

4장

연관분석(Association Analysis)

4.1 연관분석(Association Analysis)의 기본개념

▶ 대량의 데이터

- ▶ 정보화 사회가 급진전하면서 국가나 기관, 기업들은 대량의 데이터 매일 축적
- ▶ 대량의 데이터를 잘 분석하면 해당 조직에 많은 도움을 줄 수 있는 정보를 구할 수 있음
- ▶ 예) 대형 슈퍼마켓에서 매일 고객들이 구매하는 물품들에 대한 정보

[표 4.1] 한 슈퍼마켓의 5명의 고객이 구입한 상품 목록 데이터

고객 번호	상품명
1	{사과, 빵, 버터, 계란}
2	{우유, 빵, 버터, 콜라}
3	{사과, 우유}
4	{사과, 우유, 빵, 버터}
5	{사과, 우유, 빵, 콜라}

- ▶ 고객마다 구입한 상품의 수와 종류가 다른데 한 고객이 구입한 상품 모두에 대한 컴퓨터 처리를 하였을 때 : 트랜잭션(transaction) 데이터를 하나 처리하였다.
- ▶ 장바구니분석(market basket analysis) : 고객들이 어떠한 상품의 구매에 관심이 많은지 또는 구매할 때 서로 연관이 있는 상품이 있는지에 대한 분석 -> 결과를 활용하여 고객의 구매 가능성을 높이기 위해 상품의 진열을 조정, 마케팅 전략에 활용, 재고관리 또는 고객관계경영(customer relation management) 등에 이용

▶ 연관규칙을 찾는 연관분석

- ▶ 연관규칙의 예 : {빵} → {버터} (빵을 사는 고객들이 버터도 구입할 가능성이 많음)
- ▶ 연관규칙 정보를 이용하여 두 상품을 동시에 한 곳에 진열하는 판매 전략을 세울 수 있음
- ▶ 슈퍼마켓 예제에서 모든 상품들의 항목집합 : $I = \{i_1, i_2, \dots, i_m\}$
- ▶ 모든 고객 트랜잭션들의 집합 : $T = \{t_1, t_2, \dots, t_N\}$
- ▶ 각 상품의 구매여부를 1과 0으로 표시하여 이항데이터로 표현 (상품의 개수나 가격에 대한 정보는 불포함)

[표 4.2] 한 슈퍼마켓의 5명의 고객이 구입한 상품의 이항 데이터 표시

고객 번호	사과(i_1)	우유(i_2)	빵(i_3)	버터(i_4)	계란(i_5)	콜라(i_6)
t_1	1	0	1	1	1	0
t_2	0	1	1	1	0	1
t_3	1	1	0	0	0	0
t_4	1	1	1	1	0	0
t_5	1	1	1	0	0	1

- ▷ X 와 Y 를 서로 공통원소가 없는 항목들의 집합 (즉, $X \cap Y = \emptyset$)이라 할 때 연관규칙 : $X \rightarrow Y$
- ▷ $n(X)$: 전체 트랜잭션 중에서 항목집합 X 를 포함하는 트랜잭션의 수

$$n(X) = |\{t_i | X \subseteq t_i, t_i \in T\}|$$

- ▷ 전체 트랜잭션의 수가 N 일 때, 연관규칙 $X \rightarrow Y$ 의 평가

(1) 지지도

- $s(X \rightarrow Y) = \frac{n(X \cup Y)}{N}$
- 전체 트랜잭션에서 연관규칙에 해당하는 데이터의 비율
- 지지도가 낮은 연관규칙은 여러 트랜잭션에 적용되지 않은 규칙, 우연히 발생할 수 있는 규칙 또는 흥미가 없는 규칙일 가능성이 많음
- 최소지지도(minimum support, minsupport)를 정하여 그 이하의 규칙은 버림
- 좋은 규칙을 찾거나 흥미 없는 규칙을 버릴 때의 기준으로 사용

(2) 신뢰도: $c(X \rightarrow Y) = \frac{n(X \cup Y)}{n(X)}$

- 항목 X 를 포함하는 트랜잭션 중에서 항목 Y 도 포함하는 트랜잭션의 비율
- 연관규칙의 신뢰성에 대한 측도
- $X \rightarrow Y$ 가 있을 때, 신뢰도가 높을수록 항목 X 를 포함하는 트랜잭션은 항목 Y 도 포함할 가능성이 많음
- 최소신뢰도(minconfidence)를 정하여 그 이하의 규칙은 버림

연관규칙 탐색 원칙
지지도 \geq minsupport, 신뢰도 \geq minconfidence인 모든 가능한 연관규칙을 찾는다.

▶ 예제 4.1.1 연관규칙 {우유, 빵} \rightarrow {버터}의 지지도와 신뢰도

- ▷ 지지도 : {우유, 빵} \cup {버터} = {우유, 빵, 버터}를 모두 포함하고 있는 트랜잭션은 고객2, 고객4로써 지지도는 $2/5=0.4$
- ▷ 신뢰도 : {우유, 빵}을 포함하고 있는 트랜잭션은 고객2, 고객4, 고객5으로 신뢰도는 $2/3=0.67$

▶ 효율적인 탐색 알고리즘

- ▷ 모든 가능한 연관규칙을 찾기 위한 경우의 수가 너무 많아서, 효율적인 탐색 알고리즘에 대한 연구가 지속
- ▷ 연관규칙 $X \rightarrow Y$ 의 지지도와 신뢰도는 $n(X \cup Y)$ 와 관련, 이 숫자가 작으면 연관규칙이 될 가능성 거의 없음

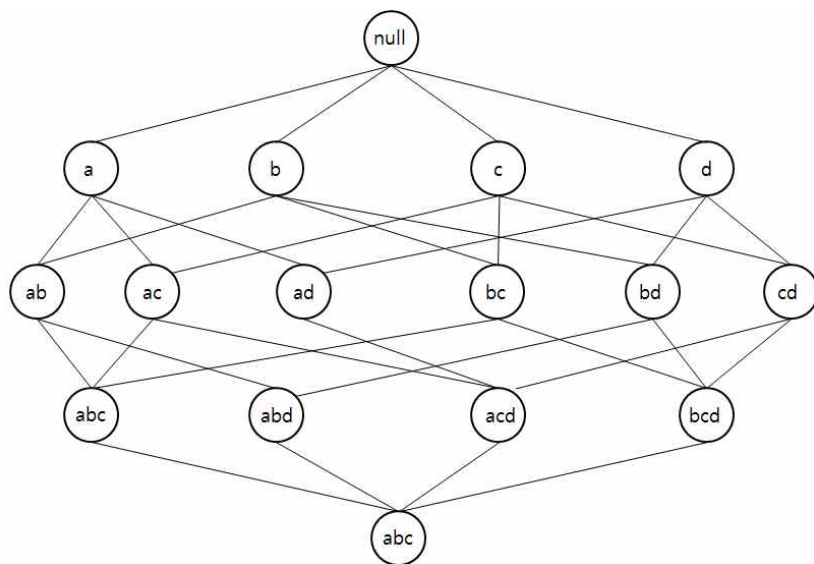
연관규칙 탐색 전략
단계 1 : 빈발항목집합의 생성(Frequent Item Set Generation)
최소지지도(minsupport)를 만족하는 모든 빈발항목집합을 찾는다
단계 2 : 연관규칙의 생성(Rule Generation)
단계 1에서 찾은 빈발항목집합에서 최소신뢰도(minconfidence)가 높은 모든 연관규칙을 찾는다

4.2 이항 데이터의 연관분석 - 선형적 알고리즘(Apriori Algorithm)

▶ 빈발항목집합

- ▷ 최소지지도 이상을 갖는 항목집합
- ▷ 트랜잭션에 나타나는 모든 항목들의 집합 : $I = \{i_1, i_2, \dots, i_m\}$
- ▷ 모든 가능한 부분집합의 수는 공집합을 제외하고 $M = 2^m - 1$
- ▷ 예) 네 원소 항목 집합 $I = \{a, b, c, d\}$ 의 모든 가능한 항목집합은 $2^4 - 1 = 15$

