Proceedings of Joint Conference of Korean Data And Information Science Society and Korean Data Analysis Society, June 21-22, 2002, pp. 169-182

Statistical Decision making of Association Threshold in Association Rule Data Mining

Hee Chang Park[1], *Geum Min Song*[2]

Abstract

One of the well-studied problems in data mining is the search for association rules. In this paper we consider the statistical decision making of association threshold in association rule. A chi-squared statistic is used to find minimum association threshold. We can calculate the range of the value that two item sets are occurred simultaneously, and can find the minimum confidence threshold values.

Keywords : 연관성규칙, 데이터마이닝, 연관기준값, 최소신뢰도

1. 서론

데이터 마이닝(data mining)의 사전적인 의미는 채굴(mine)한다는 뜻으로 자료를 거대한 더미 속에서 가치 있는 무언가를 캐낸다라는 것이다. 즉, 방대한 양의 데이터로부터 쉽게 드러나지 않는 유용한 정보들을 추출하는 과정을 뜻한다. 여기서 정보는 묵시적이고 잘 알려져 있지 않지만 잠재적으로 활용가치가 있는 정보를 말한다. 연관성 규칙(association rule)은 이러한 데이터 마이닝에서 사용하는 여러 방법들 중의 하나이며, 의사결정나무(decision tree), 신경망(neural network), 클러스터링(clustering), 유전자 알고리즘(genetic algorithm), 베이지안 네트워크(bayesian network), 메모리-기반 추론(memory-based reasoning)등과 같은 관련 기법들이 존재한다. 이중 연관성 규칙은 둘 또는 그 이상의 품목들 사이의 수많은 연관성 규칙 중에서 지지도(support), 신뢰도(confidence), 향상도(lift)에 근거해 일반화 할 수 있는 의미 있는 규칙을 탐색하는 방법이다. 데이터 마이닝에 있어서 이러한 접근 방법들은 분석할 데이터의 유형 및 크기, 데이터 마이닝의 목적에 따라 선택하거나 혹은 두 가지 이상의 분석 방법들을 적용하여 분석 방법에 따른 오류를 계산하여 적절한 모형을 만들게 된다.

연관성 규칙은 Agrawal 등(1993)에 의해 처음 소개된 이후, Agrawal 등(1994)은 후보 항목 집합을 구성하고, 발생 빈도 수를 계산하고 난 후에 사용자가 정의한 최소 지지도를 기초로 빈발 항목 집

[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.

합들을 결정하는 Apriori, 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, 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 하였고, Sergey 등(1997)에 의해 데이터가 데이터베이스 전체에 골고루 퍼져 있을 경우 적절한 간격을 이용한 DIC(dynamic itemset counting)등의 발전된 알고리즘들이 연구되었다. Liu 등(1999)은 후보 항목 집합들을 효율적으로 작게 구하여 이것을 기초로 전체 트랜잭션의 크기와 개수를 줄여나가는 DHP(direct hashing and pruning) 알고리즘을 제안하였다. Saygin 등(2002)은 트랜잭션에 있는 데이터를 미지의 상태로 대치시키고 이때의 최소 지지도와 최대 지지도의 범위를 조정하면서 고려할 항목 집합의 수를 줄이는 방법을 제안하였다. 이들 연구들은 주로 대용량 데이터베이스에서 효율적인 연관성을 찾아내고자 하는 Fast 알고리즘을 중심으로 연구가 진행되었다. 한편 연관성 규칙에서 카이제곱 통계량의 제안은 Silverstein 등(1997)이 일반적인 연관성 규칙이 두 항목 집합의 동시 발생(1)만을 고려하여 지지도와 신뢰도를 계산함으로써, 두 항목 집합의 비발생(0)에 대한 고려를 하지 않아서 발생할 수 있는 문제를 지적하기도 하였다.

그러나 기존의 연구에서는 연관성 규칙에 사용되는 최소 지지도(minimum support), 최소 신뢰도(minimum confidence) 및 동시에 고려할 품목 수를 분석자가 임의로 지정하거나, 전체 데이터베이스를 스캔하여 품목들간의 지지도 및 신뢰도를 계산한 후에, 상위에 위치하는 지지도 및 신뢰도를 이용하여 두 품목간의 관련성을 판단하는 근거로 삼고 있기 때문에 동일한 데이터에 대해 분석자마다 연관성 규칙이 서로 다른 결과를 나타내고 있는 한계점을 가지고 있다. 본 논문에서는 기존의 연관성 규칙에서 연관의 근거로 삼고 있는 최소 신뢰도를 통계적 관점에서 접근하여 파악함으로써, 보다 객관적인 연관성 규칙의 연관 기준값을 제안하고자 한다.

본 연구의 구성은 제2절에서 일반적인 연관성 규칙 및 카이제곱 통계량에 대해 알아보고, 제3절에서는 본 연구를 위한 기본적인 가정 및 용어 정의, 동시 발생 빈도와 카이제곱 통계량, 동시 발생 빈도와 신뢰도의 관계를 유도하고, 카이제곱 통계량과 신뢰도의 관계를 알아봄으로써 연관성 규칙에서의 연관 기준값에 대한 통계적 접근을 시도하였다. 제4절에서는 모의 실험을 통해 본 연구에서 제시한 최소 신뢰도를 산출하여, 기존의 연관성 규칙을 이용한 최소 신뢰도와 비교 분석함으로써 본 연구의 결과가 보다 객관적임을 확인하였다. 마지막으로 제5절에서는 본 연구의 결론 및 향후 연구 과제에 대해서 언급하고자 한다.

