14장 연관규칙과 협업필터링

1. 연관규칙들

연관규칙 연관성분석 - '무엇이 무엇과 잘 어울리는지'를 밝혀내는 것.

↳ 서로 다른 아이템의 구매 사이에 의존성을 결정하기 위해 고객 데이터 분석에서부터 비롯

⇒ 장바구니 분석 으로도 불림.

- 후보규칙의 생성

연관규칙의 아이디어 - 아이템들 사이의 모든 가능한 규칙들을 If~ then 형식으로 열거한 후에 가장 실제 의존성을 잘 표현하는 것들만 선택.

If 부분 - 조건부 (antecedent)

Then부분 - 결론부 (Consequent) 연관분석에서 조건부와 결론부는 공통아이템을 갖지않는 아이템들의 집합

연관규칙의 첫 단계는 아이템들 사이의 연관성을 표시하는 후보가 될 수 있는 모든 규칙들을 생성

ㄴ 이상적으로는 P개의 서로 다른 아이템들로 데이터베이스 내의 모든 아이템들의 가능한 조합을 검토

→ 시간이 너무 많이 걸림.

ㄴ 현실적인 해결책은 데이터 베이스 내에서 보다 빈도수가 높은 조합만을 고려. (이걸 빈발 아이템세트라고 함)

빈발 아이템세트를 구성하는 요소의 결정 - 지지도(support)의 개념과 관련

한 규칙의 지지도 - 단순히 조건부와 결론부 모두에 포함된 아이템 세트의 거래수.

ㄴ 얼마나 많은 데이터가 해당 규칙의 타당성을 '지지'하는지를 계량

지지도는 때때로 데이터베이스 내 레코드의 총 수의 비율로 표현

빈발 아이템세트의 구성은 사용자가 지정한 최소 지지도를 초과하는 지지도를 가진 세트로 정의

- Apriori 알고리즘

하나의 아이템만으로 이루어진 빈발 아이템 세트를 생성하면서 시작한 후,

두 개, 세 개 등으로 이루어진 빈발 아이템세트를 모든 크기의 빈발 아이템세트를 생성할 때 까지 재귀적으로 반복.

ㄴ k개의 빈발 아이템세트를 생성하기 위해서 k-1 세트를 이용

ㄴ 이런 단계에서 최소 지지도를 넘지 못한 것은 제외

- 강한 규칙들의 선택

생성된 많은 규칙들 중에서 조건부의 결론부 아이템 세트 사이에 의존성이 강한 규칙만을 찾는 것

└ 한 규칙에 내포된 연관의 강도를 측정하기 위해서 '신뢰도 (Confidence)'와 '향상비 (Lift ratio)' 사용

- 지지도와 신뢰도

규칙의 신뢰도 - 데이터 베이스 내의 조건부와 결론부에 동시에 존재하는 아이템세트와 조건부에 존재하는 아이템세트를 비교.

$$\text{신뢰도} = \frac{\text{조건부와 결론부의 모든 아이템 세트가 포함된 거래수}}{\text{조건부 아이템 세트가 포함된 거래수}}$$

지지도 - 데이터 베이스에서 임의로 선택된 거래가 조건부와 결론부의 모든 아이템을 포함할 확률.

지지도 = $\hat{P}$ (조건부 AND 결론부)

신뢰도 - 임의로 선택된 거래가 조건부의 모든 아이템을 포함한다고 할 때, 결론부의 모든 아이템들도 포함할 '조건부 확률'

$$\text{신뢰도} = \frac{\hat{P}(\text{조건부 and 결론부})}{\hat{P}(\text{조건부})} = \hat{P}(\text{결론부} \mid \text{조건부})$$

신뢰도가 높을수록 강한 연관규칙 (크게 신뢰할 수 있음)을 나타낸다.

└ 조건부나 결론부의 지지도가 크다면 조건부와 결론부가 상호 독립적일 때도 신뢰도가 커질 수 있다! → 주의!!

