 연관성 측도와의 관계를 규명하기 위해 다음과 같은 2×2 분할표를 가정한다.

<표 2> 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

여기서 t 는 전체 트랜잭션의 수, x_1 은 X의 총 발생 빈도 수, y_1 은 Y의 총 발생 빈도 수, 그리고 a 는 동시발생빈도이다. X와 Y는 명목형 속성 값을 가지며, 트랜잭션에 작다(L) 또는 크다(T)로 표현될 수 있다. 각 셀은 식 (4.1)을 만족해야 한다.

$$\begin{array}{c} 0 \le a \le x_1 \\ 0 \le a \le y_1 \\ (x_1+y_1)-t \le a \le x_1+y_1 \end{array} \tag{4.1}$$

<표 2>에서 t, x_1, y_1은 단 한번의 데이터베이스 스캔으로 알 수 있으므로, 사전에 이미 알려져 있다고 가정한다. 이 절에서는 항목 집합 X와 Y의 동시 발생 빈도 a와 명목형 자료에서의 연관성 측도와의 관계를 알아보고, 기존의 연관규칙으로는 제시할 수 없었던 객관적인 기준을 제시하고자 한다.

4.2 동시발생 빈도와 연관성 측도와의 관계

4.2.1 동시발생빈도와 파이제곱 계수와의 관계

동시발생빈도와 파이계수와 관계는 식 (4.2)과 같이 표현된다.

$$\hat{\phi} = \left(\frac{t}{\sqrt{x_1y_1(t-x_1)(t-y_1)}}\right)a-\left(\frac{x_1y_1}{\sqrt{x_1y_1(t-x_1)(t-y_1)}}\right) \tag{4.2}$$

식 (4.2)에서 보듯이 동시발생빈도와 파이계수와의 관계는 선형 관계를 가짐을 알 수 있다. 또한 동시발생빈도와 파이제곱 계수와의 관계는 식 (4.3)와 같이 표현되는데, 이 식에서 보는 바와 같이 이들은 2차 곡선 관계를 가진다.

$$\widehat{\phi^2} = \left(\frac{t^2}{x_1y_1(t-x_1)(t-y_1)}\right)a^2-\left(\frac{2t}{(t-x_1)(t-y_1)}\right)a+\frac{x_1y_1}{(t-x_1)(t-y_1)} \tag{4.3}$$

4.2.2 동시발생빈도와 분할계수와의 관계

동시발생빈도와 분할계수와의 관계는 식 (4.4)과 같이 표현된다.

$$\hat{P} = \sqrt{\left(\frac{t^2a^2-2tx_1y_1a+(x_1y_1)^2}{t^2a^2-2tx_1y_1a+(x_1y_1)^2+x_1y_1(t-x_1)(t-y_1)}\right)} \tag{4.4}$$

식 (4.4)에서 보듯이 동시발생빈도와 분할계수와의 관계는 비선형 관계를 가짐을 알 수 있다.

4.2.3 동시발생빈도와 Cramer의 V와의 관계

동시발생빈도와 Cramer의 V와 관계는 식 (4.5)와 같이 표현된다.

$$\hat{V} = \left(\frac{t}{\sqrt{x_1y_1(t-x_1)(t-y_1)}}\right)a-\left(\frac{x_1y_1}{\sqrt{x_1y_1(t-x_1)(t-y_1)}}\right) \tag{4.5}$$

식 (4.5)에서 보듯이 동시발생빈도와 Cramer의 V와의 관계는 선형 관계를 가짐을 알 수 있다.