2. 연관성 규칙

연관성 규칙은 항목 집합으로 표현된 트랜잭션에서 각 항목간의 연관성을 반영하는 규칙으로서,

기본적으로 미리 결정된 최소 지지도 이상의 트랜잭션 지지도를 가지는 항목 집합들의 모든 집합들인 빈발 항목 집합들을 찾아내어 연관규칙을 생성하는 단계로 이루어진다.

연관성 규칙에서 주로 사용되는 용어들을 정리하면 다음과 같다.

、트랜잭션(transaction) : 발생된 데이터를 저장하는 단위이며, 여러 가지 항목들을 가질 수 있다.

、항목 집합(item set) : 개별 트랜잭션에 포함된 단일 항목, 혹은 복수 항목의 집합이다.

、후보 항목 집합(candidate item set) : 개별 항목 집합의 결합을 통해 생성된 항목 집합으로서 후보 항목 집합에 대한 지지도 및 신뢰도를 계산한다.

、빈발 항목 집합(frequent item set) : 후보 항목 집합에서 최소 지지도 및 신뢰도를 가지는 항목 집합으로서 연관성 규칙에서 생성되는 규칙의 단위가 된다.

k-항목 집합을 가진 $\Im = \{i_{1,} i_{2,}, \ldots, i_k\}$에서 $T \subseteq \Im$ 인 각 트랜잭션을 정의하면, 트랜잭션 T 는 TID 라 불리는 고유한 트랜잭션 아이디를 가진다. 이때 $A \subseteq T$ 일 경우 트랜잭션 T 는 A 를 포함한다고 말한다.

연관성 규칙은 $R: X \Rightarrow Y$ 은 다음과 같이 정의된다.

지지도 : $\mathrm{Sup}(X \Rightarrow Y) = P(X \bigcap Y)$

신뢰도 : $Conf(X \Rightarrow Y) = P(Y|X) = \dfrac{P(X \bigcap Y)}{P(X)}$

향상도 : $Lift(X \Rightarrow Y) = \dfrac{P(Y|X)}{P(Y)} = \dfrac{P(X \bigcap Y)}{P(X)P(Y)}$

여기서 $X \subset T, Y \subset T, \mathrm{and}\, X \bigcap Y = \varnothing$ 이다.

연관성 규칙 $R: X \Rightarrow Y[support = s\%,\ confidence = c\%]$는 X와 Y가 동시에 포함된 트랜잭션이 s% 이고, X가 포함된 트랜잭션 중에서 X와 Y가 동시에 포함된 트랜잭션이 c% 임을 의미한다.

연관성 규칙은 이러한 지지도와 신뢰도를 계산하여, 사전에 분석자가 지정한 최소 지지도 기준(minimum support threshold : min_sup) 및 최소 신뢰도 기준(minimum confidence threshold : min_conf)을 모두 만족하는 규칙을 가진 두 항목 집합 (item-sets)을 강한 연관규칙을 가진 것으로 판단한다.

연관성 규칙을 해석할 때 신뢰도의 값이 크면 좋지만 신뢰도가 크다고 모두 최선의 연관성 규칙은 아니며, 두 항목 집합의 기본적인 지지도가 어느 정도 수준 이상일 경우만 고려해야 한다. 또한 신뢰도와 지지도는 자주 발생되는 항목 집합에 대해서는 연관성 때문이 아니라, 우연하게 높게 나올 수도 있으므로 향상도(Lift)를 잘 관찰 할 필요가 있다.

연관성 규칙의 대표적인 Apriori 알고리즘은 k-항목집합을 결정하기 위해 k-1 항목 집합을 서로 결합(join step)하여 후보 항목 집합(candidate k-itemsets)을 생성한다. 생성된 후보 항목 집합을 대상으로 지지도와 신뢰도가 계산되며, 여기서 분석자가 지정한 최소 지지도 및 신뢰도 기준보다 적은

값을 가지는 항목 집합은 가지치기(prune step)되고 전체 항목 집합에 대해 이 과정을 반복하여 최종적으로 강한 연관성을 가진 연관성 규칙을 생성하게 된다.

3. 카이제곱 통계량과 신뢰도의 관계

3.1 기본 가정

범주형 자료에서 두 개의 범주형 변수간에 연관성(association)이 존재하는지의 여부를 알고자 하는 경우 독립성 검정(test of independence)이 적용된다. 본 논문에서는 다음과 같은 2×2 분할표(contingency table)를 고려한다.

