데이터마이닝에서 수량연관규칙 탐사방법

박 원 환 1)

요 약

연관규칙은 데이터베이스에 잠재되어 있는 유용한 정보를 탐사하는 방법으로 데이터마이닝의 한 분야이다. 이는 항목의 발생유무만을 고려하는 이진연관규칙에 대한 연구가 주였으나, 최근에는 항목의 수량까지 고려하는 수량연관규칙 탐사가 소개되고 있다. 수량연관규칙은 수량속성 항목을 임의의 방법으로 여러 개의 소구간 항목으로 분할한 후, 각각을 이진항목으로 취급하여 연관규칙을 탐사하는 방법이다. 본 논문에서는 분할된 여러 소구간 분할항목들 중에서 필요 소구간 항목만을 선택적으로 탐사하는 방법을 제안한다. 제안방법은 블린항목제약식을 사용하여 수량항목의 탐사범위를 제한함으로써 모든 분할을 탐사하지 않고 필요 소구간만을 탐사하기 때문에 탐사시간을 단축할 수 있다.

주요용어 : 데이터마이닝, 연관규칙탐사, 수량연관규칙, 블린항목제약식

1. 서 론

데이터마이닝은 대량의 실제 데이터, 즉 트랜잭션(transaction)을 발생시킨 특성을 효과적으로 반영한 유용한 정보를 탐사하여 의사결정(decision making)을 위해 제공한다. 예를 들어 대형 할인매장에서 "노트"와 "연필"을 구매한 고객의 85%가 "지우개"를 같이 구매한다는 연관규칙을 판매 데이터베이스의 자료에서 찾을 수 있다. 이를 위한 방법으로 연관규칙(association rules) 탐사에 관한 문제는 R. Agrawal[1]에 의해 처음 제안된 이후 최근까지 연구가 활발하다.

연관규칙 탐사는 현실세계에서 발생하는 트랜잭션에 나타내는 항목(item)들간의 상호관련성을 발견하는 작업을 말하며, 이는 "한 트랜잭션 내에서 특정 항목이 나타나면 반드시 다른 항목도 그 트랜잭션 내에 함께 나타난다." 라는 형태의 규칙성을 발견하는 이진 연관규칙(binary association rule) 탐사에 대한 연구가 주를 이루고 있으며, 특히 취급 항목이 많아지면 탐사시간이 기하급수적으로 증가하므로 이를 최소화하는데 중점적으로 연구하고 있다[2][3].

센서스 자료와 같이 수량적 속성자료의 경우에도 이진연관규칙 탐사도 가능하지만, 연령, 자녀 수, 주택의 연건평 등과 같은 수량적 속성이 강한 자료에서 발견할 수 있는 연관규칙이 극히 제한적이거나 불가능하다. 예들 들어 사람의 연령은 0세에서 120세 내외의 범위에서 나타나지만, "연령"이라는 항목만으로 연관규칙을 탐사할 경우 발견할 수 있는 규칙은 극히 제한적이다. 반대로 "연령0", "연령1", "연령2" ... "연령120"과 같이 각각을 항목으로 취하여 연관규칙을 탐사한다면 탐사공간이 너무 넓어 탐사시간이 엄청나게 증가할 뿐만 아니라 데이터 발생이 분산되어 있어 탐사가 불가능한 문제가 있다.

이와 같은 문제해결을 위한 관련 연구로는 수량항목의 도메인을 일정한 작은 간격으로 분할(partition)과 이웃한 소구간을 병합(merge)하는 일정간격 분할 및 병합방법[4], 수량적 속성의 분포를 고려하여 소간격으로 분할하는 유동적 분할 및 병합방법[5]과 도메인 내의 단위구간들의 발생빈도를 고려하는 최빈수 기반의 분할법[6] 등이 소개되고 있다. 이들 방법들은 수량항목의 도메인을 제시된 최소분할지지도를 기초로 여러 개의 빈발구간 항목들을 생성할 때, 양질의 빈발구간항목을 생성하는 하는 연구

1) 통계청 서기관/공학박사/기술사/whpark@nso.go.kr/정부대전청사 3동

들이며, 양질의 빈발구간항목을 생성하였다 하더라도 생성된 모든 수량항목에 대하여 연관규칙탐사를 통하여 사용자에게 불필요한 규칙까지 탐사하는 문제는 아직도 남아 있다. 따라서 본 논문에서는 이의 해결을 위하여 이진연관규칙 탐사에 사용되는 Direct 알고리즘의 블린항목제약식을 수량연관규칙에 적용할 수 있도록 확장한다. 논문의 구성은 제2장에서 연관규칙에 대하여 알아보고, 3장에서는 수량연관규칙 탐사방법과 제안방법과 적용 예를 보이며, 마지막으로 결론 및 향후과제를 밝힌다.