4.3 연관규칙의 평가기준과 연관성 측도와의 관계

이 절에서는 이러한 연관성 평가 기준인 지지도, 신뢰도, 향상도과 연관성 측도와의 관계식을 유도하고자 한다. 먼저, 연관성 규칙을 동시발생빈도와의 관계식으로 나타내면 식 (4.8)과 같다.

$$\begin{aligned} S_{(X \Rightarrow Y)} &= \frac{a}{t} \\ C_{(X \Rightarrow Y)} &= \frac{a}{x_1} \\ L_{(X \Rightarrow Y)} &= \frac{ta}{x_1 y_1} \end{aligned} \tag{4.6}$$

4.3.1 연관규칙의 평가기준과 파이제곱 계수와의 관계

연관규칙의 평가기준과 파이계수와의 관계는 식 (4.7)과 같이 표현된다.

$$\begin{aligned} \hat{\phi} &= \left(\frac{t^2}{\sqrt{x_1 y_1 (t-x_1)(t-y_1)}}\right) S_{(X \Rightarrow Y)} - \left(\frac{x_1 y_1}{\sqrt{x_1 y_1 (t-x_1)(t-y_1)}}\right) \\ &= \left(\frac{t x_1}{\sqrt{x_1 y_1 (t-x_1)(t-y_1)}}\right) C_{(X \Rightarrow Y)} - \left(\frac{x_1 y_1}{\sqrt{x_1 y_1 (t-x_1)(t-y_1)}}\right) \\ &= \left(\frac{x_1 y_1}{\sqrt{x_1 y_1 (t-x_1)(t-y_1)}}\right) L_{(X \Rightarrow Y)} - \left(\frac{x_1 y_1}{\sqrt{x_1 y_1 (t-x_1)(t-y_1)}}\right) \end{aligned} \tag{4.7}$$

식 (4.7)에서 보는 바와 같이 연관규칙의 평가기준인 지지도, 신뢰도, 향상도 모두 파이계수와 선형 관계를 가짐을 알 수 있다. 또한 연관규칙의 평가기준과 파이제곱 계수와의 관계는 식 (4.8)과 같이 표현되며, 이를 통해 평가기준들과는 2차 곡선 관계를 가짐을 알 수 있다.

$$\begin{aligned} \widehat{\phi^2} &= \left(\frac{t^4}{x_1 y_1 (t-x_1)(t-y_1)}\right) S^2_{(X \Rightarrow Y)} - \left(\frac{2t^2}{(t-x_1)(t-y_1)}\right) S_{(X \Rightarrow Y)} + \frac{x_1 y_1}{(t-x_1)(t-y_1)} \\ &= \left(\frac{t^2 x_1^2}{x_1 y_1 (t-x_1)(t-y_1)}\right) C^2_{(X \Rightarrow Y)} - \left(\frac{2t x_1}{(t-x_1)(t-y_1)}\right) C_{(X \Rightarrow Y)} + \frac{x_1 y_1}{(t-x_1)(t-y_1)} \\ &= \left(\frac{t x_1 y_1}{x_1 y_1 (t-x_1)(t-y_1)}\right) L^2_{(X \Rightarrow Y)} - \left(\frac{2 x_1 y_1}{(t-x_1)(t-y_1)}\right) L_{(X \Rightarrow Y)} + \frac{x_1 y_1}{(t-x_1)(t-y_1)} \end{aligned} \tag{4.8}$$

4.3.2 연관규칙의 평가기준과 분할계수와의 관계

연관규칙의 평가기준과 분할계수와의 관계는 식 (4.9)와 같이 표현된다.

$$\hat{P} = \sqrt{\left(\frac{t^4 S^2_{(X \Rightarrow Y)} - 2t^2 x_1 y_1 S_{(X \Rightarrow Y)} + (x_1 y_1)^2}{t^4 S^2_{(X \Rightarrow Y)} - 2t^2 x_1 y_1 S_{(X \Rightarrow Y)} + (x_1 y_1)^2 + x_1 y_1 (t - x_1)(t - y_1)}\right)}$$
$$= \sqrt{\left(\frac{t^2 x_1 C^2_{(X \Rightarrow Y)} - 2t x_1^2 y_1 C_{(X \Rightarrow Y)} + (x_1 y_1)^2}{t^2 x_1 C^2_{(X \Rightarrow Y)} - 2t x_1^2 y_1 C_{(X \Rightarrow Y)} + (x_1 y_1)^2 + x_1 y_1 (t - x_1)(t - y_1)}\right)} \quad (4.9)$$
$$= \sqrt{\left(\frac{t x_1 y_1 L^2_{(X \Rightarrow Y)} - 2x_1^2 y_1^2 L_{(X \Rightarrow Y)} + (x_1 y_1)^2}{t x_1 y_1 L^2_{(X \Rightarrow Y)} - 2x_1^2 y_1^2 L_{(X \Rightarrow Y)} + (x_1 y_1)^2 + x_1 y_1 (t - x_1)(t - y_1)}\right)}$$