- 향상비

연관규칙의 강도를 가늠하는 방법 - 규칙의 신뢰도를 기준값과 비교

기준값은 거래 내의 결론부 아이템 세트가 각 규칙의 조건부와 독립적이라고 가정하여 얻어진 것

조건부 - 결론부가 독립적이라면

지지도 - $P(\text{조건부 and 결론부}) = P(\text{조건부}) \times P(\text{결론부})$

$$\text{기준신뢰도} = \frac{P(\text{조건부 and 결론부})}{P(\text{조건부})} = P(\text{결론부})$$

$$= \frac{\text{결론부 아이템세트가 포함된 거래수}}{\text{데이터베이스 내의 거래수}}$$

향상비 - 신뢰도와 기준 신뢰도의 비율

$$= \frac{\text{신뢰도}}{\text{기준 신뢰도}}$$

→ 향상비가 1.0 보다 큰 규칙은 뭔가 유용

└ 독립적일 때보다 기대할 수 있는게 많다.

향상비가 클수록 연관도가 더 커짐

- 규칙 선택과정

강한 규칙을 선택하는 과정은 명기된 지지도와 신뢰도 요구조건에 맞는 모든 연관규칙들을 생성하는 것에 기반

1단계: 요구되는 지지도를 가진 '빈발 아이템세트' 찾기

2단계: 이 빈발 아이템세트에서 신뢰도 요구조건에 맞는 연관규칙들을 생성

ㄴ 1단계는 데이터베이스에서 드물게 발생하는 조합을 제거

2단계는 남은 규칙들을 선별하여 높은 신뢰도를 갖는 것들만 선택

많은 계산량이 소모되는 것은 Apriori 알고리즘을 쓰는 첫 번째 단계.

• 결과의 해석

결과를 해석할 때 다양한 측도를 살펴보는 것이 유용

- 규칙의 지지도는 전체적인 크기에 대한 영향력을 시사

ㄴ 얼마나 많은 거래에 영향을 미치는가? 적은 양만 영향을 받는다면

그 규칙은 (결론부가 매우 귀중하거나 규칙 찾기에 매우 효율적이지 않는 한) 유용성이 떨어질 수 있다.

- 향상비는 무작위 선택과 비교해서 해당 규칙이 결론부를 찾는데 얼마나 효율적인지를 보여줌.

ㄴ 매우 효율적인 규칙이 선호되지만 어떤히 지지도 고려 필요.

ㄴ 매우 낮은 지지도를 갖는 매우 효율적인 규칙은 훨씬 더 큰 지지도를 갖는 덜 효율적인 규칙만큼 바람직하지 않을 수 있다.

지지도는 어느 정도로 결론부가 찾아질 지 알려줘서 해당 규칙의 실질적 유용성을 결정하는데 유용

• 규칙과 우연

우연성에 의해서 유발될 수 있는 가짜 연관성을 평가하는 데 다음과 같은 2가지 원칙

① 보다 많은 레코드에 기반한 규칙일수록 결론이 좀 더 견고하다.

② 더 많은 분명한 규칙들을 세밀하게 고려할수록 적어도 일부가 우연한 표본추출의 결과에 근거할 가능성이 더 크다.

2. 협업필터링

사용자들의 다양한 선호도("협업")를 고려하여 방대한 양의 항목집합("필터링")으로부터 연관성이 있는 항목들을 특정 사용자에게 알려준다는 개념에 기반

ㄴ ex) 구글의 연관 검색, 쇼핑앱의 연관추천등 → 사용자의 정보뿐만 아니라 비슷한 다른 사용자의 정보에 기반하여 개인 맞춤화 제공

- 데이터 종류 및 형태

  모든 항목 - 사용자 평점가 요구됨. - 각각의 항목-사용자 조합에서 그 항목에 대한 사용자의 선호도를 측정하는 척도가 필요.

