

13강. 연관규칙분석

◆ 담당교수 : 김 동 하

■ 학습개요

데이터 안에 존재하는 항목 사이에 존재하는 유의미한 규칙을 발견하는 연관규칙분석 방법론에 대해 학습한다. 연관성을 측정하기 위해 자주 사용하는 지지도, 신뢰도, 향상도 등의 측도에 대해 공부하며, 연관규칙을 효율적인 계산을 통해 구할 수 있도록 도와주는 Apriori 알고리즘에 대해서도 다루도록 한다.

■ 학습목표

1	연관성 분석의 개념에 대해 학습한다.
2	다양한 연관성 측도에 대해 학습한다.
3	Apriori 알고리즘에 대해 학습한다.

■ 주요용어

용어	해설
연관규칙	상품을 구매하거나 서비스를 받는 등의 일련의 거래나 사건들 사이에 발생하는 연관성에 대한 규칙
지지도, 신뢰도, 향상도	두 항목 혹은 사건 사이의 연관성 정도를 측정하기 위해 많이 사용하는 측도들.
Apriori 알고리즘	최소지지도 이상의 지지도를 갖는 빈발품목집합을 효율적으로 계산하도록 도와주는 방법.

■ 학습하기

01. 연관규칙분석

연관성 분석이란

- 데이터 안에 존재하는 항목간의 연관규칙 (association rule)을 발견하는 과정
- 연관 규칙

- > 상품을 구매하거나 서비스를 받는 등의 일련의 거래나 사건들의 연관성에 대한 규칙
- > 연관성 분석을 마케팅에서 손님의 장바구니에 들어있는 품목간의 관계를 분석
- > 장바구니분석 (market basket analysis)
- 슈퍼마켓에서 구입한 고객의 물건들이 담겨져 있는 장바구니 정보를 생각하자.
- 특정한 상품을 구입한 고객이 어떤 부류에 속하는지, 그들이 왜 그런 구매를 했는지를 알기 위해서 고객들이 구매한 상품에 대한 자료를 분석하는 것.
- > 이러한 분석을 통해 효율적인 매장진열, 패키지 상품의 개발, 교차판매전략 구사, 기획상품의 결정 등에 응용할 수 있다.

연관성 분석의 응용

- 백화점이나 호텔에서 고객들이 다음에 원하는 서비스를 미리 알 수 있다.
- 신용카드, 대출 등의 은행서비스 내역으로 부터 특정한 서비스를 받을 가능성이 높은 고객의 탐지 가능.
- 의료보험금이나 상해보험금 청구가 특이한 경우 보험사기의 징조가 될 수 있고 추가적인 조사 필요.
- 환자의 의무기록에서 여러 치료가 같이 이루어진 경우 합병증 발생의 징후 탐지.

02. 연관규칙

연관 규칙의 예

- 목요일 식료품 가게를 찾는 고객은 아기 기저귀와 맥주를 함께 구입하는 경향이 있다. (유용)
- 한 회사의 전자제품을 구매하던 고객은 전자제품을 살 때 같은 회사의 제품을 사는 경향이 있다. (상식적)
- 새로 연 건축 자재점에서는 번기덮개가 많이 팔린다. (유용하지 않음.)
- 첫 번째 규칙은 유용한 규칙으로 이를 이용하여 식료품 가게의 매출을 증가시킬 수 있다.
- 두 번째 규칙은 자명한 규칙으로, 대부분의 사람들이 이미 알고 있다. 자명한 규칙의 발견은 기존의 정보를 재 확인 하는 의미가 있다.
- 세 번째 규칙은 설명이 불가능한 규칙이며, 좀더 세밀한 조사가 필요하다.
- 규칙들 중에서 유용한 규칙을 발견하는 것은 분석자의 몫.

동시 구매표 작성

- 연관성분석은 하나 이상의 제품이나 서비스를 포함하는 거래 내역을 가지고 시작.
- 연관성 분석은 분석 목적상 제조업에서 생성된 제품이나 서비스를 품목 (item)이라 한다.
- 다음의 표는 5개 제품을 취급하는 편의점에 대한 5번의 거래 내역.