분할계수는 연관규칙의 평가기준인 지지도, 신뢰도, 향상도 모두와 비선형 관계를 가짐을 식 (4.9)를 통해 알 수 있다.

4.3.2 연관규칙의 평가기준과 Cramer의 V와의 관계

연관규칙의 평가기준과 Cramer의 V와의 관계는 식 (4.10)과 같이 표현된다.

$$\hat{V} = \left(\frac{t^2}{\sqrt{x_1 y_1 (t - x_1)(t - y_1)}}\right) S_{(X \Rightarrow Y)} - \left(\frac{x_1 y_1}{\sqrt{x_1 y_1 (t - x_1)(t - y_1)}}\right)$$
$$= \left(\frac{t x_1}{\sqrt{x_1 y_1 (t - x_1)(t - y_1)}}\right) C_{(X \Rightarrow Y)} - \left(\frac{x_1 y_1}{\sqrt{x_1 y_1 (t - x_1)(t - y_1)}}\right) \quad (4.10)$$
$$= \left(\frac{x_1 y_1}{\sqrt{x_1 y_1 (t - x_1)(t - y_1)}}\right) L_{(X \Rightarrow Y)} - \left(\frac{x_1 y_1}{\sqrt{x_1 y_1 (t - x_1)(t - y_1)}}\right)$$

식 (4.10)에서 보듯이 연관규칙의 평가기준인 지지도, 신뢰도, 향상도 모두 Cramer's V와 선형 관계를 가짐을 알 수 있다.

5. 모의실험

본 절에서는 4절에서 논의된 관계식을 기반으로, t, x_1, y_1을 모두 고정한 후 모의 실험을 실시하였다. 본 절에서의 모의실험은 카이제곱 검정 결과가 유의한 경우에서의 동시발생 빈도와 연관성 측도와의 관계를 규명하고, 연관 규칙의 평가기준과 각각의 연관성 측도와의 관계를 규명하고자한다. 본 절에서 사용된 모의실험 데이터는 <표 3>와 같다.

<표 3> 모의 실험 데이터

		Y		합
		B_1	B_2	
X	A_1	a	45 - a	45
	A_2	30 - a	25 + a	55
합		30	70	100

실험은 $t=100, x_1=45, y_1=30$ 인 경우에 대해서 동시발생빈도와 연관성 측도와의 관계 및 연

관규칙의 평가기준과 연관성 측도와의 관계를 알아보았다. 아래의 <표 4>은 동시발생빈도, 연관 규칙의 평가기준과 연관성 측도값을 나타낸 표이다.