[그림 4.1] 네 원소 항목집합에 대한 모든 가능한 항목집합의 격자 형태 그림



1-항목집합 : $\{a\}, \{b\}, \{c\}, \{d\}$ 와 같이 한 항목으로 이루어진 집합

2-항목집합 : $\{a, b\}, \{a, c\}, \{a, d\}, \{b, c\}, \{b, d\}, \{c, d\}$

3-항목집합 : $\{a, b, c\}, \{a, b, d\}, \{a, c, d\}, \{b, c, d\}$

4-항목집합 : $\{a, b, c, d\}$

k -항목집합 : 일반적으로 k 개의 항목으로 이루어진 집합

- ▷ 원시적으로 연관규칙을 찾기 위해서 모든 가능한 부분집합에 대해 전체 트랜잭션에 대한 지지도를 계산하여야 한다. 즉, 각각의 트랜잭션에서 모든 가능한 항목들의 부분집합이 있는지 일일이 비교하여 빈도수를 조사하고, 이 중에서 어떠한 부분집합이 빈발항목인지 알아보아야 한다.
- ▷ N 을 전체 트랜잭션의 수라 하고 w 를 한 트랜잭션의 최대 항목수라 할 때, 모든 가능한 부분집합의 지지도를 계산하는 작업은 (NMw) 에 비례하여 많은 비교가 필요. 항목의 수 m 은 실제 응용에서 매우 클 수 있는데, m 에 지수적으로 비례해서 증가하는 부분집합의 수 $M = 2^m - 1$ 때문에 전체 지지도 계산은 더욱 많은 비교를 필요로 한다. 이와 같은 원시적인 빈발항목집합 생성방법의 문제점을 줄이기 위해서 다음과 같은 두 가지 접근방식에 의해 여러 가지 연구가 진행

(1) 모든 가능한 항목집합의 수(M)를 줄이는 방식

- 선형적(apriori)인 규칙을 이용하여 지지도 계산이 불필요한 항목집합을 제거하는 방식(4.2.1절)

(2) 비교하는 수를 줄이는 방식

- 자료구조를 효율적으로 하여 각각의 항목집합을 각각의 트랜잭션과 일일이 비교하는 수를 줄이는 방식

4.2.1 선형적 규칙(Apriori Principle)

정리 4.1 빈발항목집합 추출의 선형적 규칙(Apriori principle)

(1) 한 항목집합이 빈발하다면 이 항목집합의 모든 부분집합은 역시 빈발항목집합이다.


(2) 한 항목집합이 비빈발하다면 이 항목집합을 포함하는 모든 집합은 비빈발항목집합이다.

▶ 선형적 규칙의 예

- ▷ 모든 항목들의 집합을 $I = \{a, b, c, d\}$, 만약 $\{b, c, d\}$ 가 빈발항목집합이라면 이 항목의 부분집합 $\{b, c\}, \{b, d\}, \{c, d\}, \{b\}, \{c\}, \{d\}$ 는 빈발항목집합이 됨.
- ▷ 만약, $\{a, b\}$ 가 최소지지도 기준을 넘지 못한 비빈발항목집합이라면 이 집합을 포함하는 $\{a, b, c\}, \{a, b, d\}, \{a, b, c, d\}$ 는 빈발항목집합이 될 수 없음.
- ▷ 최소지지도 기준을 넘지 못하는 항목집합들을 쉽게 가지치기 가능

▶ 선형적 알고리즘을 이용한 빈발항목집합의 생성

- (1) 1-항목집합에 대하여 지지도를 계산 (1-항목집합의 지지도가 최소지지도를 넘지 못하면 모두 버림)
- (2) 최소지지도를 넘는 1-항목 빈발항목집합을 이용하여 불필요한 2-항목집합에 대한 가지치기
- (3) 2-항목 빈발항목집합 후보 생성
- (4) 이 빈발항목집합 후보들 중에서 최소지지도 기준을 넘는 2-항목 빈발항목집합을 확정하고, 이들을 이용하여 불필요한 3-항목집합에 대한 가지치기
- (4) 3-항목 빈발항목집합 후보 생성

 같은 방법을 k-항목 빈발항목집합을 확정할 때까지 반복

▶ 선형적 알고리즘의 특징

- ▷ 1-항목집합, 2-항목집합, ..., k-항목집합 등 각 수준별(level-wise)로 접근하는 방식
- ▷ 각 수준에서 빈발항목집합 후보를 생성, 이 후보들이 지지도를 만족하는지 시험

[알고리즘 4.1] 선형적 알고리즘

```

단계 1:  $k = 1$ ,  $I = \text{전체항목집합}$ 
2:  $F_1 = \text{모든 1-항목 빈발항목집합}$ 
3: repeat
4:    $k = k + 1$ 
5:    $C_k = \text{apriori-gen}(F_{k-1})$   $\{(k-1)\text{-빈발항목집합으로 } k\text{-빈발항목집합 후보 생성}\}$ 
6:   for each transaction  $t \in T$  do
7:      $C_t = \text{subset}(C_k, t)$   $\{\text{트랜잭션에 있는 모든 } k\text{-빈발항목집합 후보를 찾음}\}$ 
8:     for each candidate itemset  $c \in C_t$  do
9:        $n(c) = n(c) + 1$   $\{\text{지지도를 증가}\}$ 
10:    end for
11:  end for
12:   $F_k = \{c \mid c \in C_k \text{ and } n(c) \geq N \times \text{minsupport}\}$   $\{k\text{-항목 빈발항목집합 추출}\}$ 
13: until  $F_k = \emptyset$ 
14: 전체 빈발항목집합 =  $\cup F_k$ 

```

- ▷ 알고리즘에서 C_k : k -빈발항목집합 후보, F_k : k -빈발항목집합
- ▷ 먼저 1-항목집합 전체에 대하여 지지도를 계산하고 모든 1-항목 빈발항목집합 F_1 을 찾는다 (단계1~2)
- ▷ 알고리즘은 반복적으로 $(k-1)$ -빈발항목집합 F_{k-1} 을 이용하여 k -빈발항목집합 후보 C_k 생성(단계5)
- ▷ 빈발항목집합 후보 C_k 의 각 부분집합에 대하여 모든 트랜잭션을 조사해 지지도를 계산(단계6~10)
- ▷ 최소지지도(minsupport)를 넘지 못하는 빈발항목집합 후보를 제거, k -항목 빈발항목집합 F_k 를 확정 (단계12)
- ▷ 새로운 빈발항목집합이 안 나올 때까지($F_k = \emptyset$) 반복한다 (단계 13)

- ▶ 예제 4.2.1 선형적 알고리즘으로 지지도가 60% 이상인 빈발항목집합을 구하라.

1-항목집합	도 수	비 고
{사과}	4	
{우유}	4	
{빵}	4	
{버터}	3	
{계란}	1	비빈발
{콜라}	2	비빈발



2-항목집합	도 수	비 고
{사과, 우유}	3	
{사과, 빵}	3	
{사과, 버터}	2	비빈발
{우유, 빵}	3	
{우유, 버터}	2	비빈발
{빵, 버터}	3	



3-항목집합	도 수	비 고
{사과, 우유, 빵}	2	비빈발

[그림 4-3] [표 4.1] 데이터에 대한 선형적 알고리즘 적용 예

- ▶ [표 4.1]의 데이터는 5개의 트랜잭션, 최소지지도 60%를 정하면 각 집합을 포함하고 있는 트랜잭션이 3개 이상 되어야 빈발항목집합. 1-항목집합에 대해 도수를 계산해 보면 {계란}과 {콜라}가 각각 1과 2로서 비빈발항목집합, 1-항목 빈발항목집합 $F_1 = \{\text{사과, 우유, 빵, 버터}\}$
- ▶ 1-항목 빈발항목집합들로서 2-항목 빈발항목집합들의 후보를 생성, 1-항목 빈발항목집합 후보의 수는 ${}_4C_2 = 6$, $C_2 = \{\{\text{사과, 우유}\}, \{\text{사과, 빵}\}, \{\text{사과, 버터}\}, \{\text{우유, 빵}\}, \{\text{우유, 버터}\}, \{\text{빵, 버터}\}\}$
- ▶ {사과, 버터}, {우유, 버터}의 도수가 2로 비빈발항목집합으로 제거, 2-항목 빈발항목집합 $F_2 = \{\{\text{사과, 우유}\}, \{\text{사과, 빵}\}, \{\text{우유, 빵}\}, \{\text{빵, 버터}\}\}$
- ▶ 2-항목 빈발항목집합에서 3항목 빈발항목집합 후보를 생성할 경우 비빈발 2-항목집합이 들어 있으면 3-항목 빈발항목집합의 후보가 될 수 없다. 3-항목 빈발항목집합 후보 $C_3 = \{\{\text{사과, 우유, 빵}\}\}$
- ▶ 3-항목 빈발항목집합 후보는 지지도가 2이므로 빈발항목집합이 될 수 없다. 즉, $F_3 = \emptyset$