고객번호	품목
1	오렌지 주스,사이다
2	우유, 오렌지 주스, 식기세척제
3	오렌지 주스, 세제
4	오렌지 주스, 세제, 사이다
5	식기 세척제, 사이다

- 동시 구매표 작성

-> 동시구매표는 대칭 행렬의 모형을 보인다.

-> 동시구매표를 보면 두 상품이 몇 번이나 함께 팔렸는지 알 수 있다.

	오렌지 주스	식기 세척제	우유	사이다	세제
오렌지 주스	4	1	1	2	2
식기 세척제	1	2	1	1	0
우유	1	1	1	0	0
사이다	2	1	0	3	1
세제	2	0	0	1	2

- 예를 들면, 사이다 행과 오렌지 주스 열이 교차하는 값을 살펴보면 두 상품이 두 번 같이 구매되었음을 알 수 있다.

- 동시구매표의 대각선 상의 자료 값은 바로 그 품목을 포함하는 총 거래 수를 나타낸다.
예를 들면, 오렌지 주스는 4번 구매되었다.

연관 규칙의 조건

- 동시구매표로 부터 간단한 규칙 (예: 사이다를 구입하는 고객은 오렌지 주스를 산다)을 만들 수 있다.

- 두 품목을 함께 산 경우는 총 5번의 구매 중 2번 일어났으며 사이다를 산 3번의 구매 중 오렌지 주스가 2번 구매되었다.

- 연관 규칙은 “If A, then B”와 같은 형식으로 표현된다.

- 모든 “if-then” 규칙이 유용한 규칙이 아니다.

- 어떤 조건을 이용하여 유용한 규칙을 추출할 수 있을까?

- 찾아진 규칙이 유용하게 사용되기 위해서는

-> 두 품목 (품목 A와 품목 B) 이 함께 구매한 경우의 수가 일정 수준 이상 이어야 하며, (일정 이상의 지지도)

-> 품목 A를 포함하는 거래 중 품목 B를 구입하는 경우의 수가 일정수준 이상 이어야 한다. (일정 이상의 신뢰도)

03. 연관규칙분석의 척도

지지도 (Support)

- 두 품목 A와 B의 지지도는 전체 거래항목 중 항목 A와 항목 B가 동시에 포함하는 거래의 비율

$$\text{지지도} = P(A \cap B) = \frac{A, B \text{가 동시에 포함된 거래 수}}{\text{전체 거래 수}}$$

신뢰도 (Confidence)

- 연관성 규칙 “If A, then B”의 신뢰도는

$$\text{신뢰도} = P(B | A) = \frac{A, B \text{가 동시에 포함된 거래 수}}{A \text{를 포함하는 거래 수}}$$

예제 1.

- 연관성 규칙 “오렌지 주스를 사면 사이다를 구매한다”의 지지도와 신뢰도를 구해보자.

-> 지지도=2/5

-> 신뢰도=2/4

연관성 규칙 “우유와 오렌지 주스를 사면 식기세척제를 산다”의 지지도와 신뢰도를 구하면

-> 지지도=1/5

-> 신뢰도=1/1

고객번호	품목
1	오렌지 주스, 사이다
2	우유, 오렌지 주스, 식기세척제
3	오렌지 주스, 세제
4	오렌지 주스, 세제, 사이다
5	식기 세척제, 사이다

예제 2.

- 세 품목 A,B,C의 동시 거래 내역 (전체: 2,000회)

항목	거래의 수	항목	거래의 수
A	100	A+C	300
B	150	B+C	200
C	200	A+B+C	100
A+B	400	추가 안함	550

- 각 품목의 조합에 대한 지지도

항목	품목이 포함된 총 거래의 수	확률	항목	품목이 포함된 총 거래의 수	확률
A	900	0.450	A+C	400	0.200
B	850	0.425	B+C	300	0.150
C	800	0.400	A+B+C	100	0.05
A+B	500	0.250			

- 모든 연관성 규칙에 대한 신뢰도

규칙	P(A*B)	P(A)	신뢰도
A B	25	45	0.556
B A	25	42.5	0.588
C B	15	40	0.375
B C	15	42.5	0.353
A C	25	45	0.556

규칙	P(A*B)	P(A)	신뢰도
C A	20	40	0.500
(A+B) C	5	25	0.200
(B+C) A	5	15	0.333
(A+C) B	5	20	0.250

- 3가지 품목을 포함하는 연관성 규칙 중 가장 높은 신뢰도를 갖는 규칙은 “B,C를 구입하면 A도 구매한다.” (신뢰도 0.333)
- 그러나, 이 연관성 규칙 실질적으로 의미 있는 규칙이 아닐 수도 있다.
- 전체 거래에서 품목 {B,C}의 거래가 일어날 가능성(지지도)은 0.15로 작기 때문.
- 지지도와 신뢰도를 함께 사용해야 함.