2. 연관규칙

2.1 정의

항목들의 집합 $I = \{i_1, i_2, i_3, \cdots, i_m\}$이라 하면, 트랜잭션 T는 I의 부분집합이다. 데이터베이스 D에는 n개의 T가 저장되어 있다. 여기서 연관규칙을 탐사한다고 하자. 항목들의 집합 X, Y에 대한 연관규칙 탐사의 결과는 사용자에 의하여 주어지는 최소지지도(minimum support, S_{min}) 이상의 지지도와 최소신뢰도(minimum confidence, C_{min}) 이상의 신뢰도를 갖고, 다음의 성질을 갖는 연관규칙 $R : X \rightarrow Y$의 집합이다.

- $X \subseteq I$인 X에 대해, $X \subseteq T$이면, T는 X를 만족한다고 정의하고, D에서 X를 만족하는 트랜잭션의 수는 *freq*(X)로 표기한다.
- $X, Y \subseteq I$이고, $X \cup Y = \phi$ 이면, 규칙 $R : X \rightarrow Y$의 지지도(S) = *freq*(X∪Y)/n 과 신뢰도 C = *freq*(X∪Y)/*freq*(X)를 갖는다.
- 최소지지도(S_{min}) 이상을 갖는 항목열 $X \subseteq I$를 빈발항목집합(large itemset)이라 정의한다. 이 때 X의 부분집합도 빈발항목집합이다.

연관규칙 탐사는 R. Agrawal[1]에 의해 처음 소개된 이후, 새로운 알고리즘이 많이 소개되고 있으나 이들의 스키마는 대부분 유사하다. 즉, 데이터베이스에 있는 모든 빈발항목들의 지지도를 계산하여 빈발항목집합을 찾고, 이로부터 주어진 신뢰도를 바탕으로 실제의 규칙을 탐사하는 과정으로 이루어지는 2단계 구조이다.

2.2 탐사 알고리즘

[그림2-1]의 Apriori[1]은 연관규칙 탐사의 대표적인 알고리즘이며, [그림3]의 Direct[7]는 항목을 선별하여 탐사하는 탐사범위제한 알고리즘이다. 이들은 모두 이진연관규칙 탐사에 사용한다.

```
// DB 검색하여 C1, L1 생성
 L1 = { large 1-itemsets}

 for ( k=2 ; Kk-1 ≠ φ ; k++ ) do begin
    Ck = apriori-gen(Lk-1) ; // New candidates

    forall transaction t ∈ D do begin
      Ct = subset(Ck, t) ;
         forall candidates c ∈ Ct do
              c.count++ ;
    end
         Lk = { c∈ Ck | c.count ≥ Smin }
 end
 Answer = Uk Lk ;
```

[그림2-1] Apriori 알고리즘

```
1) Join 단계
   insert into Ck
   select p.item1, p.item2, p.item3, ··· p.itemk-1,
          q.itemk-1
   from Lk-1 p, Lk-1 q        // self join
   where p.item1 = q.item1, ··· p.itemk-2 =
          q.itemk-2, p.itemk-1 < q.itemk-1 ;

2) Prune 단계
   forall itemset c ∈ Ck do
       forall (k-1) - subsets s of c do
            if ( s ¬∈ Lk-1 ) then
                delete c from Ck ;
```

[그림2-2] Apriori-gen 함수

Apriori는 첫 단계에서는 항목별로 빈도를 계산하여 빈발 1-항목집합을 결정하고, k(k≥2)번째는 두 단계로 분할하여 알고리즘이 진행된다. 먼저, (k-1)번째 검색에서는 발견된 빈발항목집합 L_{k-1}를 후보항목집합 C_k

으로 만들어 지지도를 계산한 후, 최소지지도를 만족하는 항목만 L_k에 진입시킨다. 이러한 시행은 L_k가 더 이상 발견되지 않을 때까지 반복한다. 이상에서 알 수 있듯이 Apriori는 길이가 l인 빈발항목집합을 찾기 위하여 2^l개의 빈발항목 부분집합(sub-itemset)을 만들고, 또한 각 길이의 항목집합에 대하여 빈발여부를 확인하기 위해 DB를 검색하므로 항목 수 증가에 대하여 탐사공간이 기하급수적으로 증가한다.