<표 1> 본 연구에서 사용된 2×2 분할표

		Y		합
		1	0	
X	1	a	$x_1 - a$	x_1
	0	$y_1 - a$	$t - (x_1 + y_1)$	$+ ax_0 = t - x_1$
합		y_1	$y_0 = t - y_1$	t

여기서

t : 전체 트랜잭션의 수

x_1 : 항목 집합 X의 총 발생 빈도 수

y_1 : 항목 집합 Y의 총 발생 빈도 수

a : 항목 집합 X와 Y의 동시 발생 빈도 수

$x_0 = t - x_1$: 항목 집합 X가 전체 트랜잭션에서 발생하지 않은 빈도 수

$y_0 = t - y_1$: 항목 집합 Y가 전체 트랜잭션에서 발생하지 않은 빈도 수

이다. 또한 각 셀은 식(3.1)의 조건을 만족해야 한다.

$$\left.\begin{array}{l} 0 \le a \le t \\ 0 \le x_1 - a \le t \\ 0 \le y_1 - a \le t \\ 0 \le t - (x_1 + y_1) + a \le t \\ 0 \le a \le x_1 \\ 0 \le a \le y_1 \end{array}\right\} \qquad (3.1)$$

조건(3.1)을 정리하면 다음과 같다.

$$\left.\begin{array}{l} 0 \le a \le x_1 \\ 0 \le a \le y_1 \\ (x_1 + y_1) - t \le a \le x_1 + y_1 \end{array}\right\} \qquad (3.2)$$

<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 항목 집합 X와 Y의 동시 발생 빈도인 a에 따라 다른 셀 들은 위의 <표 1>과 같이 나타낼 수 있다. <표 1>을 대상으로 한 독립성 검정은 표본의 크기가 클 경우 근사적으로 $\chi^2(1)$의 분포를 근사적으로 사용할 수 있으며, 검정통계량으로 피어슨의 카이제곱 통계량을 사용하였다. Cochran(1954)에 의해 제안된 연속성의 수정을 한 카이제곱 통계량 및 Fisher의 정확한 검정(Fisher's exact test)은 데이터 마이닝의 대상 데이터가 주로 대용량 데이터베이스를 대상으로 한 분석이므로, 표본의 크기가 크다고 가정하여 피어슨의 카이제곱 통계량을 사용하였다.

본 연구에서는 t, x_1, y_1을 고정시키고, 항목 집합 X와 Y의 동시 발생 빈도 a와 독립성 검정의 기각 한계 값의 관계를 알아보고, 동시 발생 빈도 a와 신뢰도의 관계를 살펴본 다음, 최종적으로 독립성 검정의 기각 한계 값과 최소 신뢰도 기준값의 관계를 알아봄으로써, 기존에 분석자가 임의로 설정했던 최소 지지도 및 최소 신뢰도 기준값을 객관적인 방법으로 제시하고자 한다.

3.2 동시 발생 빈도와 피어슨 카이제곱 통계량

<표 1>의 데이터에 대한 피어슨의 카이제곱 통계량은 다음과 같이 표현된다.

$$\begin{aligned}\chi^2 &= \sum_{i,j}\frac{(E_{ij}-O_{ij})^2}{E_{ij}} \\ &= \left(\frac{t}{x_1y_1(t-x_1)(t-y_1)}\right)\left(t^2a^2-2tx_1y_1a+(x_1y_1)^2\right)\end{aligned} \tag{3.3}$$

a에 대한 2차항의 계수를 α_2, 1차항의 계수를 α_1, 상수항을 α_0라 두면 식(3.3)은 다음과 같이 표현된다.

$$\chi^2 = \alpha_2 a^2 + \alpha_1 a + \alpha_0 \tag{3.4}$$

여기서

$$\alpha_2 = \frac{t^3}{x_1y_1(t-x_1)(t-y_1)}$$

$$\alpha_1 = \frac{-2t^2}{(t-x_1)(t-y_1)}$$

$$\alpha_0 = \frac{tx_1y_1}{(t-x_1)(t-y_1)}$$

이다. 식(3.4)에 조건(3.2)를 적용하면 다음과 같은 조건이 성립한다.

$$\left.\begin{array}{l}\alpha_2 \geq 0 \\ \alpha_1 \leq 0 \\ \alpha_0 \geq 0\end{array}\right\} \tag{3.5}$$

<표 1>의 분할표에서 독립성 검정에서의 귀무가설과 대립가설은 다음과 같다.

$$\begin{array}{l} H_0 : p_{ij} = p_{i\cdot} p_{\cdot j} \, (i = 1, 2 \, ; j = 1, 2) \\ H_1 : H_0\text{가 아니다}. \end{array} \tag{3.6}$$

여기서

p_{ij} = 범주 x_i 와 범주 y_j 에 동시에 속할 확률

$p_{i\cdot}$ = 범주 x_i 에 속할 확률

$p_{\cdot j}$= 범주 y_j 에 속할 확률

이다. 귀무가설 H_0를 기각할 수 있는 a의 범위를 찾기 위해서는 다음의 부등식을 풀면 된다.

$$\alpha_2 a^2 + \alpha_1 a + \alpha_0 \geq \chi_\alpha^2(1) \tag{3.7}$$

여기서 $\chi_\alpha^2(1)$는 유의 수준 α이고, 자유도가 1인 카이제곱 통계량 값이다. a_L을 a의 하한값, a_U을 a의 상한값이라 두면, a_L과 a_U의 값은 다음과 같다.

$$a_L = \frac{-\alpha_1 - \sqrt{\alpha_1^2 - 4\alpha_2(\alpha_0 - \chi_\alpha^2(1))}}{2\alpha_2}$$

$$a_U = \frac{-\alpha_1 + \sqrt{\alpha_1^2 - 4\alpha_2(\alpha_0 - \chi_\alpha^2(1))}}{2\alpha_2}$$

따라서 식(3.7)의 부등식을 만족하는 a의 범위는 다음과 같다.

$$a \leq a_L \, , \, a \geq a_U \tag{3.8}$$

두 항목 집합 X와 Y가 상호 연관성을 갖기 위해 귀무가설 H_0를 기각할 수 있는 a의 범위는 조건(3.5)에서 2차항의 계수 $\alpha_2 \geq 0$이므로, 조건(3.8)과 a 의 제한 조건인 조건(3.2)를 동시에 만족해야 한다.