향상도 (Lift)

- 연관성 규칙 “A이면 B이다.”의 향상도는 다음과 같이 정의된다:

$$\text{향상도} = \frac{\text{품목 } A, B \text{를 포함하는 거래수} \times \text{전체 거래수}}{\text{품목 } A \text{를 포함하는 거래수} \times \text{품목 } B \text{를 포함하는 거래수}}$$

- 즉, 향상도는 품목 A가 주어지지 않았을 때의 품목 B의 확률 대비 품목 A가 주어졌을 때의 품목 B의 확률의 증가 비율이다.
- 품목 A와 품목 B의 구매가 상호 관련이 없다면 $P(B|A)$ 와 $P(B)$ 와 같게 되어 향상도가 1이 된다.
- 어떤 규칙의 향상도가 1보다 크면, 이 규칙은 결과를 예측하는데 있어서 우연적 기회(random chance)보다 우수하다는 것을 의미한다.
- 1보다 작으면 이 규칙이 결과를 예측하는데 있어서 우연적 기회보다 나쁘다는 것을 의미한다.

04. 연관성분석의 절차

Apriori 알고리즘

- 품목의 개수가 k 가지이면 모든 가능한 품목 조합의 수는 2^k .
- k 가 아주 큰 경우에 이 모든 집합 중에서 지지도가 높은 집합을 찾는 것은 계산적으로 불가능.
- 최소지지도보다 큰 집합 (빈발품목집합, frequent item set)만을 대상으로 높은 지지도를 갖는 품목 집합을 찾음.
- 최소지지도를 넘는 모든 빈발품목집합(frequent itemset)을 생성한다.
- 빈발품목집합에서 최소 신뢰도를 넘는 모든 규칙을 생성한다.
- 최소지지도를 넘는 모든 빈발품목집합(frequent itemset)을 생성한다.
- 빈발품목집합에서 최소 신뢰도를 넘는 모든 규칙을 생성한다.

빈발품목집합의 생성

- 개별 품목 중에서 최소 지지도를 넘는 모든 품목을 찾는다.
- 위에서 찾은 개별 품목만을 이용해서 최소 지지도를 넘는 2가지 품목 집합을 찾는다.
- 위의 두 단계에서 찾은 품목 집합을 결합하여 최소 지지도를 넘는 3가지 품목 집합을 찾는다.
- 이런 방법을 반복적으로 사용하여 최소지지도가 넘는 빈발 품목 집합들을 찾을 수가 있다.

빈발품목집합의 생성 예제

- 최소 지지도가 30% 인 빈발품목집합을 생성하자.

거래	품목
1	F, K, N
2	E, F
3	E, S
4	E, F, N
5	C, E, F, K, N
6	C, K, N
7	C, K, N

- 각 품목 C, E, F, K, N, S의 빈도가 각각 3, 4, 4, 4, 5, 1이므로 지지도는 $0.3(7 \text{ transaction} \times 0.3 = 2.1)$ 을 넘는 빈발 품목 집합(여기서는 1-빈발 품목 집합으로 표현)은

$$F_1 = \{C, E, F, K, N\}$$

- 1-빈발 품목 집합의 원소들로 구성된 가능한 2-품목 조합으로 이루어진 2-후보 품목 집합은

$$C_2 = \{\{C, E\}, \{C, F\}, \{C, K\}, \{C, N\}, \{E, F\}, \{E, K\}, \{E, N\}, \\ \{F, K\}, \{F, N\}, \{K, N\}\}$$

- C_2 의 각 품목 집합의 빈도는 1, 1, 3, 3, 3, 1, 2, 2, 3, 4이므로 최소지지도를 넘는 2-빈발 품목 집합은

$$F_2 = \{\{C, K\}, \{C, N\}, \{E, F\}, \{F, N\}, \{K, N\}\}$$

- 3-후보 품목 집합은 다음과 같다.