4.2.2 선형적 알고리즘에서 빈발항목집합 후보의 생성

- ▶ 효율적인 C_k 를 생성하는 방법

- ▶ 전체 항목의 수가 m 일 때 모든 가능한 k -항목 빈발항목집합 후보들의 수는 ${}_m C_k$, k -항목까지 가능한 빈발항목집합 후보의 수는 ${}_m C_1 + {}_m C_2 + \dots + {}_m C_k$
- > 모든 경우의 수에 대하여 지지도를 계산하여 비빈발항목집합 후보를 제거하는 원시적인 방법은 수 많은 계산이 필요

- ▷ 선형적 알고리즘은 $(k-1)$ -항목 빈발항목집합 F_{k-1} 을 이용하여 k -항목 빈발항목집합 후보 C_k 를 생성, 하지만 모든 가능한 후보를 생성만 하는 것이 아니라 불필요한 후보는 가지치기를 하여 제거
- ▷ 선형적 규칙으로 $(k-1)$ -항목 빈발항목집합 F_{k-1} 을 이용하여 k -항목 빈발항목집합 후보 C_k 를 생성하는 방법에도 여러 가지가 있음. (1) $(k-1)$ -항목 빈발항목집합 F_{k-1} 각각에 1-항목 빈발항목집합 F_1 을 추가하여 조사하는 $F_{k-1} \times F_1$ 이다. 하지만 이 방법은 동일한 후보 집합이 여러번 생성될 수 있는 문제점

2-항목 빈발항목집합 F_2		1-항목 빈발항목집합 F_1		3-항목 빈발항목집합 후보 C_3
{사과, 우유}	\times	{사과}	\Rightarrow	{사과, 우유, 빵}
{사과, 빵}		{우유}		{사과, 우유, 버터}
{우유, 빵}		{빵}		{사과, 빵, 버터}
{빵, 버터}		{버터}		{우유, 빵, 버터}

[그림 4-4] [표 4.1] 데이터에 대한 $F_2 \times F_1$ 적용 예

▶ 가. $F_{k-1} \times F_{k-1}$ 빈발항목집합 후보 생성방법

- ▷ 이 방법에서는 $(k-1)$ -항목 빈발항목집합에서 처음 $(k-2)$ 항목이 같은 항목들만 혼합하여 k -항목 빈발항목집합 후보를 생성

2-항목 빈발항목집합 F_2		2-항목 빈발항목집합 F_2		3-항목 빈발항목집합 후보 C_3
{사과, 우유}	\times	{사과, 우유}	\Rightarrow	{사과, 우유, 빵}
{사과, 빵}		{사과, 빵}		
{우유, 빵}		{우유, 빵}		
{빵, 버터}		{빵, 버터}		

[그림 4-5] [표 4.1] 데이터에 대한 $F_2 \times F_2$ 적용 예

- ▷ C_3 를 생성하기 위해 2-항목 빈발항목집합 중 첫 항목이 {사과}인 {사과, 우유}와 {사과, 빵}만을 혼합
- ▷ 이 방법에서는 {사과, 우유}와 {우유, 빵}를 혼합할 필요는 없다.
- ▷ $(k-1)$ -항목 빈발항목집합 $A = \{a_1, a_2, \dots, a_{k-1}\}$ 와 $B = \{b_1, b_2, \dots, b_{k-1}\}$ 가 있을 때 두 집합은 다음 조건을 만족할 때만 혼합하게 되어 불필요한 후보 생성을 줄일 수 있다.

$$a_i = b_i \quad (i = 1, 2, \dots, k-2), \quad a_{k-1} \neq b_{k-1}$$

- ▷ 단, 생성된 후보의 $(k-2)$ -항목 부분집합이 모두 빈발항목집합인지는 조사

4.2.3 빈발항목집합을 이용한 연관규칙 생성

▶ 효율적으로 연관규칙을 추출하는 방법

- ▶ 빈발항목집합 F 의 항목들을 공집합이 아닌 두 개의 서로 다른 부분집합 X 와 Y 로 나누어 ($X \cap Y = \emptyset, F = X \cup Y$) 하나의 연관규칙 $X \rightarrow Y$ 를 만들었다고 하자.
- ▶ k -항목 빈발항목집합 F 는 최대 $2^k - 2$ (\emptyset 과 전체집합 제외됨)개의 연관규칙을 만들 수 있음.
- ▶ 빈발항목집합 F 를 이용하여 만든 모든 연관규칙은 이미 최소지지도를 만족시키기 때문에 이 중에서 최소신뢰도를 만족하는 연관규칙을 찾아야 한다.

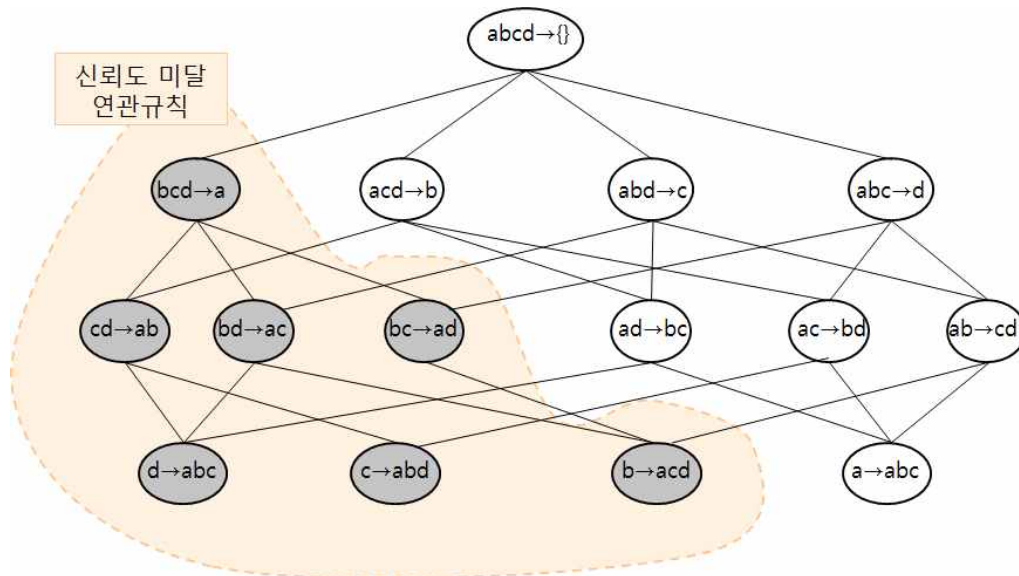
▶ 예제 4.2.2 연관규칙의 생성

- ▶ $F = \{1, 2, 3\}$ 을 빈발항목집합이라 할 때 가능한 연관규칙을 나열하라.
- ▶ 3-항목 빈발항목집합이므로 $2^3 - 2 = 6$ 개의 연관규칙이 가능.
- ▶ $\{1, 2\} \rightarrow \{3\}, \{1, 3\} \rightarrow \{2\}, \{2, 3\} \rightarrow \{1\}, \{1\} \rightarrow \{2, 3\}, \{2\} \rightarrow \{1, 3\}, \{3\} \rightarrow \{1, 2\}$.

▶ 연관규칙 생성의 선험적 규칙

- ▶ 최대 $2^k - 2$ 개의 연관규칙 중 최소신뢰도를 만족시키지 않는 연관규칙은 생성할 필요 없음.
- ▶ 신뢰도에 대한 선험적 규칙을 이용하면 불필요한 연관규칙을 생성하지 않을 수 있음.

정리 4.2 연관규칙 생성의 선험적 규칙(Apriori Principle)
빈발항목집합 F 에 대하여 연관규칙 $X \rightarrow Y (X \cap Y = \emptyset, F = X \cup Y)$ 가 최소신뢰도 기준을 만족하지 않으면 X 의 어떠한 부분집합 X' 에 대한 연관규칙 $X' \rightarrow F - X'$ 도 최소신뢰도 기준을 만족할 수 없다.