3.3 동시 발생 빈도와 신뢰도

아래 <표 2>는 <표 1>의 데이터를 대상으로 각각의 셀에 대하여 연관성 규칙에서 사용하는 지지도, 신뢰도, 향상도를 식으로 표현한 것이다. 2×2 분할표 에서는 총 8개의 지지도, 신뢰도, 향상도를 가질 수 있으며, 편의상 ⓐ ⓗ로 구분하여 표시하고, 연관 규칙 $R:X \Rightarrow Y$에 대한 지지도, 신

뢰도, 향상도를 각각 $Sup(X \Rightarrow Y)$, $Conf(X \Rightarrow Y)$, $Lift(X \Rightarrow Y)$ 로 표시하였다.

ⓐⓑⓖⓗ 영역에서는 두 항목 집합 X와 Y가 동시에 발생한 빈도수 a가 증가함에 따라 지지도와 신뢰도, 향상도가 단조 증가하는 1차 함수의 형태를 보이고 있으며, ⓒⓓⓔⓕ 영역에서는 a 가 증가함에 따라 각각 단조 감소하는 1차 함수의 형태를 보이고 있다.

각각 ⓐ ⓗ 영역에 표시된 각각의 연관성 규칙 $R: X \Rightarrow Y$ 은 $Sup(X \Rightarrow Y) \geq min\_sup$ 및 $Conf(X \Rightarrow Y) \geq min\_conf$ 인 강한 연관 규칙을 만족하기 위해 이를 만족하는 최소의 a (영역 ⓐⓑⓖⓗ) 및 최대의 a (영역 ⓒⓓⓔⓕ) 값을 정할 수 있다. ⓐⓑⓖⓗ 영역에서는 각각 지지도와 신뢰도가 a 에 대해 단조 증가 1차 함수이므로 a 의 값이 커질수록 지지도와 신뢰도가 커져서 강한 연관성 규칙을 만족하고, ⓒⓓⓔⓕ 영역에서는 각각 지지도와 신뢰도가 a 에 대해 단조 감소 1차 함수이므로 a 의 값이 작아질 수록 지지도와 신뢰도가 커져서 강한 연관성 규칙을 만족할 수 있지만, 두 항목 집합 X와 Y가 서로 연관성을 갖기 위해서는 귀무가설을 기각해야 하므로 조건(3.8)과 조건(3.2)에 의해 a 의 범위가 제한되므로 강한 연관 규칙을 가지는 최소 지지도 기준과 최소 신뢰도 기준을 정할 수 있다.

<표 2> 2×2 분할표를 이용한 지지도, 신뢰도, 향상도

$X \Rightarrow Y$	ⓐ	$Y \Rightarrow X$	ⓑ
$Sup(X \Rightarrow Y) = \frac{a}{t}$ $Conf(X \Rightarrow Y) = \frac{a}{x_1}$ $Lift(X \Rightarrow Y) = \frac{ta}{x_1 y_1}$		$Sup(Y \Rightarrow X) = \frac{a}{t}$ $Conf(Y \Rightarrow X) = \frac{a}{y_1}$ $Lift(Y \Rightarrow X) = \frac{ta}{x_1 y_1}$	
$X \Rightarrow \sim Y$	ⓒ	$\sim Y \Rightarrow X$	ⓓ
$Sup(X \Rightarrow \sim Y) = \frac{x_1 - a}{t}$ $Conf(X \Rightarrow \sim Y) = \frac{x_1 - a}{x_1}$ $Lift(X \Rightarrow \sim Y) = \frac{t(x_1 - a)}{x_1(t - y_1)}$		$Sup(\sim Y \Rightarrow X) = \frac{x_1 - a}{t}$ $Conf(\sim Y \Rightarrow X) = \frac{x_1 - a}{t - y_1}$ $Lift(\sim Y \Rightarrow X) = \frac{t(x_1 - a)}{x_1(t - y_1)}$	
$\sim X \Rightarrow Y$	ⓔ	$Y \Rightarrow \sim X$	ⓕ
$Sup(\sim X \Rightarrow Y) = \frac{y_1 - a}{t}$ $Conf(\sim X \Rightarrow Y) = \frac{y_1 - a}{t - x_1}$ $Lift(\sim X \Rightarrow Y) = \frac{t(y_1 - a)}{y_1(t - x_1)}$		$Sup(Y \Rightarrow \sim X) = \frac{y_1 - a}{t}$ $Conf(Y \Rightarrow \sim X) = \frac{y_1 - a}{y_1}$ $Lift(Y \Rightarrow \sim X) = \frac{t(y_1 - a)}{y_1(t - x_1)}$	
$\sim X \Rightarrow \sim Y$	ⓖ	$\sim Y \Rightarrow \sim X$	ⓗ
$Sup(\sim X \Rightarrow \sim Y) = \frac{t - (x_1 + y_1) + a}{t}$ $Conf(\sim X \Rightarrow \sim Y) = \frac{t - (x_1 + y_{1)} + a}{t - x_1}$ $Lift(\sim X \Rightarrow \sim Y) = \frac{t(t - (x_1 + y_1) + a)}{(t - x_1)(t - y_1)}$		$Sup(\sim Y \Rightarrow \sim X) = \frac{t - (x_1 + y_1) + a}{t}$ $Conf(\sim Y \Rightarrow \sim X) = \frac{t - (x_1 + y_{1)} + a}{t - y_1}$ $Lift(\sim Y \Rightarrow \sim X) = \frac{t(t - (x_1 + y_1) + a)}{(t - x_1)(t - y_1)}$	

* : not을 의미한다.