<표 4> 동시발생빈도, 연관 규칙의 평가기준과 연관성 측도

동시발생빈도	$S(X=>Y)$	$C(X=>Y)$	$L(X=>Y)$	ϕ^2	P	V
0	0.00	0.000	0.000	0.351	0.510	-0.592
1	0.01	0.022	0.074	0.301	0.481	-0.548
2	0.02	0.044	0.148	0.254	0.450	-0.504
3	0.03	0.067	0.222	0.212	0.418	-0.461
4	0.04	0.089	0.296	0.174	0.385	-0.417
5	0.05	0.111	0.370	0.139	0.349	-0.373
6	0.06	0.133	0.444	0.108	0.313	-0.329
7	0.07	0.156	0.519	0.081	0.274	-0.285
8	0.08	0.178	0.593	0.058	0.235	-0.241
9	0.09	0.200	0.667	0.039	0.194	-0.197
10	0.10	0.222	0.741	0.024	0.152	-0.154
11	0.11	0.244	0.815	0.012	0.109	-0.110
12	0.12	0.267	0.889	0.004	0.066	-0.066
13	0.13	0.289	0.963	0.000	0.022	-0.022
14	0.14	0.311	1.037	0.000	0.022	0.022
15	0.15	0.333	1.111	0.004	0.066	0.066
16	0.16	0.356	1.185	0.012	0.109	0.110
17	0.17	0.378	1.259	0.024	0.152	0.154
18	0.18	0.400	1.333	0.039	0.194	0.197
19	0.19	0.422	1.407	0.058	0.235	0.241
20	0.20	0.444	1.481	0.081	0.274	0.285
21	0.21	0.467	1.556	0.108	0.313	0.329
22	0.22	0.489	1.630	0.139	0.349	0.373
23	0.23	0.511	1.704	0.174	0.385	0.417
24	0.24	0.533	1.778	0.212	0.418	0.461
25	0.25	0.556	1.852	0.254	0.450	0.504
26	0.26	0.578	1.926	0.301	0.481	0.548
27	0.27	0.600	2.000	0.351	0.510	0.592
28	0.28	0.622	2.074	0.405	0.537	0.636
29	0.29	0.644	2.148	0.462	0.562	0.680
30	0.30	0.667	2.222	0.524	0.586	0.724

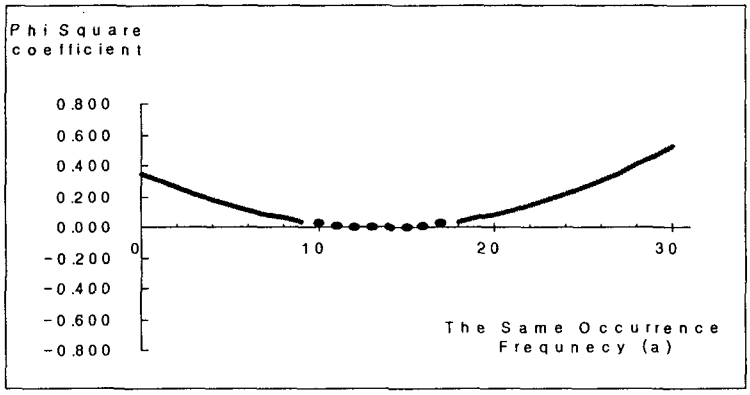
<표 4>에서 보듯이 $t=100, x_1=45, y_1=30$인 경우, a가 취할 수 있는 정수 값의 범위는 조건 (5.1)와 같다.

$$0 \le a \le 30 \tag{5.1}$$

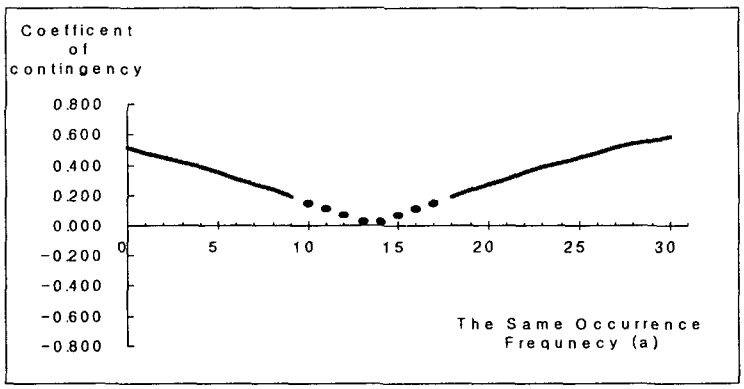
또한 유의수준 $\alpha=0.05$ 하에서 $\chi^2(1)=3.84146$이고, <표 3>에서 카이제곱 검정 결과가 유의한 경우의 a의 범위는 $a \le 9.0317$, $a \ge 17.9683$이다. 또한 a는 정수값만 허용되므로 a가 취할 수 있는 값은 식 (5.2)와 같다.