임의의 항목에만 관심이 있는 사용자에게 관심 없는 연관규칙까지 탐사하여 제공한다면, 실질적으로 탐사공간이 넓어져 탐사시간만 길어지는 한계가 있다. 이에 대하여 Direct는 블린항목제약식을 사용하여 탐사해야할 항목을 제한하여 탐사하도록 하여 탐사속도를 개선한 알고리즘이다.

Direct의 초기에만 지지도를 계산하고, 이후에는 최소지지도를 만족한 항목에 대하여 주어진 블린항목제약식을 만족하는 항목들만 선택해서 빈발항목집합을 탐사한다. 따라서 최소지지도와 블린항목제약식은 주어지며, 다음의 기호를 사용한다.

- $B = D_1 \vee D_2 \vee D_3 \cdots \vee D_m$
- $D_i = \alpha_{i1} \wedge \alpha_{i2} \wedge \alpha_{i3} \cdots \wedge \alpha_{i\,ni}$
- α_{ij} : $l_{ij} \in L$ 인 항목에 대하여 l_{ij} 또는 $\neg\, l_{ij}$
- $L = \{ l_1, l_2, \cdots, l_m \}$; 항목이라는 리터럴 집합
- S : B를 만족하는 모든 항목집합의 항목들 집합
- *적중항목집합* : S에 있는 한 항목을 포함하는 항목집합
- *k-항목집합* : k 항목으로 이루어진 집합
- L^s_k : S 에 있는 항목들로 구성된 빈발 k-항목집합들 집합
- L^b_k : B를 만족하는 빈발 k-항목집합들의 집합
- C^s_k : S 에 있는 항목들로 구성된 후보 k-항목집합들 집합
- C^b_k : B를 만족하는 후보 k-항목집합들 집합
- F : 모든 빈발 항목들 집합

1. DB를 검색하여 최소지지도를 만족하는 항목들의 집합, 즉 빈발항목집합을 생성(F)하고, 이중 주어진 B(제약식)를 만족하는 원소들만 L^b_1로 사용한다.
2. $C^s_{k+1} := L^s_k \times F$를 계산한다.
3. C^s_{k+1} 의 모든 후보 항목들 중에서 B를 만족하지 않는 항목집합을 삭제한다.
4. 잔여 C^s_{k+1} 의 모든 후보 항목들 중에서 B를 만족하지만 최소지지도를 만족하지 않는 k-부분집합을 가진 후보 항목들은 삭제한다.
5. (k+1)개의 부정부호(¬) 없는 원소로 구성된 B에 있는 각 Di, 즉 ($\alpha_{i1} \wedge \alpha_{i2} \wedge \cdots \wedge \alpha_{ini}$)에 대하여, 모든 원소(α_{ip})가 빈발하다면, 항목집합 $\{\alpha_{ip1} \wedge \alpha_{ip2} \wedge \cdots \wedge \alpha_{ipk+1}\}$를 C^s_{k+1} 에 추가한다.
6. C^b_{k+1} 의 각 항목집합들의 지지도를 계산하여 최소지지도를 만족하지 않는 항목을 제외하고 잔여 항목집합을 L^b_{k+1} 에 포함시킨다.
7. L^b_{k+1} 가 공집합이면 종료하고, 그렇지 않으면 k=k+1하여 2번으로 간다.