$$C_3 = \{\{C, K, N\}\}$$

- $\{C, K, N\}$ 의 빈도는 3으로 최소 지지도를 넘고, 따라서 3-빈발 품목 집합은 다음과 같다.

$$F_3 = \{\{C, K, N\}\}$$

연관규칙의 생성

- 빈발품목집합에 대하여 연관규칙을 생성하기 위해, 공집합을 제외한 빈발품목집합의 모든 부분 집합을 대상으로 신뢰도를 계산하고 주어진 최소 신뢰도를 넘는 연관규칙을 찾음.
- 예제에서 빈발품목집합 F_1, F_2, F_3 이 생성되면 모든 가능한 연관규칙을 생성한 후 정해진 최소 신뢰도를 넘는 연관규칙을 찾음.

05. 연관성분석의 고려사항**올바른 품목 선택**

- 어떤 품목을 선택할 것이냐는 문제는 전적으로 분석의 목적에 달려있다.
 - > 예를 들면, 대형 할인점에서는 술을 하나의 상위 품목으로 고려할 수 있다.
 - > 그러나, 어떤 경우에는 술을 세분화 하여 술을 소주, 양주, 맥주, 포도주, 막걸리 등의 술의 종류로 선택 할 수 있다.

- > 더욱 세분화하여 포도주를 적색포도주와 백색포도주로 분류할 수 있고, 또한, 제조사의 상호를 기반으로 하여 분류할 수도 있다.
- 품목의 수를 줄이는 방법으로, 품목의 분류를 상위수준으로 일반화 한다. (모든 종류의 술을 하나의 품목으로 분석)
 - 품목을 세분화 하면 결과의 활용성이 높아진다. 예를 들면, 특정한 상표의 술에 대한 정보는 그 회사의 미래 마케팅 전략에 사용될 수 있다.
 - 일반적인 방법은, 일차단계에서 상위수준의 품목 분류를 이용하여 규칙을 찾은 후 이를 바탕으로 세분화된 품목으로 분석을 진행시켜 나간다.

■ 연습문제

(객관식)1. 다음 중 연관규칙에 대해 옳지 않게 설명한 것을 고르시오.

- ① 마케팅에서 손님의 장바구니에 들어있는 품목간의 관계를 분석한다고 해서 ‘장바구니 분석’이라고도 불린다..
- ② if-then의 단순 명료한 규칙을 가지고 있다.
- ③ 연관성분석으로부터 얻어지는 모든 규칙은 유의미하다고 할 수 있다.
- ④ 품목의 수가 증가하면 연관규칙을 계산하기 위한 계산량이 폭증한다.

정답 : ③

해설 : 연관성분석으로부터 얻는 모든 규칙이 유의미한 것은 아니다. 분석자가 잘 파악하여 유의미한 규칙을 추출해야 한다.

(주관식)2. 전체 20회의 거래 중 품목 A와 품목B를 함께 산 거래 횟수가 10회라 하자. 이 때, 품목 A와 B의 지지도는 얼마인가?

정답) 0.5

해설) 10 나누기 20이므로 답은 0.5이다.

(O/X)3. Apriori 알고리즘은 최소 향상도 이상의 품목 집합들을 효율적으로 구하기 위해 고안된 알고리즘이다.

정답 : X

해설 : Apriori 알고리즘은 최소 ‘지지도’ 이상의 빈발품목집합을 효율적으로 구하기 위해 고안되었다.

■ 정리하기

1. 연관규칙분석은 구매, 서비스 이용 등의 데이터에서 항목간의 유의미한 관계를 찾아내기 위해 실시하는 분석 방법이다.
2. 연관규칙의 유의미함을 측정하기 위해 지지도, 신뢰도, 향상도 등의 측도가 사용된다.
3. 빈발품목집합을 효율적으로 계산하기 위해서 Apriori 알고리즘이 고안되었다.

■ 참고자료 (참고도서, 참고논문, 참고사이트 등)

박창이, 김용대, 김진석, 송종우, 최호식. 『R을 이용한 데이터마이닝』. 서울:교우사, 2018.