$$0 \le a \le 9 \tag{5.2}$$

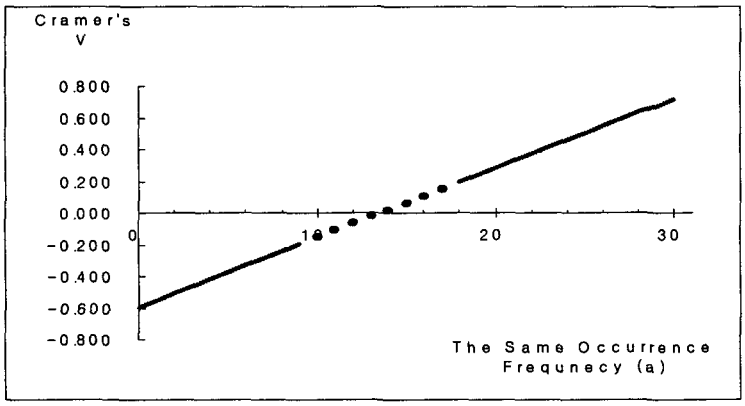
이 조건에 따라 <표 4>에서 a가 취할 수 있는 값을 진하게 표현해 놓았다. <표 4>에서 보듯이 동시발생빈도가 증가할수록 연관성 측도값들이 증가함을 알 수 있다. 4절에서 동시발생빈도와 연관성 측도와의 관계식을 그림으로 나타내면 다음과 같다.



