Proceedings of Korean Data And Information Science Society,
November 17-18, 2006, pp. 51-58

A Study for Antecedent Association Rules

Kwang-Hyun Cho[1], Hee-Chang Park[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 association rule mining based antecedent variables. We call these rules to antecedent association rules. An antecedent variable is a variable that occurs before the independent variable and the dependent variable. For example, in politics, a special interest group may want to support a politician who backs their cause. The group would look for a candidate who supports their views and support his election. Once in office, the politician would then conduct policy that supports the interest group.

keywords : antecedent variable, association rules, confidence, lift, support

1. 서론

연관 규칙(association rule)은 하나의 거래나 사건에 포함되어 있는 둘이상의 품목들의 경향을 파악해서 상호 관련성을 발견하는 것으로 대용량 데이터베이스에 존재하는 항목간의 관련성을 찾아내는 작업을 말한다. 마케팅에서는 고객이 동시에 구매한 장바구니를 살펴봄으로써 거래되는 상품들의 관계를 발견 또는 분석한다는 의미에서 장바구니분석(market basket analysis)이라고 한다. 연관 규칙은 교차판매, 매장 진열, 카탈로그 디자인, 장바구니 분석 등에 사용된다. 각 항목간의 연관성을 반영하는 규칙으로 둘 또는 그 이상의 품목들 사이의 지지도(support), 신뢰도(confidence), 향상도(lift)를 바탕으로 관련성 여부를 측정한다. 연관 규칙은 탐색적이며, 비목적성 분석이며, 기존의 데이터를 특별한 변형 없이 계산이 용이하게 사용 가능하다는 장점을 가지고 있으며, 계산 과정이 길고, 반복된 계산이 많으며, 적절한 품목의 결정이 어렵고, 각 품목의 단위에 따른 표준화가 어렵다는 단점을 아울러 가지고 있다. 연관 규칙은 이러한 단점에도 불구하고 두 품목간의 관계를 명확히 수치화함으로써 두 개 이상의

1) Graduate Student, Department of Statistics, Changwon National University, Changwon, Gyeongnam, 641-773, Korea
E-mail : cho1023@changwon.ac.kr
2) Corresponding author : Professor, Department of Statistics, Changwon National University, Changwon, Gyeongnam, 641-773, Korea
E-mail : hcpark@changwon.ac.kr

품목간의 관련성을 나타내 주기 때문에 현업에서 많이 활용되고 있다.

연관 규칙은 Agrawal 등(1993)에 의해 처음 소개된 이후, 많은 학자들에 의해 여러 가지 알고리즘이 연구되고 있다(Agrawal 등(1994), Cheung 등(1996), Park 등(1995), Sergey 등(1997), Toivonen(1996), Saygin 등(2002) 등). 또한 연관 규칙에는 일반적으로 가장 많이 사용되는 순차적 연관 규칙(sequential association rule)을 비롯하여, 두 변수 간의 역의 관련성을 규명할 수 있는 역 연관 규칙(negative association rule), 그리고 연관 규칙 중 독특하거나 드문 규칙을 찾아낼 수 있는 변칙 연관 규칙(anomalous association rule) 등 다양한 종류들이 있다.

연관 규칙 분석에 의해 생성된 규칙에 대하여 변수들 간의 선행 관계가 존재한다면 그 관계에 미치는 영향을 명확히 할 필요성이 있다. 이에 본 논문에서는 연관 규칙에서 변수들 간의 선행관계를 규명할 수 있는 선행 연관 규칙(antecedent association rule)에 관하여 연구하고자 한다. 이러한 선행 연관 규칙은 변수들 간의 관계를 보다 정확하고 명료하게 이해할 수 있도록 해준다. 본 논문의 2절에서는 선행 연관 규칙에 대하여 기술하고 3절에서는 예제 적용 결과에 대하여 기술하며, 4절에서 결론을 맺는다.