[그림 4-6] 연관규칙의 생성 및 신뢰도 미달 연관규칙의 제거

- ▶ 빈발항목집합이 $\{a, b, c, d\}$ 라고 하면 연관규칙 $\{b, c, d\} \rightarrow \{a\}$ 가 원하는 신뢰도 수준을 만족하지 않는다면, $\{b, c, d\}$ 의 부분집합 $\{b, c\}, \{b, d\}, \{c, d\}, \{b\}, \{c\}, \{d\}$ 를 포함하는 연관규칙들은 정리에 의해 최소신

뢰도를 만족하지 않으므로 고려 대상에서 제외

▶ 연관규칙 생성의 선형적 알고리즘

- (1) 연관규칙의 우측 항목이 하나인 규칙들을 생성
- (2) 이 중에서 최소신뢰도를 만족하지 않은 연관규칙이 있으면 제거
- (3) [정리 4.2]에 의하여 연관규칙 좌측 항목의 모든 부분집합들을 고려되는 연관규칙 후보에서 제외
→ 최소신뢰도를 만족하는 연관규칙들은 새로운 후보규칙을 만드는데 이용

- ▶ 예) $\{a,c,d\} \rightarrow \{b\}$ 와 $\{a,b,d\} \rightarrow \{c\}$ 가 최소신뢰도를 만족한다면 두 연관규칙 결과를 섞어 $\{a,d\} \rightarrow \{b,c\}$ 연관규칙 생성

[알고리즘 4.2] 연관규칙 생성을 위한 선형적 알고리즘(Apriori Algorithm)

```
단계 1: for 각 k-항목 빈발항목집합  $f_k$ ,  $k \geq 2$  do
2:    $H_1 = \{i | i \in f_k\}$    {연관규칙의 우측항이 1-항목인 규칙}
3:   call ap-genrule( $f_k, H_1$ )   {연관규칙의 생성 프로시저}
4: end for
```

- ▶ 빈발항목집합 생성 알고리즘과 유사
- ▶ 다른 점은 후보규칙의 지지도를 계산할 필요가 없고 각 규칙의 신뢰도를 구한다는 것

[알고리즘 4.3] Procedure *ap-genrule*(f_k, H_1)

```
단계 1:  $k$  = 빈발항목집합의 크기
2:  $m$  = 연관규칙 우측 항목의 크기
3: if  $k < m + 1$  then
4:    $H_{m+1} = \text{apriori-gen}(H_m)$ 
5:   for each  $h_{m+1} \in H_{m+1}$  do
6:     confidence =  $n(f_k) / n(f_k - h_{m+1})$ 
7:     if confidence  $\geq$  minconfidence
8:       output the rule  $(f_k - h_{m+1}) \rightarrow h_{m+1}$ 
9:     else
10:      delete  $h_{m+1}$  from  $H_{m+1}$ 
11:    end if
12:  end for
13:  call ap-genrule( $f_k, H_{m+1}$ )
14: end if
```

4.3 연관규칙의 평가

- ▶ 트랜잭션의 연관분석은 수많은 연관규칙을 생성할 가능성이 많음.
- ▶ 흥미 있는 연관규칙을 가려내기 위해 일반적으로 받아들일 수 있는 연관규칙의 평가기준을 설정하는 일이 매우 중요.
- ▶ 평가기준 - 객관적 관점: 객관적 흥미측도(interestingness measure)를 적용하여 흥미없는 규칙 제거
 - 주관적 관점: 한 연관분석이 분석자에게 기대하지 않았던 정보를 준다면 이는 주관적인 관점에서 흥미 있는 규칙이 됨. 그러나 이러한 주관적 관점을 연관규칙의 평가기준으로 하기 위해서는 많은 사전연구가 필요하며, 적절한 주관적 흥미측도를 개발해야 함.

ex) {빵} → {버터} : 너무 당연히 알려져 있는 규칙.

높은 지지도와 신뢰도를 가지고 있더라도 별로 흥미가 없을 수 있음.

{기저귀} → {맥주} : 전혀 기대하지 않았던 정보를 줌

-> 마케팅 전략 수립에 흥미 있는 규칙이 될 수 있다.

4.3.1 객관적 흥미측도-두 항목의 경우

- ▶ 연관규칙 $A \rightarrow B$ 에서 A와 B가 각각 한 개의 항목으로 이루어져 있을 때 사용되는 객관적인 흥미측도
- ▶ 두 항목이 각 트랜잭션이 있는 지(=1) 없는 지(=0) 조사하여 2×2 교차표 만들.

[표 4.3] 연관규칙 $A \rightarrow B$ 를 평가하기 위한 2×2 교차표

		항목 B		행의 합
		B = 1	B = 0	
항목 A	A = 1	f_{11}	f_{10}	$f_{1.}$
	A = 0	f_{01}	f_{00}	$f_{0.}$
열의 합		$f_{.1}$	$f_{.0}$	N

- ▶ N : 전체 트랜잭션의 수
- ▶ f_{11} : 전체 트랜잭션 N 개 중에서 항목 A와 B가 동시에 나타난 트랜잭션의 수
- ▶ f_{01} : 전체 트랜잭션 N 개 중에서 항목 A는 없고 B만 있는 트랜잭션의 수
- ▶ $f_{1.}$: 항목 A가 있는 모든 트랜잭션 수의 합
- ▶ 비대칭 이항변수(asymmetric binary variable): 두 항목이 동시에 나타난 도수 f_{11} 중시.
대칭 이항변수(symmetric binary variable): 한 항목이 있고 없는 것이 동일한 비중.
- ▶ 연관규칙의 흥미측도들은 대칭적 측도와 비대칭적 측도로 구분할 수 있음.
대칭적측도(symmetric measure): $\mu(A \rightarrow B) = \mu(B \rightarrow A)$
항목집합을 평가하는데 많이 이용.
- 비대칭적측도(asymmetric measure): 신뢰도 $c(A \rightarrow B)$ 는 비대칭적 측도.
신뢰도 $c(A \rightarrow B)$ 와 $c(B \rightarrow A)$ 가 항상 같지는 않음.
연관규칙을 설명하는데 적합.
- ▶ 선형적 알고리즘에서는 연관규칙의 평가를 지지도와 신뢰도로 함. 최소지지도와 최소신뢰도를 넘지 못하는 항목집합은 빈발항목집합이 될 수 없음.

[예제 4.3.1] 100명의 사람들에게 대하여 커피와 차의 선호여부를 조사하였더니 [표 4.4]와 같다. 연관규칙 {차}→{커피}에 대한 지지도와 신뢰도를 구하고 이 연관규칙이 의미가 있는지 분석하라.

[표 4.4] 100명에 대한 연관규칙 {차}→{커피}의 2×2 교차표

		커피 B		행의 합
		B = 1	B = 0	
차 A	A = 1	15	5	20
	A = 0	65	15	80
열의 합		80	20	100

▶ 지지도 = $\frac{n(\text{차}, \text{커피})}{N} = \frac{15}{100} = 0.15$ (i.e., 15%), 신뢰도 = $\frac{n(\text{차}, \text{커피})}{n(\text{차})} = \frac{15}{20} = 0.75$ (i.e., 75%)

▶ 차를 마시는 사람의 비율이 20%밖에 되지 않기 때문에, {차}→{커피}는 신뢰도가 높음에도 불구하고 흥미 없는 연관규칙이 된다.

▶ 지지도와 신뢰도 모두를 고려해야 함.

▶ 지지도와 신뢰도를 함께 고려하는 다양한 흥미측도들이 연구됨

▶ 가장 많이 이용되는 측도: 리프트, 파이계수, IS측도, 기타 측도

가. 리프트(lift)

▶ 연관규칙 A→B의 지지도와 신뢰도를 동시에 고려하는 측도로 많이 사용.

▶ $Lift(A, B) = \frac{c(A \rightarrow B)}{s(B)}$ c(A→B): 연관규칙의 신뢰도, s(B): 항목 B의 지지도

$$Lift(A, B) = \frac{f_{11}/f_{1\cdot}}{f_{\cdot 1}/N} = \frac{Nf_{11}}{f_{1\cdot} \cdot f_{\cdot 1}} \quad [\text{표 4.3}]$$

▶ 전체 트랜잭션에서 항목 B가 출현하는 비율(지지도)에 대비하여, 항목 A를 가지고 있는 트랜잭션의 비율(신뢰도)이 얼마나 향상되었는지를 측정.

▶ A와 B가 독립인 경우,

$$\frac{f_{11}}{N} = \frac{f_{1\cdot}}{N} \times \frac{f_{\cdot 1}}{N} \Rightarrow Nf_{11} = f_{1\cdot} \times f_{\cdot 1}$$

▶ A와 B가 서로 독립: Lift=1

A와 B가 양의 상관: Lift>1

A와 B가 음의 상관: Lift<1

▶ [예제 4.3.1] Lift(차, 커피) = 0.75/0.8=0.9375: 차와 커피는 거의 독립인 경우로 약한 음의 상관을 보임.

▶ 연관규칙의 리프트는 높을수록 좋음.