[그림3] Direct 알고리즘

위의 Direct는 이진연관규칙에서 탐사항목을 제한하여 탐사시간 단축하고, 또한 불필요한 규칙발견의 배제가 가능하다. 다만, 이를 위하여 식(2-1)과 같이 주어진 블린항목제약식을 사용한다.

$$B = (l_{1\ 1} \wedge l_{1\ 2} \wedge \cdots \wedge l_{1\ ni}) \vee (l_{2\ 1} \wedge lc_{2\ 2} \wedge \cdots \wedge l_{2\ ni}) \vee \cdots \vee (l_{m\ 1} \wedge l_{m\ 2} \wedge \cdots \wedge l_{m\ ni}) \quad \text{-----} \ (2\text{-}1)$$

여기서 m은 합집합 수, ni는 교집합 수이다.

3. 탐사범위제한 수량연관규칙

3.1 수량연관규칙

수량연관규칙(quantitative association rules)은 수량속성을 포함하는 규칙을 말한다. 이를 위하여 관련 용어를 도메인, 소구간 분할 집합을 다음과 정의한다.

- 도메인(domain)은 수량속성 항목의 데이터가 발생하는 영역
- 소구간 분할집합은 도메인을 임의의 방법으로 분할된 빈발한 소구간들 집합

수량 연관규칙을 예를 통하여 구체적으로 알아보자. 표1은 예제 데이터베이스이고, 표2는 표1의 데이터에 대하여 최소지지도 40%, 최소신뢰도 60%일 때의 수량 연관규칙의 예이다.

표1. 데이터베이스의 예(사람)

Tid	나이	결혼	차동차수
1	23	no	1
2	25	yes	1
3	29	no	0
4	34	yes	2
5	38	yes	2

표2. 수량 연관규칙의 예

연관규칙(예)	지지도	신뢰도
R_1 : <나이 : 30~39> and <결혼 : yes> → <자동차 : 2대 >	40%	100%
R_2 : <자동차 : 0~1> → <결혼 : no>	40%	66.6%

위 예의 R_1은 '나이 30세에서 39세 중에서 기혼자는 자동차 2대를 보유하고 있다' 라는 의미의 수량연관규칙이다. 여기서 '나이' 항목만 포함한다면, 규칙은 이진연관규칙(binary association rules)이다. 이와 같이 항목을 수량까지 고려했을 때, 발견되는 규칙은 보다 고급의 정보가 된다.

3.2 탐사범위제한 수량연관규칙

항목집합(L), 블린제약식(B) 등 관련용어와 기호는 다음과 같이 정의한다.

- $L = Lc+Lq$ (Lc : 범주속성(categorical attribute) 항목, Lq : 수량 속성(quantitative attribute) 항목)
- $B = D_1 \vee D_2 \vee D_3 \cdots \vee D_m$이며, m개의 합집합(conjunction)이다.
- $D_i = \alpha_{i\,1} \wedge \alpha_{i\,2} \wedge \alpha_{i\,3} \cdots \wedge \alpha_{i\,ni}$이며, ni개의 교집합(disjunction)이다.
- 범주속성 항목일 경우의 α_{ij} : $lc_{ij} \in Lc$에 대하여 lc_{ij} 또는 $\neg\, lc_{ij}$ (범주 항목의 존재유무)
- $Lc = \{ lc_1, lc_2, \ldots, lc_{m-k} \}$: 범주항목이라는 리터럴 집합이며, k(≤m)는 수량 항목 수이다.
- 수량속성의 소구간 분할일 경우의 α_{ij} : $lq_{kij} \in Lq_k$에 대하여 lq_{kij} 또는 $\neg\, lq_{kij}$ (소구간 분할의 존재유무)
- $Lq_k = \{ lq_{k\,1}, lq_{k\,2}, \ldots, lq_{k\,P} \}$: 수량속성항목의 소구간 분할의 리터럴 집합, p는 소구간 분할 수이다. □

위에서 정의에 의하면 L에는 m개의 항목이 Lc와 Lq에 나누어 분포하고 있다. 따라서 Lc와 Lq를 다음과 같이 나열하고, 이를 이용하여 Lc + Lq에 적용할 새로운 블린제약식(B')을 유도하자.