2. 선행 연관 규칙

연관 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. 연관규칙을 평가하는 기준에는 지지도(support), 신뢰도(confidence), 향상도(lift) 등이 있으며, 다음과 같이 정의한다.

$$\text{지지도} : S_{(X \Rightarrow Y)} = P(X \cap Y) \tag{2.1}$$

$$\text{신뢰도} : C_{(X \Rightarrow Y)} = P(Y|X) = \frac{P(X \cap Y)}{P(X)} \tag{2.2}$$

$$\text{향상도} : L_{(X \Rightarrow Y)} = \frac{P(Y|X)}{P(Y)} = \frac{P(X \cap Y)}{P(X)P(Y)} \tag{2.3}$$

한편, 선행변수(antecedent variable)는 인과관계에서 독립변수에 선행하면서 독립변수에 대해 유효한 영향력을 행사하는 변수를 의미한다. 선행변수는 매개변수(intervening variable)와는 달리 독립변수와 종속변수간의 관계를 설명하는 것이 아니라, 그 관계에 미치는 영향을 명확히 하고자 할 때 도입한다. 선행변수가 의미를 가지기 위해서는 다음과 같은 조건을 만족해야 한다. 여기서 독립변수를 전항변수(preceding variable)라고 하며, 종속변수를 후항변수(consequent variable)라고 한다.

첫째, 선행변수, 독립변수, 종속변수 간에 상호관련을 맺고 있어야 한다.

둘째, 선행변수를 통제할 때에 독립변수와 종속변수의 관계가 사라져서는 안 된다.

셋째, 독립변수를 통제할 때 선행변수와 종속변수와의 관계는 사라져야 한다.

예를 들면, 만약 사교육비와 예체능대학 진학률 사이에 인과 관계가 나타나고 있는 것을 알고 있다면, 사교육비를 이끌어내는 변수가 또 무엇인지에 대하여 생각해 볼 수 있다. 이때 재산이 사교육비에 영향을 미치는 것을 찾아낸다면, 재산은 선행변수가 된다.

또한, 어떤 제품의 고객 만족도와 제품 구매도 사이에 관련성이 있다고 하자. 이때 고객 만족도에 영향을 미치는 변수가 제품의 질 및 서비스의 질이라는 것을 알 수 있었다면, 여기서 제품의 질 및 서비스의 질은 고객 만족도와 제품 구매도 사이의 선행변수가 된다.

연관 규칙 분석에 의해 생성된 규칙에 대하여 변수들 간의 선행 관계가 존재한다면 그 관계에 미치는 영향을 명확히 할 필요성이 있다. 연관 규칙에서 나타난 규칙들 중 변수들 간의 선행관계를 명확하게 규명하고자 하는 것이 선행 연관 규칙이다. 변수 X를 전항변수(독립변수), Y를 후항변수(종속변수), Z를 선행변수라고 가정했을 때, 선행 연관 규칙의 조건은 다음과 같다.

(조건 1) 변수 X와 변수 Y에 대한 연관성이 존재한다.

$$X \Rightarrow Y$$

(조건 2) 변수 Z와 변수 X에 대한 연관성이 존재한다.

$$Z \Rightarrow X$$

(조건 3) 변수 Z와 변수 Y에 대한 연관성이 존재한다.

$$Z \Rightarrow Y$$

(조건 4) 변수 Z를 통제했을 때, 변수 X와 변수 Y에 대한 연관성이 존재한다.

$$X_{|Z} \Rightarrow Y_{|Z}$$

(조건 5) 변수 X를 통제했을 때, 변수 Y와 변수 Z에 대한 연관성이 존재하지 않는다.

$$Z_{|X} \not\Rightarrow Y_{|X}$$

위의 5가지 조건이 만족하면 선행 연관 규칙이 성립한다고 할 수 있다. 선행 연관 규칙이 성립할 경우, 선행 변수 Z가 변수 X와 변수 Y의 관련성에 대하여 선행적으로 영향을 미친다고 할 수 있다.

2.3 선행 연관 규칙 적용 방법

선행 연관 규칙은 연관성 규칙 생성, 선행 변수 추출, 선행변수 통제 후 연관 규칙 생성, 후항변수 통제 후 연관성 규칙 생성, 선행 연관 규칙 파악의 순으로 진행되며, 그 내용은 다음과 같다.