3.4 카이제곱 통계량과 신뢰도

조건(3.8)과 조건(3.2)를 동시에 만족하는 a를 이용하여 카이제곱 통계량과 <표 2>의 영역 ⓐ에서의 신뢰도 $Conf(X \Rightarrow Y)$와의 관계를 알아보자. 편의상 $Conf(X \Rightarrow Y)$ 를 $Conf_{xy}$ 라 표시하였다. ⓐ영역에서 $Conf_{xy} = \frac{a}{x_1}$ 이고, 여기서 a는 두 항목 집합 X와 Y가 상호 독립이 아닐 범위인 조건(3.8)과 조건(3.2)를 동시에 만족해야 한다.

$a = x_1 Conf_{xy}$ 를 식(3.3)에 대입하여 $Conf_{xy}$ 에 대해 전개하면 다음과 같다.

$$\begin{aligned} \chi^2 &= \frac{t}{x_1 y_1 (t - x_1)(t - y_1)} (t^2 x_1^2 \; Conf_{xy}{}^2 - 2t x_1 y_1 x_1 Conf_{xy} + x_1^2 y_1^2) \\ &= \frac{t x_1}{y_1 (t - x_1)(t - y_1)} (t^2 \; Conf_{xy}{}^2 - 2t y_1 Conf_{xy} + y_1^2) \end{aligned} \tag{3.9}$$

$Conf_{xy}$ 에 대한 2차항의 계수를 β_2, 1차항의 계수를 β_1, 상수항을 β_0라 두면 식(3.9)는 다음과 같이 표현된다.

$$\chi^2 = \beta_2 \; Conf_{xy}{}^2 + \beta_1 Conf_{xy} + \beta_0 \tag{3.10}$$

여기서

$$\beta_2 = \frac{t^3 x_1}{y_1 (t - x_1)(t - y_1)}$$

$$\beta_1 = \frac{-2 \; t^2 x_1}{(t - x_1)(t - y_1)}$$

$$\beta_0 = \frac{t y_1}{(t - x_1)(t - y_1)}$$

이다. 식(3.10)에 조건(3.2)를 적용하면 다음과 같은 조건이 성립한다.

$$\beta_2 \geq 0, \; \beta_1 \leq 0, \; \beta_0 \geq 0 \tag{3.11}$$

귀무가설을 기각하기 위한 조건은 식(3.10)이 $X_\alpha^2(1)$ 보다 크거나 같아야 하므로 다음의 부등식을 풀면 된다.

$$\beta_2 \; Conf_{xy}{}^2 + \beta_1 Conf_{xy} + \beta_0 \geq X_\alpha^2(1) \tag{3.12}$$

여기서 $\chi_\alpha{}^2(1)$는 유의 수준 α이고, 자유도가 1인 카이제곱 통계량이다.

$Conf_{xy_L}$ 을 $Conf_{xy}$ 의 하한값, $Conf_{xy_U}$ 를 $Conf_{xy}$ 의 상한값이라 두면, $Conf_{xy_L}$ 와

$Conf_{xy_U}$ 의 값은 다음과 같다.

$$Conf_{xy_L} = \frac{-\beta_1 - \sqrt{\beta_1^2 - 4\beta_2(\beta_0 - \chi_\alpha^2(1))}}{2\beta_2}$$

$$Conf_{xy_U} = \frac{-\beta_1 + \sqrt{\beta_1^2 - 4\beta_2(\beta_0 - \chi_\alpha^2(1))}}{2\beta_2}$$

따라서 식(3.12)의 부등식을 만족하는 $Conf_{xy}$ 의 범위는 다음과 같다.

$$Conf_{xy} \le Conf_{xy_L}, \; Conf_{xy} \ge Conf_{xy_U} \tag{3.13}$$

두 항목 집합 X와 Y가 상호 연관성을 갖기 위해 귀무가설을 기각할 수 있는 $Conf_{xy}$ 의 범위는 조건(3.11)에서 2차항의 계수 $\beta_2 \ge 0$이므로, 조건(3.13)과 a의 제한 조건인 조건(3.2)에 대해 $Conf_{xy} = a/x_1$ 을 동시에 만족해야 한다.

위의 결과로부터 ⓐ영역에서 귀무가설을 기각할 수 있는 최소 신뢰도 기준을 정할 수 있으며, ⓑ ⓗ영역에서도 동일한 방법으로 귀무가설을 기각할 수 있는 최소 신뢰도 기준을 정할 수 있다.