① ϕ^2과 동시발생빈도



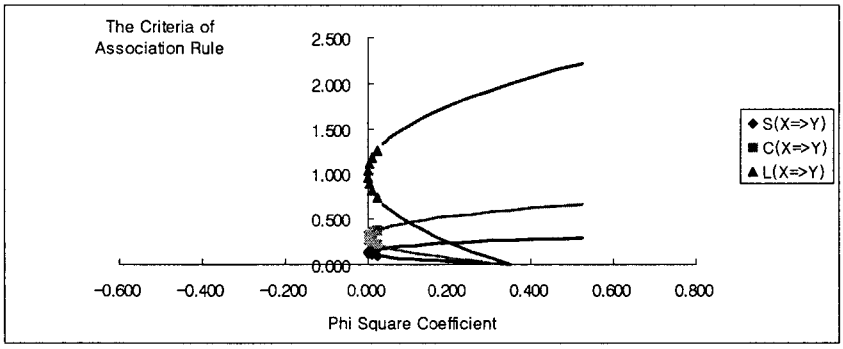
② P와 동시발생빈도



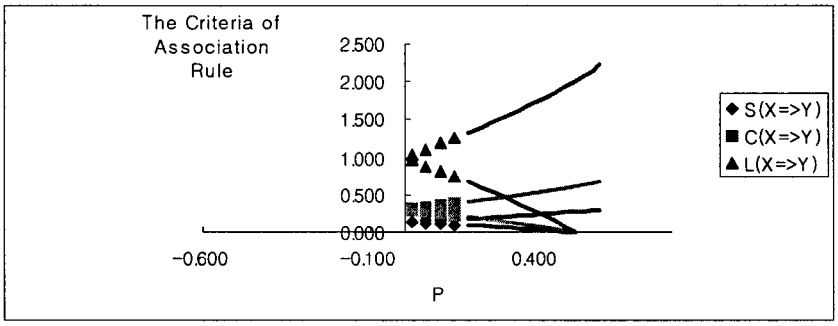
③ V와 동시발생빈도

<그림 2> 동시발생빈도에 따른 연관성 측도값의 변화

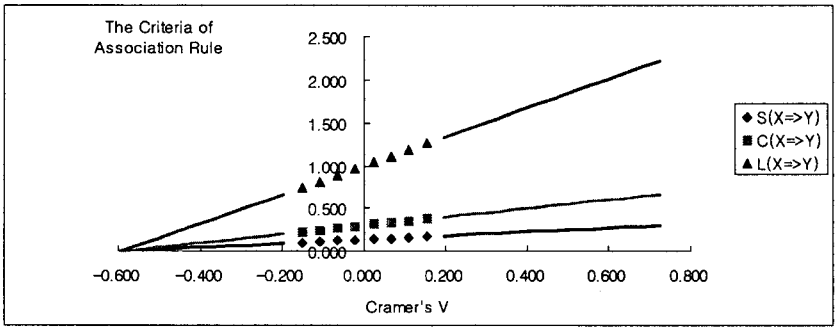
본 논문의 4절에서 연관규칙의 평가기준과 연관성 측도와의 관계식을 그림으로 나타내면 다음과 같다.



① ϕ^2와 연관규칙의 평가기준들과의 관계



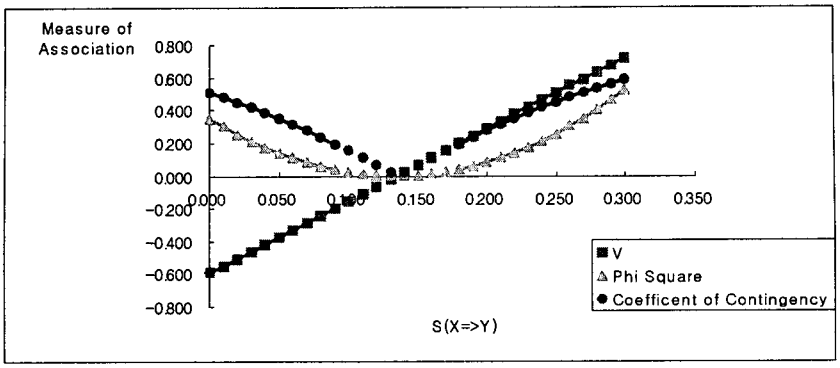
② P와 연관규칙의 평가기준들과의 관계



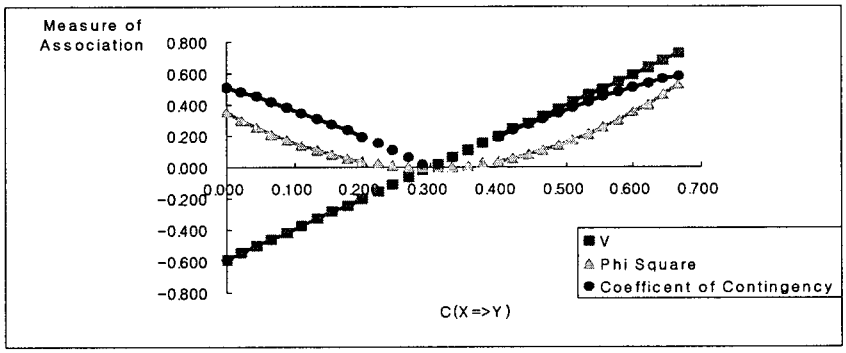
③ V와 연관규칙의 평가기준들과의 관계

<그림 3> 연관성 측도에 따른 연관규칙의 평가기준값의 변화

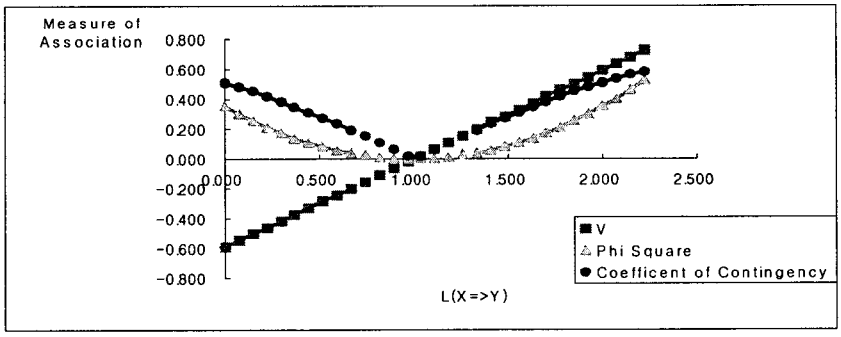
다음은 <표 4>에 대해서 연관성 측도들과 연관규칙의 평가기준과의 관계를 그림으로 나타낸 것이다.



