J. of Basic Sciences, Cheju Nat'l Univ.
16(2), 75~81, 2003

기초과학연구 제주대학교
16(2), 75~81, 2003

연관 규칙 마이닝 안에서 오즈비 사용

강 형 창 · 김 철 수 · 이 봉 규
제주대학교 자연과학대학 전산통계학과

요 약

데이터 마이닝(data mining)은 방대한 양의 데이터로부터 유용한 정보를 추출하는 과정이다. 연관 규칙 마이닝(association rule mining)은 데이터 마이닝에서 사용하는 여러 방법들 중의 하나이다. 연관 규칙 마이닝은 둘 또는 그 이상의 항목들 사이에 존재하는 수많은 연관 규칙 중에서 지지도(support), 신뢰도(confidence) 그리고 향상도(lift)에 근거한 연관 규칙을 일반화하는 규칙 탐색 방법이다. 연관 규칙은 신뢰도가 높으면 좋으나, 신뢰도가 크다고 모두 최선의 연관 규칙은 아니며, 지지도가 어느 정도 이상일 경우에 의미가 크다. 본 논문에서는 2×2 분할표에서 두 항목집합(itemsets)간에 연관성을 찾아내는 방법으로 오즈비(odds ratio)라는 연관성 측도를 사용한다. 향상도를 오즈비를 통하여 계산 한 후 두 항목집합간에 강한 연관 규칙을 찾아내기 위한 최소 지지도(minimum support)를 계산한다.

1. 서 론

연관 규칙 마이닝(association rule mining)은 데이터 항목집합 사이에서 유용한 연관성을 찾는 방법이다. 이때 데이터 항목집합 사이에는 수많은 연관 규칙들이 존재하게 되는데, 연관 규칙들 중에서 유용한 연관성을 찾아내기 위해 지지도(support), 신뢰도(confidence), 그리고 향상도(lift)에 근거하여 일반화하는 규칙 탐색 방법이다.

현재까지 알려진 연관 규칙 마이닝 알고리즘들은 agrawal 등(1994)이 제안한 후보 항목집합들을 구성하고 후보 항목집합들의 발생 빈도 수를 계산하여 사용자가 정의한 최소 지지도를 기초로 빈발 항목집합을 결정하는 Apriori 알고리즘을 비롯하여, Park 등(1995)은 데이터베이스를 중복되지 않는 크기로 분할한 후 한번에 한 개의 분할 영역만을 고려하여 그 안에서 빈발 항목집합을 생성하는 Partition 알고리즘을 제안하였으며, Toivonen(1996)는 무작위로 선정된 표본을 이용하여 빈발 항목집합 찾은 후 그 결과를 데이터베이스의 나머지 부분에 적용하여 증명하는 방법인 Sampling 알고리즘을 제안하였다. 그리고 Cheung 등(1996)은 갱신된 데이터베이스에서는 이전에 빈발 항목을 다루었던 항목집합은 데이터베이스 스캔을 생략하는 FUP 알고리즘을 연구하였고, Liu 등(1999)은 후보 항목집합들을 효율적으로 작게 구하여 이를 기초로 전체 트랜잭션의 크기와 개수를 줄이는 방법인 DHP (Direct Hashing and Pruning)알고리즘을 제안하였다. 이들 연구들은 주로 대용량 데이터베이스에서 효율적인 연관성을 찾기 위한 연구로서 효율적인 연관성을 찾기 위해서 후보 항목집합 구성 시간과 후보 항목집합의 크기 및 빈발 항목집합을 구성하기 위한 시간을 줄이는 알고리즘을 중심으로 연구되어졌다. 한편, Silverstein 등(1997)은 연관 규칙에서 지지도와 신뢰도를 계산하는 경우에 두 항목집합이 발생하지 않는 경우는 고려하지 않고, 두 항목집합이 동시에 발생하는 경우만 고려함으로써 발생할 수 있는 문제를 지적하면서 연관 규칙에서 카이제곱 통계량의

사용을 제안하였다.