[예제 4.3.2] 텍스트 마이닝

신문기사와 같은 텍스트 데이터에서는 두 관련 있는 단어가 여러 문서에 나타날 수 있다. 100개의 문서를 조사하여 단어집합 {p,q}와 {r,s}가 나타난 도수를 조사해보니 [표 4.5]와 같다. 두 단어집합의 리프트를 조사하여 비교하라.

[표 4.5] 100개의 문서를 조사하여 단어집합 {p,q}와 {r,s}가 나타난 도수의 교차표

		단어 p		행의 합
		p=1	p=0	
단어 q	q=1	88	5	93
	q=0	5	2	7
열의 합		93	7	100

		단어 r		행의 합
		r=1	r=0	
단어 s	s=1	2	5	7
	s=0	5	88	93
열의 합		7	93	100

$$Lift(p,q) = \frac{0.88}{0.93 \times 0.93} = 1.02$$

\Rightarrow {p,q}가 동시에 나타난 문서가 88%임에도 불구하고 리프트는 1에 가까워 두 단어가 서로 독립.

$$Lift(r,s) = \frac{0.02}{0.07 \times 0.07} = 4.08 \Rightarrow \{r,s\} \text{가 동시에 나타난 문서는 단지 2\%임에도 불구하고 리프트 높음.}$$

$$c(p \rightarrow q) = \frac{88}{93} = 94.6\% \quad , \quad c(r \rightarrow s) = \frac{2}{7} = 28.6\%$$

Lift측도 보다는 신뢰도 측도가 더 적절하다고 볼 수 있음.

리프트 하나만으로 연관규칙의 좋고 나쁨을 평가해서는 안 된다는 예

나. 파이계수

▶ 두 이항변수의 관련성을 측정.

$$\phi = \frac{f_{11}f_{00} - f_{01}f_{10}}{\sqrt{f_{1.} \cdot f_{.1} f_{0.} \cdot f_{.0}}} \quad , \quad -1 \leq \phi \leq 1$$

▶ $\phi = -1$: 완전 음의 상관

$\phi = 1$: 완전 양의 상관

$\phi = 0$: 두 이항변수는 독립

▶ 대칭적 이항변수에 대해 적합한 측도

[예제 4.3.2] 단어집합 {p,q}와 {r,s}의 파이계수:
$$\phi(p,q) = \frac{88 \times 2 - 5 \times 5}{\sqrt{93 \times 7 \times 93 \times 7}} = 0.232 = \phi(r,s)$$

단어집합 {p,q}가 단어집합 {r,s}보다 더 많은 문서에 나타났음에도 불구하고, 파이계수가 같은 이유는 대칭적 측도이기 때문(즉, 단어의 출현과 비출현이 동일한 비중으로 여겨짐.)

다. IS측도 (Interest-Support)

▶ 비대칭적 이항변수의 경우 적합한 측도

▶ $IS(A, B) = \sqrt{Lift(A, B) \times s(A, B)} = \frac{s(A, B)}{\sqrt{s(A)s(B)}}$: 리프트와 지지도가 모두 높을 때 큰 값 가짐.

$= \sqrt{\frac{s(A, B)}{s(A)} \times \frac{s(A, B)}{s(B)}} = \sqrt{c(A \rightarrow B) \times c(B \rightarrow A)}$: 연관규칙 $A \rightarrow B$ 와 $B \rightarrow A$ 중에서 하나의 신뢰도가 작을 때 값이 작게 됨.

▶ A, B가 독립일 때,

$IS(A, B) = \frac{s(A, B)}{\sqrt{s(A)s(B)}} = \frac{s(A)s(B)}{\sqrt{s(A)s(B)}} = \sqrt{s(A) \times s(B)}$: 두 변수가 상관이 없거나, 음의 상관을 갖더라도 커질 수 있는 문제점 있음.

라. 기타 측도

[표 4.6] 두 항목집합 {A,B}에 대한 대칭적 측도

측도	정의
오즈비(odds ratio)	$\alpha = \frac{f_{11}f_{00}}{f_{10}f_{01}}$
카파(kappa)	$\kappa = \frac{Nf_{11} + Nf_{00} - f_{1 \cdot} \cdot f_{\cdot 1} - f_{0 \cdot} \cdot f_{\cdot 0}}{N^2 - f_{1 \cdot} \cdot f_{\cdot 1} - f_{0 \cdot} \cdot f_{\cdot 0}}$
피아테츠키-샤피로(Piatetsky-Shapiro)	$PS = \frac{f_{11}}{N} - \frac{f_{1 \cdot} \cdot f_{\cdot 1}}{N^2}$
자카르트(Jaccard)	$\zeta = \frac{f_{11}}{\sqrt{f_{1 \cdot} + f_{\cdot 1} - f_{11}}}$

[표 4.7] 연관규칙 $A \rightarrow B$ 에 대한 비대칭적 측도

측도	정의
굿맨-크루스칼(Goodman-Kruskal)	$\lambda = \frac{\sum_j (\max_k f_{jk} - \min_k f_{\cdot k})}{N - \min_k f_{\cdot k}}$
상호정보(mutual information)	$M = \frac{\sum_i \sum_j \frac{f_{ij}}{N} \log \frac{Nf_{ij}}{f_{i \cdot} \cdot f_{\cdot j}}}{-\sum_i \frac{f_{i \cdot}}{N} \log \frac{f_{i \cdot}}{N}}$
J-측도(J-measure)	$J = \frac{f_{11}}{N} \log \frac{Nf_{11}}{f_{1 \cdot} \cdot f_{\cdot 1}} + \frac{f_{10}}{N} \log \frac{Nf_{10}}{f_{1 \cdot} \cdot f_{\cdot 0}}$
라플라스(Laplace)	$L = \frac{f_{11} + 1}{f_{1 \cdot} + 2}$
설득도(conviction)	$V = \frac{f_{11}f_{\cdot 0}}{Nf_{10}}$

각 측도별로 장단점이 있으므로 분석자의 판단에 의해 결정할 수밖에 없음. 이를 위해 여러 가지 측도들을 적용해 보는 것이 바람직함.

4.3.2 객관적 흥미측도 - 두 항목 이상의 경우

[표 4.8] 세 항목집합 {A, B, C}에 대한 3차원 도수분포표

		항목 B				행의 합
		B=1		B=0		
		C=1	C=0	C=1	C=0	
항목 A	A=1	f_{111}	f_{110}	f_{101}	f_{100}	$f_{1 \dots}$
	A=0	f_{011}	f_{010}	f_{001}	f_{000}	$f_{0 \dots}$
열의 합		$f_{\cdot 11}$	$f_{\cdot 10}$	$f_{\cdot 01}$	$f_{\cdot 00}$	N
		$f_{\cdot 1 \cdot}$		$f_{\cdot 0 \cdot}$		

$$f_{.1.} = f_{.11} + f_{.01}, \quad f_{.0.} = f_{.10} + f_{.00}$$

- ▶ $Lift(A, B, C) = \frac{N^2 \times f_{111}}{f_{1..} \times f_{.1.} \times f_{.1.}}$
- ▶ 세 항목이 독립이라면, $\frac{f_{111}}{N} = \frac{f_{1..}}{N} \times \frac{f_{.1.}}{N} \times \frac{f_{.1.}}{N}$
- ▶ m개의 항목집합 $\{i_1, i_2, \dots, i_m\}$ 의 도수를 f_{i_1, i_2, \dots, i_m} 로 표시할 때

$$Lift = \frac{N^{m-1} \times f_{i_1 i_2 \dots i_m}}{f_{i_1 \dots} \times f_{i_2 \dots} \times \dots \times f_{i_m \dots}}$$

m개의 항목들이 서로 독립이라면,

$$\frac{f_{i_1, i_2, \dots, i_m}}{N} = \frac{f_{i_1 \dots} \times f_{i_2 \dots} \times \dots \times f_{i_m \dots}}{N^m}$$

- ▶ 다른 방법으로는 여러 개의 항목들을 두 항목씩 쌍으로 하여 흥미측도를 구한 후 이들의 최대값, 최소값 또는 평균을 구하여 여러 항목의 측도로 하기도 함.

4.3.3 객관적 흥미측도 - 항목들의 지지도가 매우 다를 경우

▶ 대부분의 항목지지도가 매우 낮거나 중간정도, 일부 항목지지도가 매우 높은 경우의 흥미측도에 대해 알아보자.

예) 항목이 2000개 있을 때, 지지도가 1%미만인 항목이 1600개, 1%~90%인 항목이 380개, 90%이상인 항목 20개 있음: 지지도가 매우 다른 경우.