$$Lc = (lc_1, \; lc_2, \; \ldots, \; lc_{m-k})$$

$$\begin{aligned} Lq_1 &= (lq_{1\,1}, \; lq_{1\,2}, \ldots, lq_{1\,p}) \\ Lq_2 &= (lq_{2\,1}, \; lq_{2\,2}, \ldots, lq_{2\,p}) \\ &\cdots \\ Lq_{k-1} &= (lq_{k-1\,1}, lq_{k-1\,2}, \ldots, lq_{k-1\,p}) \\ Lq_k &= (lq_{k\,1}, lq_{k-1\,2}, \ldots, lq_{k\,p}) \end{aligned}$$

$Lq_{ki\,pi}$(ki=1, ...,k이고, pi= 1,2, ..., p)의 구간 값은 오름차순으로 정렬되어 있고, p는 동일하다고 가정하였을 때, 임의의 한 Lq_{ki}에 대하여 ($lq_{ki\,LB}$)와 ($lq_{ki\,UB}$)가 사용자에 의해 주어졌을 때, 고려해야할 탐사범위 유형은 다음과 같다. 여기서 LB와 UB는 pi이다.

① 단일범위 : 소구간 분할들 중에서 한 분할만을 지정하며, ($lq_{ki\,LB}\pm$) 또는 ($lq_{ki\,UB}\pm$)로 표기한다. 그의 부정(¬)은 해당 소구간을 제외한 전체구간을 지정

② 상향범위 : 임의의 지정 소구간을 기준으로 그 이상은 모든 구간을 포함하는 범위를 지정하며, ($lq_{ki\,LB}+$)로 표기한다. 그의 부정(¬)은 지정구간 미만 구간 모두를 지정

③ 하향범위 : 지정구간을 포함한 상향범위의 역이며, ($lq_{ki\,UB}-$)로 표기하며, 그의 부정은 상향범위를 지정

④ 상하범위 : i)과 ii)의 공통범위, ($lq_{ki\,LB}+$)∧($lq_{ki\,UB}-$), 부정은 지정구간 제외구간, $p_1, p_2 = 1, 2, \ldots, p$

위의 정의에 따라 ①, ②, ③은 ④의 특수한 경우이므로 ④에 대하여만 포함된 소구간 분할들을 대입하여 분배법칙(distributive law)에 의하여 정리하면 식 (3-1)과 같다.

$$(lq_{ki\ p1}+)\wedge(lq_{ki\ p2}-) = (lq_{ki\ LB}\vee lq_{ki\ LB+1}\vee\cdots\vee lq_{ki\ p})\ \wedge\ (lq_{ki\ UB}\vee lq_{ki\ UB-1}\vee\ \cdots\ \vee\ lq_{ki\ 1})$$
$$= \underline{(lq_{ki\ LB}\wedge lq_{ki\ UB})\vee(lq_{ki\ LB}\wedge lq_{ki\ UB-1})\vee\ \cdots\ \vee(lq_{ki\ p}\wedge\ lq_{ki\ 1})}\ \text{------}\ (3\text{-}1)$$

식(3-1)에서 밑줄 친 부분의 (a∧b) 형태는 a와 b가 같을 때만 진(true)이므로 탐사가 필요한 지정 구간만 남게 되며, 그의 부정은 해당구간을 제외한 구간이다. 이진항목 블린항목제약식 (2-1)을 Lc 형태로 변형하면 (3-2)와 같다.

$$B = (lc_{1\ 1}\wedge lc_{1\ 2}\wedge\cdots\wedge lc_{1\ ni})\vee(lc_{2\ 1}\wedge lc_{2\ 2}\wedge\cdots\wedge lc_{2\ ni})\vee\cdots\vee(lc_{m\ 1}\wedge lc_{m\ 2}\wedge\cdots\wedge lc_{m\ ni})\ \text{--}(3\text{-}2)$$

식(3-2)에서 항목들 중에서 k개를 식(3-1)로 대체하여 정리하면 블린제약식 B′ (3-3)이 된다.