기존의 연구에서는 후보 항목집합에서 빈발 항목집합을 찾기 위해 최소 지지도(minimum support)와 최소 신뢰도(minimum confidence)를 사용자가 임의로 지정하거나 전체 데이터베이스를 스캔하여 항목간의 지지도 및 신뢰도를 계산한 후 상위에 위치하는 지지도 및 신뢰도를 두 항목간의 연관성을 판단하는 기준으로 삼고 있기 때문에 동일 데이터에 대해서도 분석자마다 연관 규칙이 서로 다른 결과가 나타날 수 있다. 한편, Park 등(2002)은 연관 규칙에서 기준이 되는 최소 신뢰도를 카이제곱 통계량을 이용하여 통계적인 관점에서 파악하고 객관적인 연관 규칙의 연관 기준값을 제안하였다. 본 논문에서는 연관성을 파악할 수 있는 측도인 오즈비를 이용하여 연관 규칙을 통계적으로 접근하고, 강한 연관 규칙의 기준이 되는 향상도를 오즈비를 통하여 제안 하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 연관 규칙에 대하여 설명하고, 3장에서는 오즈비와 향상도의 관계를 유도하고, 4장에서는 적용 예를 통하여 오즈비와 향상도의 관계를 설명하였다. 마지막 5장에서는 결론을 다룬다.

2. 연관 규칙

연관 규칙은 항목집합으로 표현되는 트랜잭션에서 각 항목간의 연관성 정도를 반영하는 규칙으로 미리 결정되어진 최소 지지도를 만족하는 빈발 항목집합을 찾아내어 연관 규칙을 생성한다.

2.1 기본 개념

k개로 이루어진 항목들의 집합 $I=\{i_1, i_2, \cdots, i_k\}$이 주어지면, 트랜잭션 T는 I의 부분집합으로 정의된다($T \subseteq I$). 이때, 각 트랜잭션들은 중복된 항목을 허용하지 않으며, TID라 불리는 고유한 트랜잭션 아이디를 갖는다. 만일 트랜잭션 T가 X의 모든 항목들을 포함한다면($X \subseteq T$), T가 항목집합 X를 지지한다(support)고 한다.

항목 A와 B가 각각 $A \subseteq T$, $B \subseteq T$ 그리고, $A \cap B = \emptyset$이며 또한, $B \neq \emptyset$을 만족하는 경우 연관 규칙 $A \Rightarrow B$는 다음과 같이 정의된다.

- 지지도(support): $\text{support}(A \Rightarrow B) = P(A \cup B)$
- 지지도 개수(support frequency): 전체 트랜잭션수×지지도
- 신뢰도(confidence):
  $$\text{confidence}(A \Rightarrow B) = P(B|A) = \frac{P(A \cap B)}{P(A)}$$
- 신뢰도 개수(confidence frequency): 전체 트랜잭션수×신뢰도
- 향상도(lift):
  $$\text{lift}(A \Rightarrow B) = \frac{P(B|A)}{P(B)} = \frac{P(A \cap B)}{P(A)P(B)}$$

연관 규칙 $A \Rightarrow B$[support = s% , confidence = c%]는 항목 A와 B를 동시에 포함하는 트랜잭션이 s%이고, A가 포함된 트랜잭션 중에서 A와 B가 동시에 포함된 트랜잭션이 c%가 됨을 의미한다. 연관 규칙은 이러한 지지도와 신뢰도를 계산하여, 미리 지정한 최소 지지도 임계치(minimum support threshold)와 최소 신뢰도 임계치(minimum cofidence threshold)를 모두 만족하는 규칙을 가지는 두 항목집합을 강한(strong) 연관 규칙을 갖는 것으로 판단한다. 그리고, 지지도와 신뢰도의 값은 0과 1사이의 값보다 0%와 100%사이의 값으로 표현한다.

모든 연관 규칙들을 구성하기 위해서는 항목집합들 중에서 모든 빈발 항목집합들의 지지도를 먼저 계산하고, 그들로부터 주어진 신뢰도를 바탕으로 실제의 연관 규칙을 구성한다. 연관 규칙에서 사용되는 용어들은 다음과 같다.

- 트랜잭션(transaction): 발생된 데이터를 저장하는 단위이며, 여러 가지 항목들을 가질 수 있다.
- 항목집합(itemset): 각 개별 트랜잭션에 포함된 단일 항목 또는 복수 항목의 집합
- 후보 항목집합(candidate itemset): 개별 항목집합의 결합을 통해 생성된 항목집합으로써 후보 항목집합에 대한 지지도 및 신뢰도를 계산하여 최소 지지도 및 최소 신뢰도를 만족하

는 경우 빈발 항목집합으로 간주한다.