[단계 1] 연관 규칙 생성
선행 변수를 추출하기 위하여 연관 규칙 생성을 생성한다. 연관 규칙 생성을 위하여 전항변수와 후항변수를 선정하고 전항변수의 개수를 설정한다. 여기서는 일차적으로 각 변수들의 연관 관계를 알아보기 위함이므로 전항변수를 1개로 설정한다. 최소지지도와 최소신뢰도를 결정한 후, 설정된 전항변수의 개수와 최소지지도 및 최소신뢰도를 바탕으로 연관 규칙을 생성한다.

[단계 2] 선행 변수 추출
생성된 연관 규칙을 탐색하여 선행 변수를 추출한다. 선행 연관 규칙의 조건 1, 조건 2, 조건 3을 만족하는 변수를 선행 변수로 지정한다.

[단계 3] 선행변수 통제 후 연관 규칙 생성
선행변수를 통제 한 후 연관 규칙을 생성한다. 연관 규칙 생성 시, 전항변수의 개수 및 최소지지도와 최소신뢰도는 앞에서와 동일하게 지정한다.

[단계 4] 후항변수 통제 후 연관성 규칙 생성
후항변수를 통제 한 후 연관 규칙을 생성한다. 연관 규칙 생성 시, 전항변수의 개수 및 최소지지도와 최소신뢰도는 앞에서와 동일하게 지정한다.

[단계 5] 선행 연관 규칙 파악
선행 연관 규칙이 성립되는 지를 파악한다. 단계 3의 선행변수를 통제 한 후 연관 규칙 생성한 결과가 최소지지도보다 크고, 단계 4의 후항변수를 통제 한 후 연관 규칙 생성한 결과가 최소지지도보다 큰 경우 선행 연관 규칙이 성립한다고 한다.

3. 적용 예제

본 절에서는 사용한 예제 데이터는 2002년 경상남도에서 조사된 사회지표조사 자료 중 일부를 사용하였다. 레코드의 수는 1,000개이며, 사용한 변수는 교육서비스의 만족도, 월평균 병원 이용 횟수, 운동 활동 빈도, 주관적 사회계층, 학력, 연령 등 총 20개 문항을 추출하여 선행 연관 규칙을 적용하였다. 조사 자료에 대한 자세한 내용은 박희창과 조광현(2005)에 제시되어 있다.
연관 규칙 생성 시 최소지지도를 10, 최소신뢰도를 70으로 지정하였으며, 선행 연관 규칙의 조건 1, 조건 2, 조건 3을 만족하는 규칙을 추출하여 선행변수로 지정하였다. 조건에 의하여 추출된 후항변수, 전항변수, 선행변수를 구분하면 <표 1>과 같다.

<표 1> 변수 구분

순번	후항변수	전항변수	선행변수
1	교육서비스의 만족도	주관적 사회계층	학력
2	월평균 병원 이용횟수	운동 활동 빈도	연령

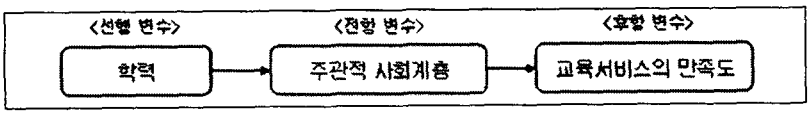
추출된 변수들에 의한 선행 연관 규칙 결과, 선행변수가 학력 및 연령인 경우 모두

선행 연관 규칙이 성립하는 것으로 나타났다. 이를 세부적으로 기술하면 <표 2> 및 <표 3>과 같다.