4. 모의 실험

본 절에서는 제3절에서 논의된 결과들을 통합하여, 다음과 같이 두 가지 모의 실험을 실시하였다. 첫 번째 실험은 각각 t, y_1, x_1 을 모두 고정한 다음, 귀무가설을 기각할 수 있는 a 의 범위를 구하고, 이를 통해 최소 신뢰도 기준을 정했고, 두 번째 실험은 t, y_1 만을 고정한 다음 x_1, a 를 적절한 비율로 변화시키면서, 귀무가설을 기각할 수 있는 최소 신뢰도 기준을 살펴보았다.

유용한 규칙의 해석을 위해 항목 집합 X, Y에 대해 다음과 같이 가정하였다. 먼저 데이터베이스에 있는 총 트랜잭션의 수(t)를 100명으로 하고, 항목 집합 X는 구매한 컴퓨터의 금액을 기준으로 300만원 이상(1) 구매한 사람수를 90명으로 하고 300만원 미만(0)을 구매한 사람수를10명으로 하였다. 또한 항목 집합 Y를 결제 방식을 기준으로 신용 카드로 결제(1)한 사람수를 25명으로 하고 신용 카드 이외의 방법으로 결제(0)한 사람의 수를 75명으로 하였다. 항목 집합 X와 Y가 동시에 발생한 빈도 수, 즉 300만원 이상의 컴퓨터를 구매하면서 신용카드로 결제한 빈도 수는 a명으로 하였다. 이를 정리하면 <표 3>과 같다.

<표 3> 모의 실험 데이터

		Y		합
		1	0	
X	1	a	$90-a$	90
	0	$25-a$	$a-15$	10
합		25	75	100

조건(3.2)에서 $t=100, x_1=90, y_1=25$ 를 적용하여 a가 취할 정수 값의 범위를 정하면,

$$15 \le a \le 19 \tag{4.1}$$

또한 식(3.4)에서 $\alpha_2, \alpha_1, \alpha_0$ 는 각각 다음과 같이 계산된다.

$$\alpha_2 = \frac{t^3}{x_1 y_1 (t - x_1)(t - y_1)} = \frac{100^3}{90 \times 25 \times (100 - 90) \times (100 - 25)} = 0.5926$$

$$\alpha_1 = \frac{-2t^2}{(t - x_1)(t - y_1)} = \frac{-2 \times 100^2}{(100 - 90)(100 - 25)} = -26.6667$$

$$\alpha_0 = \frac{t x_1 y_1}{(t - x_1)(t - y_1)} = \frac{100 \times 90 \times 25}{(100 - 90) \times (100 - 25)} = 300$$

유의 수준 $\alpha = 0.05$ 하에서 $\chi^2(1) = 3.84146$ 이고, a의 하한값과 상한값은 각각 다음과 같다.

$$a_L = \frac{-\alpha_1 - \sqrt{\alpha_1^2 - 4\alpha_2(\alpha_0 - \chi_\alpha^2(1))}}{2\alpha_2} = \frac{-(-26.6667) - \sqrt{19.095}}{2 \times 0.5926} = 19.955$$

$$a_U = \frac{-\alpha_1 + \sqrt{\alpha_1^2 - 4\alpha_2(\alpha_0 - \chi_\alpha^2(1))}}{2\alpha_2} = \frac{-(-26.5557) + \sqrt{9.095}}{2 \times 0.5926} = 25.044$$

그러므로 귀무가설 H_0 를 기각할 수 있는 a 의 범위는

$$a \le 19.955,\ a \ge 25.044 \tag{4.2}$$

조건(4.1)과 조건(4.2)에서 a는 정수값만 허용되므로 다음과 같다.

$$15 \le a \le 19 \tag{4.3}$$

즉, 두 항목 집합 X와 Y는 a가 $15 \le a \le 19$ 일 경우에 상호 독립적이지 않다는 결론에 이른다. 조건(4.3)을 <표 1>의 ⓐ 영역에 적용하면 다음과 같다.

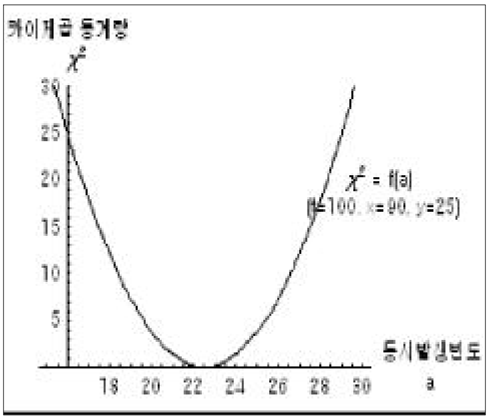
$$\frac{15}{100} \le \mathrm{Sup}(X \Rightarrow Y) = \frac{a}{t} \le \frac{19}{100}$$

$$\frac{15}{90} \le \mathrm{Conf}(X \Rightarrow Y) = \frac{a}{x_1} \le \frac{19}{90}$$

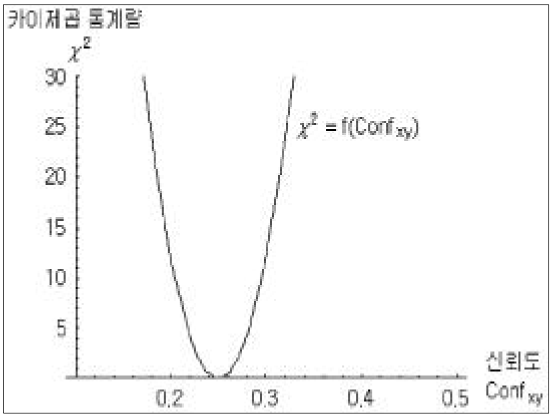
$$\frac{100 \times 15}{90 \times 25} \le \mathrm{Lift}(X \Rightarrow Y) = \frac{ta}{x_1 y_1} \le \frac{100 \times 19}{90 \times 25}$$

여기서 우리는 연관성 규칙 $R: X \Rightarrow Y$ 가 강한 연관 규칙으로 판단되기 위해서는 $a = 15$ 일 때, 최소 지지도 기준 $min\_sup = 15\%$ 이고, 최소 신뢰도 기준 $min\_conf = 16.7\%$ 임을 알 수 있다.