- ▶ 항목집합 $X = \{i_1, i_2, \dots, i_m\}$ 에 대해 의미 없는 연관규칙의 생성을 방지하기 위해

교차지지도(cross support) $r(X)$ 를 이용

$$r(X) = \frac{\min\{s(i_1), s(i_2), \dots, s(i_m)\}}{\max\{s(i_1), s(i_2), \dots, s(i_m)\}} : \text{최대지지도에 대한 최소지지도의 비율을 의미}$$

- ▶ $r(X)$ 가 매우 작으면 항목집합 X 에서 생성되는 연관규칙이 의미가 없을 가능성이 높음.
- ▶ 사용자가 정의한 수준을 교차지지도가 넘으면 교차지지 연관규칙이라고 함.

[예제 4.3.3] 교차지지 연관규칙

우유의 지지도가 70%이고, 설탕의 지지도가 10%, 캐비어의 지지도가 0.04%일 때 {우유, 설탕, 캐비어}의 교차지지도를 구하라.

$$r = \frac{\min\{0.7, 0.1, 0.0004\}}{\max\{0.7, 0.1, 0.0004\}} = \frac{0.0004}{0.7} = 0.00058$$

교차지지도가 낮음. {우유, 설탕, 캐비어}에서 나타나는 연관규칙은 의미 없는 규칙일 가능성이 높음.

- ▶ $h\text{-신뢰도} = \frac{s(\{i_1, i_2, \dots, i_k\})}{\max\{s(i_1), s(i_2), \dots, s(i_k)\}}$: 항목집합 X 의 지지도를 각 항목들의 최대지지도로 나눔
- ▶ $h\text{-신뢰도가 높으면}$ 항목집합에 있는 항목들 사이에는 서로 강한 연관이 있음.

4.4 범주형 및 연속형 데이터의 연관분석

4.4.1 범주형 데이터의 연관분석

[표 4.9] 범주형 데이터가 있는 인터넷 여론조사 데이터

성별	교육정도	출신지역	인터넷 쇼핑	인터넷 채팅
여	고졸	서울	함	함
남	고졸	경기	안함	함
남	대졸	대구	함	함
여	고졸	광주	함	안함
여	중졸	대전	안함	함
...

이항데이터: 성별, 범주형데이터: 교육정도, 출신지역

예) {교육정도 = 대졸} → {인터넷 쇼핑 = 함}

- ▶ 범주형 데이터의 연관분석은 이항 데이터로 변환하여 이항 데이터의 연관분석 방법을 적용.

[표 4.10] 범주형 데이터의 이항 데이터 변환

성별=남	성별=여	교육정도 =대졸	교육정도 =고졸	교육정도 =중졸	출신지역 =서울	출신지역 =...	인터넷 쇼핑=함	인터넷쇼 핑=안함	인터넷 채팅=함	인터넷채 팅=안함
0	1	0	1	0	1	...	1	0	1	0
1	0	0	1	0	0		0	1	1	0
1	0	1	0	0	0		1	0	1	0
0	1	0	1	0	0		1	0	0	1
0	1	0	0	1	0		0	1	1	0
...

이항 데이터로 변환된 범주형 데이터를 분석할 때 다음을 유의

- 1) 출신 지역같이 범주가 많은 변수는 한 변수 값의 빈도수가 매우 작을 수 있어, 최소지지도를 낮추면 너무 많은 빈발항목집합이 나타날 수 있는 문제가 있다. 해결방법으로는 지역을 중부권, 영남권, 호남권 등 유사한 변수 값으로 그룹화하거나 변수 값이 작은 것만 묶어서 하나의 범주로 만드는 방법이 있음.
- 2) 각각의 변수를 이항데이터로 변환하는 경우에는 한 변수 값이 너무 많이 나타나서 불필요한 연관규칙

이 많이 만들어질 가능성이 있는데, 이러한 경우에는 연관분석을 하기 전에 한 변수 값이 너무 많이 나타난 변수를 제거하는 것이 바람직.

- 3) 많은 변수들이 범주형인 경우에는 계산시간이 엄청나게 늘어날 수 있다. 계산 시간을 줄이기 위해 같은 변수의 서로 다른 값이 나타나는 빈발항목집합후보를 제거하는 방법이 있음.

예) {출신지역=서울}, {출신지역=대구}는 동시에 나타날 수 없는 빈발항목집합에서 제거하는 것이 타당.

4.4.2 연속형 데이터의 연관분석

연관분석을 원하는 데이터에 연속형 변수(예: 나이, 월수입, 쇼핑횟수)가 포함되어 있을 수 있는데, '나이가 30대이고 월수입이 300만원 이상인 사람이 인터넷 쇼핑을 월 3회 이상 한다'와 같은 정량적 연관규칙(quantitative association rule)을 기대할 수 있다.

가. 이산화 기반 방법

연속형 변수에 대한 이산화(discretization)에 많이 사용되는 방법은 인접한 연속형 값을 유한개의 구간으로 나누는 것이다.

예) 나이는 20대, 30대 등의 10살 간격으로 나눔.

구간을 나눌 때는 같은 너비로 나누는 방법, 같은 도수로 나누는 방법, 엔트로피 기반으로 나누는 방법, 군집화하여 나누는 방법 등을 생각할 수 있다.

- ▶ 이산화한 후에는 범주형 데이터처럼 이항데이터로 변환하여 연관분석 알고리즘에 적용.
- ▶ 이산화한 경우는 구간의 수가 문제가 되는데 구간수가 너무 적으면 신뢰도가 작아지고, 구간수가 너무 많으면 지지도가 작아져 의미 있는 연관규칙을 잃어버릴 수 있음.
- ▶ 구간의 수를 일단 많이 나눈 후 인접구간과 하나씩 합쳐서 의미 있는 연관규칙을 찾아냄.

나. 통계량 기반 방법

통계량 기반 연관규칙은 규칙의 결과를 통계량으로 찾는 방법

예) {월수입>300만원, 인터넷 쇼핑=함} → {나이: 평균=38세}

- ▶ 통계량 기반 연관규칙 생성 과정
 - ① 목표변수를 정하기 (예) 나이
 - ② 나머지 변수를 이항변수로 변환하여 빈발항목집합을 찾는다.
 - ③ 각각의 빈발항목집합은 전체 데이터를 분할
 - ④ 각 분할된 데이터의 평균 등의 통계량을 구하여 연관규칙 완성
- ▶ 통계량 기반 연관규칙의 수는 전체 빈발항목집합의 수와 동일함.
- ▶ 통계량 기반 연관규칙에서 지지도가 정의되지 않기 때문에, 연관규칙에 의해 설명되는 데이터의 통계량과 이 연관규칙과 관계가 없는 나머지 데이터의 통계량이 서로 유의하게 달라야 함.

예) {월수입>300만원, 인터넷 쇼핑=함} → {나이: 평균=38세}가 흥미가 있으려면 {월수입>300만원, 인터넷 쇼핑=함}이라는 조건을 만족하지 않는 데이터들에 대한 나이의 평균이 30세라고 하자. 이 나이의 차이가 의미 있는지 통계적 가설검정을 통해 알아본다.

- ▶ 통계량 기반 연관규칙을 $A \rightarrow \{t: \mu_1\}$ 라 하자.

A : 빈발항목집합, t : 연속형 목표변수

μ_1 : A 에 속하는 데이터들의 모집단 평균, μ_2 : A 에 속하지 않는 데이터들의 모집단 평균
(편의상, $\mu_1 > \mu_2$ 로 가정)

- ▶ 연관규칙이 흥미가 있다는 것은 $\mu_1 - \mu_2$ 가 분석자가 정한 Δ 보다 클 때이다. (즉, 대립가설)

- ▶ $H_0: \mu_1 - \mu_2 \leq \Delta$
 $H_1: \mu_1 - \mu_2 > \Delta$

- ▶ 두 데이터의 수가 충분히 크다면 위 가설은 다음과 같은 Z 통계량을 이용하여 검정할 수 있다.

$$Z = \frac{\overline{X_1} - \overline{X_2} - \Delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

n_1 : A 의 조건을 만족하는 데이터의 수, 평균: $\overline{X_1}$, 분산: s_1^2

n_2 : A 의 조건을 만족하지 않는 데이터의 수, 평균: $\overline{X_2}$, 분산: s_2^2

- ▶ 귀무가설 H_0 가 참일 때 위의 통계량은 정규분포를 따르므로 유의수준 α 에 대한 정규분포 기준 값이 Z_α 일 때 $Z > Z_\alpha$ 이면 H_0 가 기각된다.