빈발 항목집합(frequent itemset): 후보 항목집합에서 최소 지지도 및 최소 신뢰도를 만족하는 항목집합

연관 규칙의 해석은 위와 같이 최소 지지도와 최소 신뢰도 개념을 이용한다. 그러나 신뢰도의 값이 클수록 좋으나 신뢰도가 크다고 모두 강한 연관 규칙을 갖는 것은 아니며, 두 항목 집합의 지지도가 어느 정도 수준 이상의 경우만 고려해야 하며 또한, 신뢰도와 지지도는 자주 발생되는 항목집합에 대해서는 연관성 외에 우연성에 의해 높은 결과가 나타날 수 있기 때문에 향상도를 잘 관찰해야 한다.

3. 오즈비와 향상도의 관계

3.1 기본 가정

범주형 자료에서 두 개의 범주형 변수간에 연관성(association)이 존재하는지의 여부를 알고자하는 경우 독립성 검정 이외에 오즈비라는 연관성 측도가 있다. 본 논문에서는 다음과 같은 2×2 분할표(contingency table)를 고려한다.

〈표 1〉 2×2 분할표

		B		합
		1	0	
A	1	n_{11}	n_{12}	$n_{1.}$
	0	n_{21}	n_{22}	$n_{2.}$
합		$n_{.1}$	$n_{.2}$	T

본 논문에서 사용되는 기호는 다음과 같다.

T: 전체 트랜잭션의 수

n_{11}: 항목 집합 A와 B의 동시 발생 빈도 수

$n_{1.}$: 항목 집합 A의 총 발생 빈도 수

$n_{.1}$: 항목 집합 B의 총 발생 빈도 수

$n_{2.}$: 항목 집합 A가 전체 트랜잭션에서 발생하지 않는 빈도 수

$n_{.2}$: 항목 집합 B가 전체 트랜잭션에서 발생하지 않는 빈도 수

한 트랜잭션에서 구매한 항목들의 양은 고려하지 않으며, 각 항목은 구매 여부만을 나타내는 이진 변수이다. 또한, 분할표의 각 셀들은 다음의 조건을 만족한다. 이때 항목 집합 A와 B의 연관 규칙 A⇒B을 탐색하기 위해서는 $B \neq \varnothing$의 필요하다.

$$\begin{cases} 0<n_{11}<T \\ 0<n_{1.}-n_{11}<T \\ 0<n_{.1}-n_{11}<T \\ 0<T-(n_{1.}+n_{.1})+n_{11}<T \\ 0<n_{11}<n_{1.} \\ 0<n_{11}<n_{.1} \end{cases} \tag{3.1}$$

〈표 1〉에서 T, $n_{1.}$, $n_{.1}$은 단 한번의 데이터베이스 스캔으로 알 수 있으므로, 사전에 이미 알려져 있다고 가정할 수 있다. T, $n_{1.}$, $n_{.1}$이 알려져 있는 경우 항목 집합 A와 B의 동시 발생 빈도인 n_{11}에 따라 〈표 1〉과 같이 나타낼 수 있다. 본 연구에서는 T, $n_{1.}$, $n_{.1}$을 고정시키고, 항목 집합 A와 B의 동시 발생 빈도 수(지지도 개수) n_{11}와 향상도의 관계를 설명한다.

3.2 향상도와 오즈비

항목집합 A와 B가 발생하는 경우를 "성공"이라 하자. 항목 집합 A와 B의 성공확률을 π_1, π_2라 하면, 항목 집합 A의 성공할 오즈 $\text{odds}_1 = \pi_1/(1-\pi_1)$이고, 항목집합 B의 성공 오즈 $\text{odds}_2 = \pi_2/(1-\pi_2)$이다. 두 항목집합의 오즈값의 비율

$$\theta = \frac{\text{odds}_1}{\text{odds}_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} \tag{3.2}$$

를 오즈비(odds ratio)라 한다. 행과 열이 바뀌어도 오즈비는 달라지지 않는다. 따라서 오즈비는 다음과 같이 정의할 수 있다.

$$\theta = \frac{\text{odds}_1}{\text{odds}_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} \tag{3.3}$$

여기서,

$\pi_{11}=\frac{n_{11}}{T}$, $\pi_{12}=\frac{n_{1.}-n_{11}}{T}$, $\pi_{21}=\frac{n_{.1}-n_{11}}{T}$,

$\pi_{12}=\frac{T+n_{11}-(n_{1.}-n_{.1})}{T}$.