<표 2> 선행연관성규칙 결과 1

순번	후항변수	전항변수	선행 변수	신뢰도
1	교육서비스의 만족도 = 불만족	주관적 사회계층 = 중상류층 이상	-	72.35
2	교육서비스의 만족도 = 불만족	-	학력 = 대졸이상	78.90
3	-	주관적 사회계층 = 중상류층 이상	학력 = 대졸이상	70.36
4	교육서비스의 만족도 = 불만족	주관적 사회계층 = 중상류층 이상	-	82.42
5	교육서비스의 만족도 = 불만족	-	학력 = 대졸이상	61.35

[표 3.2]에서 보는 바와 같이 후항변수와 전항변수에 의한 연관 규칙 결과(순번 1)의 신뢰도가 72.35로 나타났고, 후항변수와 선행변수에 의한 연관 규칙 결과(순번 2)의 신뢰도가 78.90으로 나타났으며, 전항변수와 선행변수에 의한 연관 규칙 결과(순번 3)의 신뢰도가 70.36으로 나타났다. 이는 선행 연관 규칙의 조건 1, 조건 2, 조건 3을 만족(최소 신뢰도보다 큼)하므로 후항변수를 교육서비스의 만족도, 전항변수를 주관적 사회계층, 선행변수를 학력으로 지정하였다. 선행 연관 규칙이 성립되는 지를 검정하기 위하여 선행변수인 학력을 대졸로 통제한 후, 후항변수와 전항변수의 연관 규칙을 실시하였다. 규칙 결과(순번 4), 신뢰도가 82.42로 나타나 선행 연관 규칙의 조건 4를 만족(최소 신뢰도보다 큼)하는 것을 알 수 있다. 또한, 전항변수인 주관적 사회계층을 중류층과 중상류층으로 통제한 후, 후항변수와 선행변수의 연관 규칙을 실시하였다. 규칙 결과(순번 5), 신뢰도가 61.35로 나타나 선행 연관 규칙의 조건 5를 만족(최소 신뢰도보다 작음)하는 것으로 나타나 선행 연관 규칙이 성립됨을 알 수 있다. 즉, 주관적 사회계층이 높은 응답자들은 교육서비스의 만족도에 대하여 불만족으로 응답함을 알 수 있으며, 선행변수인 학력이 주관적 사회계층을 선행하면서 유효한 영향력을 미치는 것으로 나타났다. 선행 연관성 규칙에 의한 변수들의 관계를 그림으로 나타내면 <그림 1>과 같다.



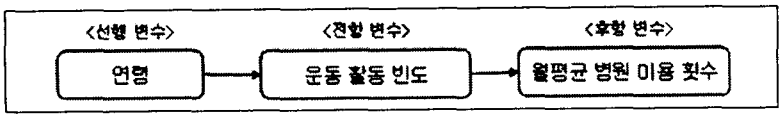
<그림 1> 결과 1에 대한 변수들의 관계

<표 3> 선행 연관 규칙 결과 2

순번	후항변수	전항변수	선행 변수	신뢰도
1	월평균 병원 이용 횟수 = 평균 이하	운동 활동 빈도 = 평균 이상	-	76.57
2	월평균 병원 이용 횟수 = 평균 이하	-	연령 = 평균 이하	74.06
3	-	운동 활동 빈도 = 평균 이상	연령 = 평균 이하	71.90
4	월평균 병원 이용 횟수 = 평균 이하	운동 활동 빈도 = 평균 이상	-	79.12
5	월평균 병원 이용 횟수 = 평균 이하	-	연령 = 평균 이하	60.86

[표 3.3]에서 보는 바와 같이 후항변수와 전항변수에 의한 연관 규칙 결과(순번 1)의 신뢰도가 76.57로 나타났고, 후항변수와 선행변수에 의한 연관 규칙 결과(순번 2)