[예제 4.4.1] 통계량 기반 연관규칙 ‘{월수입>300만원, 인터넷 쇼핑=함} → {나이: 평균=32세}’을 만족하는 데이터가 50개 있고, 나이의 평균이 32, 표준편차가 3이다. {월수입>300만원, 인터넷 쇼핑=함}을 만족하지 않는 데이터는 100개 있고, 나이의 평균은 25, 표준편차는 5이다. 두 집단의 나이 평균차가 5세 이상일 때 흥미 있는 연관규칙이라고 할 때, 위의 연관규칙이 흥미 있는지 유의수준 5%로 가설 검정하라.

$$n_1 = 50, \overline{X_1} = 32, s_1^2 = 3^2, \quad n_2 = 100, \overline{X_2} = 25, s_2^2 = 5^2$$

$$H_0: \mu_1 - \mu_2 \leq 5$$

$$H_1: \mu_1 - \mu_2 > 5$$

$$Z = \frac{32 - 25 - 5}{\sqrt{\frac{3^2}{50} + \frac{5^2}{100}}} = 3.050, \quad Z_{0.05} = 1.645$$

따라서 $Z > Z_{0.05}$ 이므로 귀무가설은 기각된다. 즉, ‘연관규칙은 흥미가 있다’라고 결론 내릴 수 있다.

4.5 항목 출현순서의 연관분석

- ▶ 트랜잭션에 포함되어 있는 항목들은 시간에 관련되어 있을 수 있음.

예) 1. 한 슈퍼마켓 방문자의 상품 구입순서

<우유>, {쥬스}, {소고기}, {생선}, {설탕}, {라면}>

-> 고객들이 물품을 사는 순서와 같은 정보가 들어있음

2. 어느 인터넷 쇼핑 웹사이트 방문자의 웹페이지 방문순서

<홈페이지>, {전자제품}, {디지털카메라}, {구매}, {주문확인}, {홈페이지}>

-> 어떤 순서로 웹페이지를 방문하였는가? 라는 정보가 있음.

- ▶ 출현순서패턴(sequence pattern)의 분석은 항목들의 출현구조를 알아볼 수 있으며, 어떠한 항목이 미래에 나타날 것인지 예측도 가능하게 해줌.

4.5.1 출현순서패턴의 기본개념

[표 4.12] 5명의 고객이 일자별로 구입한 상품 데이터의 예

고객명	일자	구입 상품
A	1	a, b, d
A	2	b, c
A	3	e
B	1	a, b
B	2	b, c, d
C	1	a, b
C	2	b, c, d
C	3	b, d, e
D	1	b
D	2	c, d
D	3	d, e
E	1	a, c
E	2	b, c, e

- ▶ 5명의 고객, 일자별로 구입한 13개의 상품 구매 트랜잭션, 구입상품 종류는 5개
출현순서(sequence): 각각의 고객에 대한 일자별 구입상품 트랜잭션의 집합을 출현순서라 함.

$s = \langle e_1, e_2, \dots, e_n \rangle$ 으로 표기

예) 고객 A: $\langle \{a, b, d\}, \{b, c\}, \{e\} \rangle$

고객 B: $\langle \{a, b\}, \{b, c, d\} \rangle$

고객 C: $\langle \{a, b\}, \{b, c, d\}, \{b, d, e\} \rangle$

고객 D: $\langle \{b\}, \{c, d\}, \{d, e\} \rangle$

고객 E: $\langle \{a, c\}, \{b, d, e\} \rangle$

- ▶ 출현순서 데이터에서는 고객 중에 상품 a와 상품 b를 같이 구입한 고객은 5명중 3명이 있다.
이를 부출현순서(subsequence)라 하고 $\langle \{a, b\} \rangle$ 로 표시.
부출현순서 $\langle \{a, b\} \rangle$ 의 지지도는 $3/5=60\%$
- ▶ 첫째 날은 상품 {a}, 둘째 날은 상품 {b}를 구입한 고객이 몇 명인지. 이를 부출현순서 $\langle \{a\}, \{b\} \rangle$ 로 표시하고, 지지도는 $4/5=80\%$
- ▶ 부출현순서의 지지도가 최소 기준값을 넘으면 빈발출현순서(frequent sequence) 또는 출현순서패턴(sequence pattern)이라 함.

[표 4.13] [표 4.12]의 데이터에 대해 지지도 기준을 50%로 했을 때 빈발출현순서

빈발출현순서	지지도
$\langle \{a,b\} \rangle$	60%
$\langle \{b,c\} \rangle$	60%
$\langle \{b,d\} \rangle$	80%
$\langle \{c\}, \{e\} \rangle$	80%
$\langle \{a\}, \{b\} \rangle$	80%
$\langle \{b\}, \{b\} \rangle$	60%
$\langle \{a\}, \{b,c\} \rangle$	60%
$\langle \{b\}, \{b,c\} \rangle$	60%
$\langle \{a,b\}, \{b,c\} \rangle$	60%

4.5.1 출현순서패턴 탐색 알고리즘

- ▶ 출현순서패턴의 원시적인 탐색은 모든 가능한 부출현순서들을 하나씩 살펴보고 각각의 지지도를 조사함
- ▶ m 개의 항목집합 $\{i_1, i_2, \dots, i_m\}$ 이 있을 때 가능한 출현순서를 1-항목 출현순서, 2-항목 출현순서, 3-항목 출현순서 등으로 나열해보면 다음과 같다.

1-항목: $\langle \{i_1\} \rangle, \langle \{i_2\} \rangle, \dots, \langle \{i_m\} \rangle$

2-항목: $\langle \{i_1, i_2\} \rangle, \langle \{i_1, i_3\} \rangle, \dots, \langle \{i_{m-1}, i_m\} \rangle$
 $\langle \{i_1\}, \{i_1\} \rangle, \langle \{i_1\}, \{i_2\} \rangle, \dots, \langle \{i_{m-1}\}, \{i_m\} \rangle$

3-항목: $\langle \{i_1, i_2, i_3\} \rangle, \langle \{i_1, i_2, i_4\} \rangle, \dots, \langle \{i_1, i_2\}, \{i_1\} \rangle, \dots,$
 $\langle \{i_1\}, \{i_1, i_2\} \rangle, \dots, \langle \{i_1\}, \{i_1\}, \{i_1\} \rangle, \dots, \langle \{i_m\}, \{i_m\}, \{i_m\} \rangle$

- ▶ 빈발 출현순서패턴의 탐색에도 선형적 규칙(apriori principle)이 적용될 수 있음.
- ▶ k -항목 출현순서는 반드시 $(k-1)$ -항목 출현순서를 포함하여야 하는데, 이를 이용하여 출현순서패턴을 찾는 선형적 알고리즘을 만들 수 있다.
- ▶ 빈발 k -항목 출현순서 후보를 생성하고, 비빈발 $(k-1)$ -항목 출현순서를 가지치기한 후 출현순서후보의 지지도를 계산하여 빈발출현순서패턴을 결정한다.

가. k -항목 출현순서패턴 후보 생성

- ▶ 출현순서패턴 s_1 과 s_2 의 혼합은 s_1 의 첫 번째 항목집합을 뺀 출현순서와 s_2 의 마지막 항목집합을 뺀 출현순서가 동일할 때 s_1 과 s_2 의 마지막 항목집합을 혼합한다.

(예) $s_1 = \langle e_1, e_2, e_3, e_4 \rangle$, $s_2 = \langle e_2, e_3, e_4, e_5 \rangle$ 이면, s_1 과 s_2 의 혼합은 $\langle e_1, e_2, e_3, e_4, e_5 \rangle$ 이다.

- ▶ 만일 s_2 의 마지막 두 개의 항목집합이 서로 다른 집합이라면 s_2 의 마지막 항목만 s_1 과 결합하고, 만일 서로 같은 집합에 속해 있다면 s_2 의 마지막 항목은 s_1 의 마지막 항목집합의 원소가 된다.

(예) $s_1 = \langle e_1, e_2, e_3, e_4 \rangle$, $s_2 = \langle e_2, e_3, e_4, e_4^* \rangle$ 일 때,

만약 $e_4^* \subset e_4$ 이면, s_1 과 s_2 의 혼합은 $\langle e_1, e_2, e_3, e_4 \rangle$ 이다.

만약 $e_4 \neq e_4^*$ 이면, s_1 과 s_2 의 혼합은 $\langle e_1, e_2, e_3, e_4, e_4^* \rangle$ 이다.

나. 후보들의 가지치기

▶ k -항목 출현순서패턴 후보에서 모든 가능한 $(k-1)$ -항목 부출현순서를 조사하여 적어도 하나 이상이 출현순서패턴이 아니라면 그 k -항목 출현순서는 가지치기 한다.