$$\begin{aligned} B' = &(((lq_{1\ LB}+)\wedge(lq_{1\ UB}-))\vee\cdots\vee((lq_{k\ LB}+)\wedge\ (lq_{k\ UB}-))\wedge\ lc_{1\ 1}\wedge lc_{1\ 2}\ \wedge\ \cdots\ \wedge lc_{1\ ni})\vee \\ &(((lq_{1\ LB}+)\wedge(lq_{1\ UB}-))\vee\cdots\vee((lq_{k\ LB}+)\ \wedge\ (lq_{k\ UB}-))\ \wedge\ lc_{2\ 1}\wedge lc_{2\ 2}\ \wedge\ \cdots\ \wedge lc_{2\ ni})\vee\cdots\vee \\ &(((lq_{1\ LB}+)\wedge(lq_{1\ UB}-))\vee\cdots\vee((lq_{k\ LB}+)\ \ \wedge\ (lq_{k\ UB}-))\ \ \wedge\ lc_{m\text{-}k\ 1}\wedge lc_{m\text{-}k\ 2}\wedge\cdots\wedge lc_{m\text{-}k\ ni})\ \text{---}(3\text{-}3) \end{aligned}$$

이상과 같이 이진(범주)항목에 대한 블린항목제약식 B로부터 수량까지 고려하여 탐사범위제한 블린항목제약식 B′을 얻었다.

3.3 적용 예

B′를 수량항목과 범주항목을 동시에 적용하는 예를 보자. 이를 위해 수량속성 항목(Lq_1, Lq_2) 및 그의 분할들, 그리고 범주항목(Lc)이 아래와 같이 주어졌을 때 사용 예이다.

Lq_1(나이) = <A0-5, A6-10, ⋯ ,A96-100> (5세 등간격 분할) = <a1, a2, ⋯ , a19, a20>	Lq_2(통근시간) = <T0-20,T21-40, ⋯ ,T181-200> (20분 등간격 분할) = <t1, t2, ⋯ , t9, t10 >
Lc = <남자, 여자, 결혼, 미혼, ...>	

[탐사조건] (예1) 81세 이상 남자와 관련된 연관규칙탐사	(예2)21-43세 여자가 통근시간이 1시간 이하 구간에서의 규칙탐사

▶ (예1)의 조건에 따른 제약식은 주어진 조건에서 수량속성(Lq_1)의 ($lq_{1\ LB}$+)= (a16 +), ($lq_{1\ UB}$-)= (a20 -) 범주속성(Lc)은 '남자' 이다, 이를 식(3-3)에 적용하면 다음과 같다.

$$\begin{aligned} B' &= ((lq_{1\ LB}+)\wedge\ (lq_{1\ UB}-)\wedge\ \text{남자}) \\ &= (a16+)\ \wedge\ (a20\ -)\ \wedge\ \text{남자} \\ &= (a16\ \vee\ a17\ \vee\ a18\ \vee\ a19\ \vee\ a20)\ \wedge\ \text{남자} \\ &= (a16\ \wedge\ \text{남자})\vee(a17\ \wedge\ \text{남자})\vee(a18\ \wedge\ \text{남자})\ \vee(a19\ \wedge\ \text{남자})\vee(a20\wedge\text{남자}) \end{aligned}$$

▶ (예2)의 조건에 따른 제약식은 Lq_1과 Lq_2의 조건 ($lq_{1\ LB}$+) = (a5+), ($lq_{1\ UB}$-)= (a9 -), ($lq_{2\ LB}$ +) = (t1+), ($lq_{1\ UB}$-) = (t3 -) 이고, Lc 조건은 '여자' 이다, 이를 식(3-3)에 적용하면 다음과 같다.

$$\begin{aligned} B' &= ((lq_{1\ LB}+)\wedge(lq_{1\ UB}-)\wedge(lq_{2\ LB}+)\wedge(lq_{2\ UB}-)\wedge\text{여자}) \\ &= ((a5+)\wedge(a9-)\wedge(t1+)\wedge\ (t3-)\wedge(\text{여자}) \\ &= (a5\vee a6\vee a7\vee a8\vee a9)\wedge(t3\vee t2\vee t1)\wedge(\text{여자}) \\ &= (a5\wedge(t3\vee t2\vee t1))\vee(a6\wedge(t3\vee t2\vee t1))(a7\wedge(t3\vee t2\vee t1))\vee(a8\wedge(t3\vee t2\vee t1))\vee(a9\wedge(t3\vee t2\vee t1))\wedge(\text{여자}) \\ &= (a5\wedge t3\wedge\text{여자})\vee(a5\wedge t2\wedge\text{여자})\vee(a5\wedge t1\wedge\text{여자})\vee(a6\wedge t3\wedge\text{여자})\vee(a6\wedge t2\wedge\text{여자})\vee(a6\wedge t1\wedge\text{여자}) \\ &\quad\vee(a7\wedge t3\wedge\text{여자})\vee(a7\wedge t2\wedge\text{여자})\vee(a7\wedge t1\wedge\text{여자})\vee(a8\wedge t3\wedge\text{여자})\vee(a8\wedge t2\wedge\text{여자})\vee(a8\wedge t1\wedge\text{여자}) \\ &\quad\vee(a9\wedge t3\wedge\text{여자})\vee(a9\wedge t2\wedge\text{여자})\vee(a9\wedge t1\wedge\text{여자}) \end{aligned}$$