또한 향상도는 $Lift = 0.667$ 임을 알 수 있다. 다시 말하면 <표 2>의 첫 번째 모의 실험 데이터 표와 같이 주어졌을 때 ⓐ영역에서 두 항목 집합 X 와 Y과 강한 연관 규칙을 가지는 조건은 $a = 15$ 일 때, 최소 신뢰도 기준 $min\_conf = 16.7\%$ 로 정해야 함을 알 수 있다.

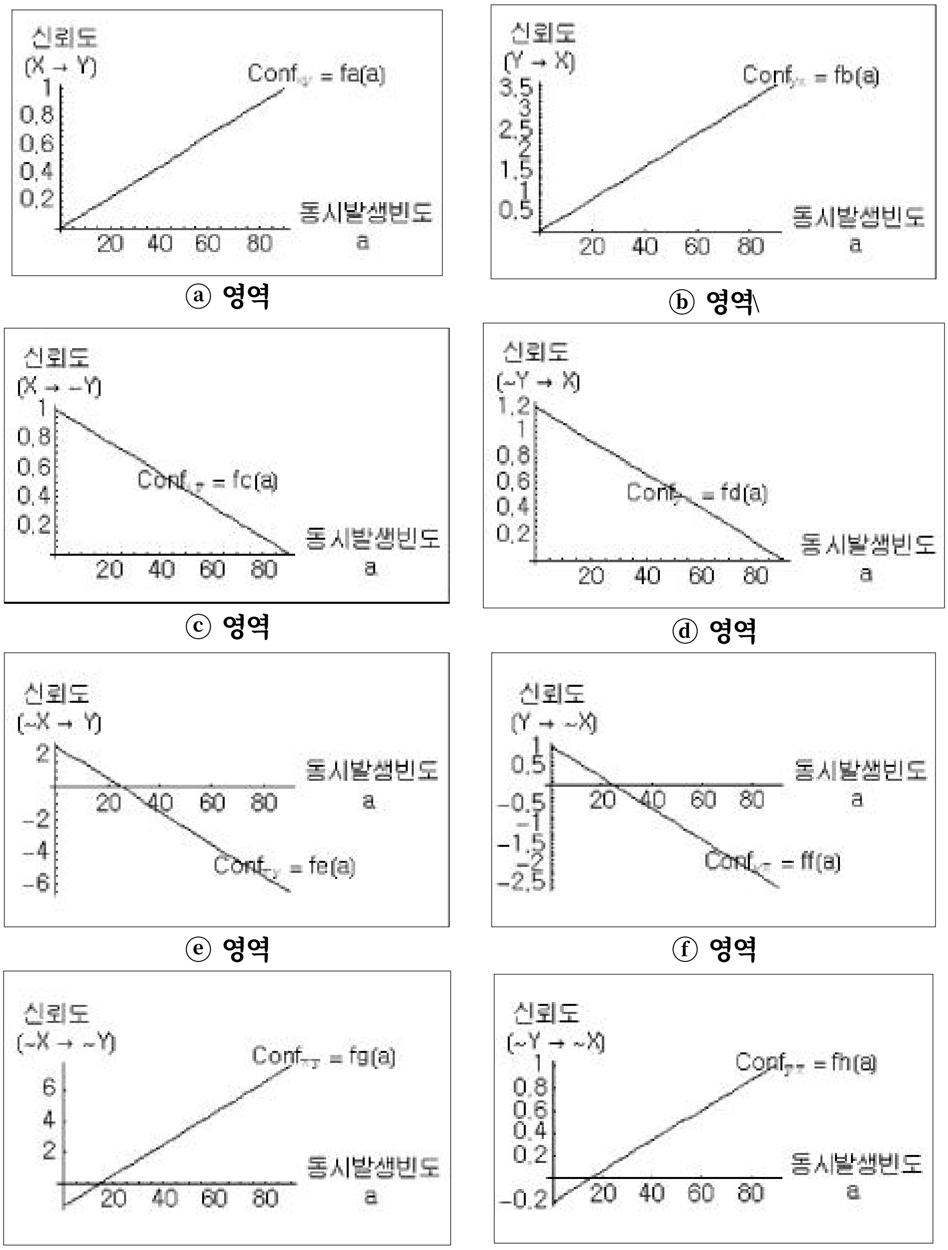


<그림 1> 동시 발생 빈도에 따른 카이제곱 통계량의 변화



<그림 2> ⓐ영역에서 신뢰도에 따른 카이제곱 통계량의 변화

<그림 1>과 <그림 2>는 <표 3>을 이용하여 각각 동시 발생 빈도 a 및 ⓐ영역에서 신뢰도에 따른 카이제곱 통계량의 변화를 표현한 것이다. 또한 <그림 3>는 <표 3>을 이용하여 동시 발생 빈도 a에 따른 각각 ⓐ ⓗ 영역에서 신뢰도의 변화를 표현한 것이다.