의 신뢰도가 74.06으로 나타났으며, 전항변수와 선행변수에 의한 연관 규칙 결과(순번 3)의 신뢰도가 71.90으로 나타났다. 이는 선행 연관 규칙의 조건 1, 조건 2, 조건 3을 만족(최소 신뢰도보다 큼)하므로 후항변수를 월평균 병원 이용 횟수, 전항변수를 운동 활동 빈도, 선행변수를 연령으로 지정하였다. 선행 연관 규칙이 성립되는 지를 검정하기 위하여 선행변수인 연령을 20대와 30대로 통제한 후, 후항변수와 전항변수의 연관 규칙을 실시하였다. 규칙 결과(순번 4), 신뢰도가 79.12로 나타나 선행 연관 규칙의 조건 4를 만족(최소 신뢰도보다 큼)하는 것을 알 수 있다. 또한, 전항변수인 운동 활동 빈도를 주 2회 이하로 통제한 후, 후항변수와 선행변수의 연관 규칙을 실시하였다. 규칙 결과(순번 5), 신뢰도가 60.86으로 나타나 선행 연관 규칙의 조건 5를 만족(최소 신뢰도보다 작음)하는 것으로 나타나 선행 연관 규칙이 성립됨을 알 수 있다. 즉, 주관적 사회계층이 높은 응답자들은 교육서비스의 만족도에 대하여 불만족으로 응답함을 알 수 있으며, 선행변수인 학력이 주관적 사회계층을 선행하면서 유효한 영향력을 미치는 것으로 나타났다. 선행 연관성 규칙에 의한 변수들의 관계를 그림으로 나타내면 <그림 2>와 같다.



<그림 2> 결과 2에 대한 변수들의 관계

4. 결론

연관 규칙이란 하나의 거래나 사건에 포함되어 있는 둘이상의 품목들의 경향을 파악해서 상호 관련성을 발견하는 것으로 데이터 안에 존재하는 항목간의 종속관계를 찾아내는 기법이다. 연관 규칙은 일반적으로 가장 많이 사용되는 순차적 연관 규칙을 비롯하여, 두 변수 간의 역의 관련성을 규명할 수 있는 역 연관 규칙, 연관 규칙 중 독특하거나 드문 규칙을 찾아낼 수 있는 변칙 연관 규칙 등의 다양한 종류들이 있다.

연관 규칙 분석 시, 생성된 규칙에 대하여 변수들 간의 선행 관계가 존재한다면 그 관계에 미치는 영향을 명확히 할 필요성이 있다. 연관 규칙에서 나타난 규칙들 중 변수들 간의 선행관계를 명확하게 규명하고자 하는 것이 선행 연관 규칙이다. 선행 연관 규칙은 변수들 간의 관계를 보다 정확하고 명료하게 이해할 수 있도록 해준다. 본 논문에서는 연관 규칙에서 변수들의 관계에 미치는 영향을 명확하게 규명할 수 있는 선행 연관 규칙을 정의하고, 예제 자료에 적용해 보았다. 적용 결과, 전항변수를 선행하면서 유효한 영향력을 미치는 선행변수들을 발견할 수 있었으며, 변수들 간의 관계를 보다 정확하고 명확하게 이해할 수 있었다.

참고 문헌

1. 박희창, 조광현(2005), Environmental Consciousness Data Modeling by Association Rules, *Journal of the Korean Data & Information Science Society*, Vol. 16, No. 3, pp.529-538.
2. Agrawal R., Imielinski R., Swami A.(1993), Mining association rules between sets of items in large databases, *In Proc. of the ACM SIGMOD Conference on Management of Data*, Washington, D.C.
3. Agrawal R., Srikant R.(1994), Fast algorithms for mining association rules, *In Proc. of the 20th VLDB Conference*, Santiago, Chile.
4. Cheung D.W., Han J., Ng V., Fu A.W., Fu Y.(1996), A Fast distribution algorithm for mining association rules, *Int's Conf. on Parallel and Distributes Information System*, Miami Beach, Florida.
5. Park J.S., Chen M.S., and Philip S.Y.(1995), An effective hash-based algorithms for mining association rules, *In Proc. of ACM SIGMOD Conference on Management of Data.*, Washington, D.C.
6. Saygin Y., Vassilios S.V., Clifton C.(2002), Using Unknowns to Prevent Discovery of Association Rules, *2002 Conference on Research Issues in Data Engineering.*
7. Sergey B., Rajeev M., Jeffrey D.U., Shalom T.(1997), Dynamic itemset counting and implication rules for market data, *In Proceedings of ACM SIGMOD Conference on Management of Data.* Washington, D.C.
8. Toivonen H.(1996), Sampling Large Database for Association Rules, *Proceedings of the 22nd VLDB Conference Mumbai(Bombay)*, India.