그리고, 두 항목집합 A와 B의 성공 오즈가 같을 때($\pi_1=\pi_2$) $odds_1=odds_2$를 만족하고, 이 경우 항목집합 A와 B는 연관성이 존재하지 않는다고 한다.

오즈비는 $0<\theta<\infty$의 값을 가지게 되며, $\theta=1$이면 두 항목집합 A, B간에는 연관성이 존재하지 않으며 $0<\theta<1$이면 B⇒A의 연관성이 존재하며, $1<\theta<\infty$이면 A⇒B인 연관성이 존재한다. 〈표 1〉에서 연관 규칙 A⇒B와 B⇒A가 존재하기 위한 조건은 $\theta\neq1$이므로 항목집합 A와 B가 동시에 발생하는 최소 지지도 개수 n_{11}은 다음과 같은 두 가지로 나타낼 수 있다.

I. $1<\theta<\infty$의 조건(A⇒B 연관 규칙)

$$\theta=\frac{n_{11}\times[T-(n_{1.}+n_{.1})+n_{11}]}{(n_{1.}-n_{11})\times(n_{.1}-n_{11})} \tag{3.4}$$

에서 A⇒B 연관 규칙이 존재하기 위해서는

$$n_{11}\times[T-(n_{1.}+n_{.1})+n_{11}]> (n_{1.}-n_{11})\times(n_{.1}-n_{11}) \tag{3.5}$$

$$n_{11}>\frac{n_{1.}\times n_{.1}}{T} \tag{3.6}$$

이 된다.

(3.6)식에서 연관성이 존재한다면 $\frac{n_{11}\times T}{n_{1.}\times n_{.1}}>1$

이 되며, $\frac{n_{11}\times T}{n_{1.}\times n_{.1}}$는 향상도 lift(A⇒B)와 같다. 즉, 항목집합 A와 B가 동시에 발생하는 빈도 수(지지도 개수)가 최소한 $\frac{n_{1.}\times n_{.1}}{T}$ 보다 크면 연관 규칙 A⇒B가 존재하게 된다.

II. $0<\theta<1$의 조건(B⇒A 연관 규칙)

식 (3.4)에서 B⇒A 연관 규칙이 존재하기 위해서는

$$n_{11}\times[T-(n_{1.}+n_{.1})+n_{11}]< (n_{1.}-n_{11})\times(n_{.1}-n_{11}) \tag{3.7}$$

$$n_{11}<\frac{n_{1.}\times n_{.1}}{T} \tag{3.8}$$

이 된다.

(3.8)식에서 연관성이 존재한다면 $\frac{n_{11}\times T}{n_{1.}\times n_{.1}}<1$

이 되며, $\frac{n_{11}\times T}{n_{1.}\times n_{.1}}$는 향상도 lift(B⇒A)와 같다. 즉, 항목집합 A와 B가 동시에 발생하는 빈도 수(지지도 개수)가 최대한 $\frac{n_{1.}\times n_{.1}}{T}$ 보다 작으면 연관 규칙 B⇒A가 존재하게 된다.

연관성의 측도인 오즈비를 통하여 향상도 A⇒B를 쉽게 계산 할 수 있고, 연관성의 방향 또한 알 수 있게 된다.

4. 적 용

본 장에서는 3장에서 논의한 결과로 다음의 2×2 분할표를 이용하여 실험을 하였다. 먼저, 다음과 같이 가정하였다. 총 트랜잭션 수(T)는 1000, 항목집합 A는 k개의 항목집합에서 $\{I_1\}$을 구입한 빈도 수로 600이라 하고 구입하지 않은 빈도 수는 400이라 하였다. 그리고, 항목집합 B는 k개의 항목집합에서 $\{I_2\}$를 구입한 빈도 수로 488이라 하고 구입하지 않은 빈도 수는 512이라 하였다. 이를 2×2 분할표로 표시하면 다음과 같다.

〈표 2〉 2×2 분할표 -실험 데이터-

		B		합
		1	0	
A	1	n_{11}	$600-n_{11}$	600
	0	$488-n_{11}$	$1000-(600+488)+n_{11}$	400
합		488	512	1,000

조건 (3.1)에서 동시 발생 빈도 수 n_{11}이 취할

값의 범위는 $88<n_{11}<488$이 되고, 식 (3.4)에서 연관 규칙 A⇒B가 존재하기 위한 n_{11}의 값을 정하면, $n_{11} > \frac{600\times488}{1000} \fallingdotseq 292.8$이므로 $292 < n_{11} < 488$가 되고, 연관 규칙 B⇒A가 존재하기 위해서 n_{11}이 취할 값을 정하면, $n_{11} < \frac{600\times488}{1000} \fallingdotseq 292.8$이므로 $88 < n_{11} \leq 292$이 된다.