ⓐ 영역

ⓑ 영역\

ⓒ 영역

ⓓ 영역

ⓔ 영역

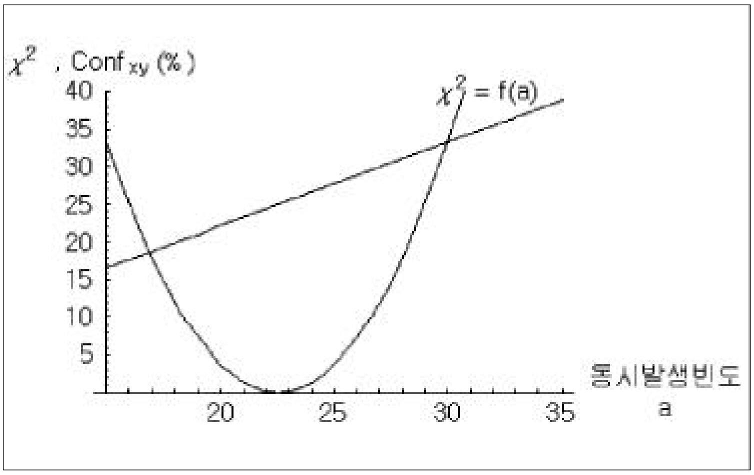
ⓕ 영역

ⓖ 영역

ⓗ 영역

<그림 3> 동시 발생 빈도에 따른 신뢰도의 변화

<그림 4>는 <표 3>을 이용하여 ⓐ영역에서 동시 발생 빈도 a에 따른 신뢰도와 카이제곱 통계량의 변화를 함께 표현한 것이고, 여기서 신뢰도는 백분율(%)로 표시하였다.



<그림 4> 동시발생 빈도에 따른 카이제곱 통계량과 신뢰도의 변화 (ⓐ 영역)

5. 결론 및 향후 과제

연관성 규칙은 탐색적이며, 비목적성 분석이며, 기존의 데이터를 특별한 변형 없이 계산이 용이하게 사용 가능하며, 연관 정도를 수치화 할 수 있어서 상호 비교가 가능하다는 장점 때문에 데이터 마이닝에서 자주 사용되는 방법이다.

본 연구에서는 연관성 규칙에서 연관 기준값으로 사용하고 있는 최소 신뢰도 기준을 카이제곱 통계량과의 관계를 통해 강한 연관 규칙을 가질 때의 최소 신뢰도 기준을 정하는 방법을 제안하였다. 모의 실험 결과 정해진 유의수준 하에서 독립성 검정의 귀무가설을 기각하는 영역에 대해 최소 신뢰도 기준을 통계적으로 결정할 수 있었고, 이는 분석자가 임의로 정한 최소 신뢰도 기준값보다는 객관적임을 확인하였다.

참고문헌

[1] 노맹석 (2000). 통계적 기법을 이용한 데이터 마이닝, *J.Basic Sci. Res. Kyungsan Univ*. 4(1), P 91-98.

[2] Han, J., Kamber M. (2001). Data Mining : *Concepts and Techniques*, Morgan Kaufmann Publishers.

[3] Reynolds, H. T. (1977). *Analysis cf Nominal Data, Quantitative Applications in the Social Sciences,* Series No. 7, Sage Publications.

[4] Stephen, E. F. (1976). *The Analysis cf Cross-Classfied Categorical Data*, Second Edition.

[5] Cochran, W. G. (1954), Some methods for strengthening the common χ^2 tests, *Biometrics*, 10.

[6] 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.

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

[8] Bing, L., Wynne, H., Yiming, M. (1999). Mining Association Rules with Multiple minimum Supports, *Proceedings of ACM KDD*-99.

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

[10] Cheung, D. W., Han, J., Ng, V., 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.

[11] Sergey, B., Rajeev, M., Jeffrey, D. U., Shalom, T. (1997). Dynamic itemset counting and implication rules for market data, *Proceedings of ACM SIGMOD Conference on Management of Data*.

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

[13] 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*.

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

[15] Cheung, D. W., Han, J., Ng, V., Fu, A. W., Fu, Y. (1996). A Fast distribution algorithm for mining association rules, *Int's Conference on Parallel and Distributes Information System*, Miami Beach, Florida.

[16] Han, J., Pei, J. (2000). Mining Frequent Patterns by Pattern-Growth : Methodology and Implications, *SIGKDD Explorations*.

[17] Hipp, J., Guntzer, U., Gholamreza (2000). Algorithms for Association Rule Mining - A General Survey and Comparison, *SIGKDD Explorations*, Vol. 2, Issue 1, P 58 - 64.

[18] Saygin, Y., Vassilios, S. V., Clifton, C. (2002). Using Unknowns to Prevent Discovery of Association Rules, *2002 Conference on Research Issues in Data Engineering*.

[19] John, F. R., Rice, S. (2001). What's Interesting About Cricket? - On Thresholds and Anticipation in Discovered Rules, *SIGKDD Explorations*, Vol.3 Issue 1, P 1-5.

[20] Khalil, M. A., Nagwa, M. E, Taha, Y. (2000). *A note on 'Beyond Market Baskets: Generalizing Association Rules to Correlations'*.