이상과 같이 수량적 속성자료의 소구간 분할에 대하여 탐사범위를 지정할 수 있다. 예의 결과는 식(2-1)과 동일한 형태이므로 [그림3]의 Direct 알고리즘에서 1과, 5단계에서 B′을 사용하고, 잔여 단계는 동일한 방법으로 탐사가 가능하다.

4. 결론 및 향후과제

본 논문에서는 센서스 자료와 같이 수량속성 항목에 대한 연관규칙을 보다 효과적으로 탐사할 수 있도록 기존의 Direct 알고리즘을 확장하여 관심구간을 지정하여 탐사할 수 있도록 하였다. 제안방법의 특징은 수량항목을 여러 개의 소구간으로 분할했을 때 탐사공간이 증가함으로써 탐사시간이 증가하는 문제를 다소 줄일 수 있고, 또한 기존의 블린항목제약식을 확장하였으므로 범주항목은 물론 수량항목까지 제한하여 탐사할 수 있어 탐사구간을 임의로 선택할 수 있어 사용의 편리함도 있다.

그러나 3.1절에서 소개한 바 있는 여러 가지 소구간 분할방법들, 그리고 향후에도 효율적인 분할방법들이 연구될 것이다. 이들 중에서 등간격 분할 및 통합방법[4]은 식(3-3)을 쉽게 응용할 수 있지만, 유동적 분할 및 결합[5] 방법과 빈도기반 분할법[6] 같이 간격을 일정하지 않게 분할하는 분할법에는 지정한 탐사구간 위치도 함께 유동적이 되므로 응용이 쉽지 않다. 따라서 제한탐사시점을 소구간 분할 후로 변경해야 응용이 가능하다. 만약 탐사범위지정과 소구간 분할법을 동시에 고려한 새로운 분할방법을 개발한다면 보다 효율적이고 편리한 방법이 될 수 있지만, 본 논문에서는 취급하지 않고 향후의 연구과제로 남겨둔다.

< 참 고 문 헌 >

[1] R. Agrawal, T. Imielinski, and A. Swami(1993), "Mining association rules between sets of items in large databases", In Proc. of the ACM SIGMOD Conference on Management Data, PP 207-216, Washington, D.C.

[2] R. Agrawal and R. Srikant(1994), "Fast Algorithms for mining association rules", *In Proceedings of the 20th VLDB Conference,* Santiago, Chile, Sept.

[3] J.S. park, M.S. Chen and P.S. Yu(1995), "An Effective hase-based algorithm for mining association rules", *In Proceedings of ACM SIGMOD Conference on Management of Data,* pp. 175-186,San Jose, California, May.

[4] R. Srikant and R. Agrawal(1996), "Mining Quantitative Association Rules in Large Relational Tables", *Proceedings of the ACM SIGMOD Conference on Management of Data,* Monreal, Canada.

[5] 최영희, 장수민, 유재수, 오재철(1999), "수량적 연관규칙탐사를 위한 효율적인 고빈도 항목열 생성기법", 한국정보처리학회 논문지 제6권 제10호, pp2597 - 2607.

[6] 박원환, 박두순(2001), "데이터의 지역성을 이용한 빈발구간 항목집합 생성방법", 멀티미디어학회 논문지 제4권 제5호, pp 465 - 475.

[7] R. Srikant, Q. Vu, and R. Agrawal(1997), "Mining Association Rules with Item Constraints", *Proceedings of the 3rd Int'l Conference on Knowledge Discovery in Databases and Data Mining,* Newport Beach, California, August.