예) 고객 A: $\langle \{a,b,d\}, \{b,c\}, \{e\} \rangle$

고객 B: $\langle \{a,b\}, \{b,c,d\} \rangle$

고객 C: $\langle \{a,b\}, \{b,c,d\}, \{b,d,e\} \rangle$

고객 D: $\langle \{b\}, \{c,d\}, \{d,e\} \rangle$

고객 E: $\langle \{a,c\}, \{b,d,e\} \rangle$

지지도 기준을 40%로 했을 때 빈발출현순서를 구하시오.

1-항목: $\langle \{a\}, \langle \{b\}, \langle \{c\}, \langle \{d\}, \langle \{e\} \rangle$

2-항목: $\langle \{a\}, \{b\} \rangle, \langle \{a\}, \{c\} \rangle, \langle \{a\}, \{d\} \rangle, \langle \{a\}, \{e\} \rangle, \langle \{b\}, \{c\} \rangle, \langle \{b\}, \{d\} \rangle, \langle \{b\}, \{e\} \rangle$

$\langle \{c\}, \{d\} \rangle, \langle \{c\}, \{e\} \rangle, \langle \{d\}, \{e\} \rangle, \langle \{b\}, \{b\} \rangle$

$\langle \{a,b\} \rangle, \langle \{b,c\} \rangle, \langle \{b,d\} \rangle, \langle \{c,d\} \rangle, \langle \{d,e\} \rangle$

3-항목: $\langle \{a\}, \{b,c\} \rangle, \langle \{a\}, \{b,d\} \rangle, \langle \{b\}, \{b,c\} \rangle, \langle \{b\}, \{c,d\} \rangle, \langle \{c\}, \{d,e\} \rangle$

$\langle \{a,b\}, \{c\} \rangle, \langle \{a,b\}, \{b\} \rangle, \langle \{b\}, \{c\}, \{e\} \rangle$

4-항목: $\langle \{a,b\}, \{b,c\} \rangle$

4.6 비빈발패턴 탐색

▶ 데이터에서 빈발하게 나타나지 않는 패턴(비빈발패턴)은 무관심하게 여기고 제거할 수 있다.

▶ 이러한 비빈발패턴은 대개 무의미한 경우가 많으나 어느 경우에는 유용할 수 있다.

▶ 예) 전자제품 매장에서 DVD와 VCR을 동시에 구매하는 사람은 거의 없기 때문에 비빈발패턴이다. DVD 판매가 늘면 VCR 판매는 줄어들 가능성이 많기 때문에, 이를 서로 음의 상관관계가 있는 비빈발패턴이라 한다. 음의 상관관계가 있는 비빈발패턴을 조사하면 어떠한 제품들이 서로 경쟁하고 있는 지 알 수 있다.

▶ 예) 차와 커피, 버터와 마아가린, 책상용 컴퓨터와 노트북 등도 음의 상관관계가 있는 경쟁제품의 예이다.

▶ ‘어떻게 흥미 있는 비빈발패턴을 정의하느냐?’, ‘어떻게 효율적으로 대량 데이터에서 비빈발패턴을 탐색하느냐?’라는 문제이다.

4.6.1 흥미 있는 비빈발패턴의 정의

1) 항목들의 집합을 $\{i_1, i_2, \dots, i_m\}$ 이라 하고, 한 트랜잭션에 항목 i_k 가 있을 때 이를 양의 항목(positive item), 이 항목이 없을 때 $\overline{i_k}$ 라 표기하고 음의 항목(negative item)이라 하자.

만일 항목집합 X 가 양의 항목집합 A 와 음의 항목집합 \overline{B} 의 합집합이고, X 의 지지도가 최소지지도보다 클 때 이를 음의 항목집합(negative item set)이라 한다.

$$X = A \cup \overline{B}, \quad s(X) \geq \text{minsupport}$$

2) 음의 항목집합에서 최소지지도와 최소신뢰도를 만족하는 연관규칙을 음의 연관규칙(negative association rule) 또는 음의 패턴(negative pattern)이라 함.

예) 음의 패턴 $\{\text{차}\} \rightarrow \{\overline{\text{커피}}\}$ 는 ‘차를 마시는 사람은 커피를 마시지 않는다’라는 연관규칙을 의미

3) $X = \{x_1, x_2, \dots, x_m\}$ 를 m -항목집합,

$P(X)$: 한 트랜잭션이 X 를 포함할 확률이라 하면, 확률 $P(X)$ 의 추정값은 지지도 $s(X)$ 이다.

항목 x_j 의 지지도가 $s(x_j)$ 일 때, 모든 항목들이 독립이면 $s(X) = s(x_1)s(x_2) \cdots s(x_m)$ 이다.

만일 m -항목집합 X 가 $s(X) < \prod_{j=1}^m s(x_j) = s(x_1) \times s(x_2) \times \cdots \times s(x_m)$ 를 만족하면 X 를 음의 상관 항목 집합(negatively correlated item set)이라 한다.

$s(X)$ 가 작으면 작을수록 음의 상관관계가 강한 패턴이 된다.

4) 항목집합 X 와 Y 가 서로 공통된 항목이 없을 때 연관규칙 $X \rightarrow Y$ 가 $s(X \cup Y) < s(X)s(Y)$ 를 만족할 때 음의 상관 연관규칙(negatively correlated association rule) 또는 음의 상관 패턴(negatively correlated pattern)이라 한다.

항목집합 X 와 Y 에 대한 음의 항목집합을 \bar{X} 와 \bar{Y} 라 하면 위의 식은 다음과 같이 표현할 수 있다.

$$s(X \cup Y)s(\bar{X} \cup \bar{Y}) < s(X \cup \bar{Y})s(\bar{X} \cup Y)$$

비빈발패턴, 음의 패턴, 음의 상관 패턴은 서로 밀접한 관련이 있음.

1) $X \cup Y$ 가 비빈발항목집합이고, 최소지지도가 높지 않을 경우 항목집합 X , Y 와 관련된 음의 항목집합이 있을 가능성이 높다. 만일 $X \cup Y$ 가 비빈발항목집합이라면 $s(X \cup Y)$ 가 작은 값이고 상대적으로 $s(X \cup \bar{Y})$, $s(\bar{X} \cup Y)$ 또는 $s(\bar{X} \cup \bar{Y})$ 중의 하나가 최소지지도보다 커야 된다.

2) 음의 상관 패턴은 관련된 음의 패턴을 가지고 있다. 항목집합 X 와 Y 가 강한 음의 상관을 가진다면 위의 식의 좌측은 우측보다 매우 작은 값을 갖는다. 따라서 X 와 Y 가 강한 음의 상관을 가진다면 $s(X \cup \bar{Y})$ 나 $s(\bar{X} \cup Y)$ 또는 두 가지도 모두 높아야 하므로 음의 패턴과 연관이 있게 된다.

3) $X \cup Y$ 의 지지도가 낮으면 낮을수록 음의 상관관계가 높은 패턴이 될 가능성이 있게 되고, 비빈발 음의 상관 패턴들은 흥미 있는 패턴이 될 가능성이 높다.

4.6.2 비빈발패턴의 탐색방법

1) 음의 패턴을 찾는 간단한 방법은 모든 항목들을 두 개의 양의 항목과 음의 항목으로 이항변수화하는 것이다. 그 후 기존의 선형적 알고리즘으로 모든 음의 빈발항목집합을 찾을 수 있다.

2) 양의 항목집합의 지지도를 이용하여 음의 항목집합의 지지도를 구함.

예를 들면, $\{p, \bar{q}, \bar{r}\}$ 의 지지도는 다음과 같이 구할 수 있다.

$$s(\{p, \bar{q}, \bar{r}\}) = s(\{p\}) - s(\{p, q\}) - s(\{p, r\}) + s(\{p, q, r\})$$

위의 식을 일반화하면 항목집합 $X \cup \bar{Y}$ 의 지지도는 다음과 같다.

$$s(X \cup \bar{Y}) = s(X) + \sum_{i=1}^n \sum_{Z \subset Y, |Z|=i} \{(-1)^i \times s(X \cup Z)\}$$

여기서, Z 는 Y 의 부분집합이고, $|Z|$ 는 이 부분집합의 원소수이다.

$$(\text{예}) \quad s(X \cup \bar{Y}) = s(X) - s(X \cup Y), \quad s(\bar{X} \cup Y) = s(Y) - s(X \cup Y)$$

$$s(\bar{X} \cup \bar{Y}) = s(\bar{X}) - s(\bar{X} \cup Y)$$