① 지지도와 연관성 측도들과의 관계



② 신뢰도와 연관성 측도들과의 관계



③ 향상도와 연관성 측도들과의 관계

<그림 4> 연관 규칙의 평가기준값에 따른 연관성 측도값의 변화

6. 결론 및 향후 과제

연관성 규칙은 둘 또는 그 이상의 품목들 사이의 관련성을 발견하고 분석하는 방법이다. 그러나 기존의 연구에서는 관련성의 여부는 파악할 수 있었으나 어느 정도의 관련성이 있는지의 정도는 파악할 수 없었으며 둘 이상의 연관규칙 간의 비교 분석을 할 수 없었다.

본 논문에서는 동시발생빈도와 연관성 측도와의 관계 및 연관규칙의 평가기준과 연관성 측도와의 관계식을 제시하고, 제시된 관계식을 기반으로 모의실험을 통해 관계식을 증명했다. 이러한 관

계를 정리하면 동시발생빈도와 명목형 자료에서의 연관성 측도들, 연관규칙의 평가기준들과 연관성 측도들이 선형관계를 가짐을 알 수 있었다. 이 결과를 통해, 본 논문에서는 기존의 연관성 규칙에서 사용하는 세 가지 평가기준과 명목형 자료에서의 연관성 측도를 관련시킴으로써 연관성 규칙에 대한 관련성 정도를 객관적으로 제시해주며 둘 이상의 연관 규칙간의 비교 분석 또한 가능하도록 하였다.

향후 연구 과제로는 N×N 분할표에서의 연관성 측도와 연관규칙의 평가기준과의 관계에 대한 연구가 필요하다.

참고문헌

[1] 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,* Washington, D.C.

[2] Agrawal, R. Imielinski, T. and Swami, A. Mining Associations Between Sets of Items in Massive Databases, *Proceedings of the ACM SIGMOD,* Washington, DC, May 1993, pp. 207-216.

[3] Agrawal, R., and John, C.S. (1996). Parallel Mining of Association Rules, *IEEE Transactions on Knowledge and Data Engineering,* Vol.8, No.6.

[4] A grawal, R., and Srikant, R. (1994). Fast algorithms for mining association rules, *Proceedings of the 20th VLDB Conference*, Santiago, Chile.

[5] Brin S., Motwani R. Ullman J.D., and Tsur S. (1997). Dynamic itemset counting and implication rules for market data, *Proceedings of ACM SIGMOD,* p255-264.

[6] Cheung, D.W., Han, J., Ng, V., and Wong, C.Y. (1996). Maintenance of discovered association rules in large databases: An incremental updating technique, *Int'l Conference on Data Engineering,* New Orleans, Louisiana.

[7] 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.

[8] Han, J., Pei, J., and Yin, Y. (2000). Mining Frequent Patterns Wirhout Candidate Generation, *Proceedings of ACM SIGMOD*, p1-12

[9] Markus, and Hegland. Algorithms for Association Rules, *Australian National University*, Canberra ACT 0200, Australia.

[10] 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, Vol. 13, No.2 pp. 115-128.

[11] 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.*

[12] Savasere A., Omiecinski E., and Navathe S. (1995). An Efficient Algorithm for Mining Association Rules in Large Databases, *Proceedings of the VLDB*, P432-444.

[13] 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.*

[14] Silverstein, C., Brin, S., and Motwani, R. (1998). Beyond Market Baskets: Generalizing Association Rules to Dependence Rules, *Data Mining and Knowledge Discovery,* No.2, P 39-68.

[15] Toivonen, H. (1996). Sampling Large Database for Association Rules, *Proceedings of the 22nd VLDB Conference,* Mumbai(Bombay), India.