오즈비 $\theta \neq 1$이 아니면, 두 항목집합 A와 B 사이에는 연관성이 존재하게 되고, 향상도 값을 계산할 수 있게 된다. 향상도는 동시 발생 빈도 수에 의해 결정이 되기 때문에 향상도 값을 결정하기 위한 동시 발생 빈도 수를 계산 할 수 있게 된다.

연관 규칙 A⇒B가 강하게 존재한다는 것은 향상도 A⇒B가 높다는 것을 의미한다. 이때 향상도 A⇒B는 연관성의 측도인 오즈비를 이용하여 표현가능하며 동시 발생 빈도 수(지지도 개수)가 결정되면 오즈비를 통하여 향상도를 쉽게 계산할 수 있다. 두 항목간에 연관성이 강하게 존재한다는 것은 두 항목간의 향상도가 높다는 것을 의미하므로 두 항목간에 오즈비를 계산하여 최소 동시 발생 빈도 수(최소 지지도 개수)를 결정하게 된다면 의미있는 연관 규칙을 생성할 수 있다.

5. 결 론

수 많은 연관 규칙에서 의미 있는 연관 규칙을 찾는 것은 최소 지지도와 최소 신뢰도에 의해서 결정된다. 이러한 최소 지지도와 최소 신뢰도는 분석자(또는 사용자)에 의해 결정되거나, 전체 데이터베이스를 스캔하여 항목들간의 지지도 및 신뢰도를 계산한 후, 상위에 위치하는 지지도 및 신뢰도를 이용하여 두 항목집합간의 연관성을 판단하는 근거로 삼고 있기 때문에 분석자마다 연관 규칙이 서로 다르게 나타날 수 있다. 향상도는 신뢰도와 지지도에 의해 구성되기 때문에 향상도가 높다는 것은 신뢰도와 지지도가 높다는 것과 동일하므로, 신뢰도를 관리함으로써 신뢰도 및 지지도를 관리할 수 있다는 것과 동일하게 생각할 수 있다. 이에 본 논문에서는 2×2분할표 연관 규칙에서 연관 규칙 A⇒B이 강하다는 것은 향상도 A⇒B가 높다는 것과 같으므로, 연관성의 측도인 오즈비를 통하여 향상도를 계산하게 되면, 사용자가 결정하는 향상도에 따른 최소 지지도와 최소 신뢰도를 계산할 수 있게 된다.

오즈비의 계산으로 동시 발생 빈도 수(지지도 개수)가 결정되면 오즈비를 통하여 향상도를 계산할 수 있게 되며, 분석자가 원하는 향상도를 얻기 위한 최소 동시 발생 빈도 수(또는 최소 지지도) 및 최소 신뢰도를 계산할 수 있게 된다. 향후 연구과제는 2×2분할표를 확장한 $2\times n$분할표 또는 2×2×2분할표에서 오즈비 또는 주변 오즈비를 이용하여 향상도와 최소 지지도 및 최소 신뢰도를 계산하는 방법에 대해 논의될 것이다.

참고문헌

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

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

[3] 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.

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

[5] Park, J.S., Chen, M.S., and Philips, S.Y. (1995). An effecvie hash-based algorithms for mining association rules, Proceedings of ACM SIGMOD conference on Management of Data.

[6] Park, H.C., 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.

[7] Silverstein. C.. Bin. S.. Motwani. R. (1997). Beyond Market Baskets: Generalizing Association Rules to Dependence Rules. Data Mining and Knowledge Discovery. No.2 P 39-68.

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

The use of odds ratio for association rule mining

Hyung Chang Kang · Chul Soo Kim · Bong Kyu Lee
Department of Computer Science and Statistics, Cheju National University

Abstract

Data mining is the discovery of interesting information among huge amounts of data. Association rule mining finds interesting association among a large set of data itemsets. Interesting association is strong association rule equals. Strong association rule is that satisfy both a minimum support threshold and minimum confidence threshold. Even if confidence higher, association rule is not best, support must be at least satisfy. In this article we consider the association in 2×2 contingency table. Odds ratio is used to find association rule. To get the range of the lift value we calculate the odds ratio two itemsets are occurred simultaneously, and find the minimum support threshold values.