  선호도는 점수화된 평가 또는 구매, '좋다' 또는 클릭과 같은 이진화의 행동이 될 수 있음.

  n명의 사용자 $(u_1, u_2, \cdots, u_n)$와 p개의 항목 $(i_1, i_2, \cdots, i_p)$에 대한 데이터는 n행 p열의 $n \times p$ 행렬로 생각할 수 있음. → 행렬의 각 셀은 각 항목에 대한 특정 사용자의 선호도를 나타냄

  └ 모든 사용자가 모든 물품에 점수를 매기는 것이 아니기에 행렬은 결측치를 많이 포함하고 있음

  └ 이 결측값은 때때로 '흥미없음'을 의미하기도 함

  n, p가 클 경우 선호도 $(r_{u,i})$를 행렬로 표현하는 것은 비효율적.

  └ 데이터를 행렬 대신에 하나의 행으로 포현하는 것이 효과적. $(u_u, I_i, r_{u,i})$ (항목, 사용자 ID, 선호도 정보)

- 사용자 기반 협업필터링: "People like You"

  비슷한 선호도를 가진 사람들을 찾는 것과 그들이 좋아하지만 아직 구매하지않은 항목들을 추천하는데 기반

  ① 관심대상의 사용자와 가장 비슷한 사용자(이웃)를 찾는다. 이를 위해선 사용자의 선호와 다른 사용자들의 선호를 비교한다.

  ② 오직 사용자가 아직 구매하지않은 항목들만을 고려하고, 그 사용자의 이웃들이 가장 선호하는 것을 추천한다.

  1단계는 사용자와 다른 사용자들 간의 거리를 측정하는 거리(또는 근접성) 측도의 선택이 필요.

  거리들이 계산된 이후에는 한계점을 거리 혹은 필요한 이웃들의 개수에 적용하여

  2단계에서 이용될 최근접이웃들을 결정하는데 사용할 수 있다. ⇒ "사용자 기반 최우선 -N추천" (user-based top N - recommendation)

  최근접이웃 방법은 k-최근접이웃 알고리즘과 유사하게 데이터베이스에 있는 다른 사용자들과 사용자들의 거리를 측정. → 기존의 유클리드거리는 성능 안좋음

  두 사용자 간의 대표적인 근접성 평가방법은 각각의 평가 간의 피어슨(Pearson) 계수

사용자 U_1의 품목 $I_1, I_2, \cdots, I_p$에 대한 선호도 $r_{1,1}, r_{1,2}, \cdots, r_{1,p}$: 평균 $\overline{r_1}$

사용자 U_2의 품목 $I_1, I_2, \cdots, I_p$에 대한 선호도 $r_{2,1}, r_{2,2}, \cdots, r_{2,p}$: 평균 $\overline{r_2}$

두 사용자 간의 상관근접성 (Correlation Proximity)

$$Corr(U_1, U_2) = \frac{\sum (r_{1,i} - \overline{r_1})(r_{2,i} - \overline{r_2})}{\sqrt{\sum (r_{1,i} - \overline{r_1})^2} \times \sqrt{\sum (r_{2,i} - \overline{r_2})^2}}$$

→ 두 사용자 모두 포함된 항목들만 계산

↳ 각 사용자의 평균은 모든 데이터에 대해. 상관 계수 계산은 두 사용자 모두에게 평가된 항목만!

또 다른 대표적 측정방법은 피어슨 상관계수에서 파생된 코사인 유사도 (cosine similarity)

↳ 평균을 빼지 않는다는 점이 상관계수 식과 다른 점

관심대상인 사용자에 대해 상관계수, 코사인 유사도 혹은 다른 측정방법을 사용하여 데이터 베이스에 있는 다른 사용자들과의 유사도를 각각 계산.

2단계에서는 k-최근접 사용자들을 관찰하고, 그들이 평가하고 구매한 다른 모든 항목들 중에서 최고 항목을 선정하여 관심대상의 사용자에게 추천.

무엇이 가장 좋은 추천인가? ⇒
- 이진화 (binary) 구매데이터의 경우 가장 많이 구매된 것
- 평가점수의 경우 가장 높은 평가나 가장 많은 평가 혹은 가중치

최근접이웃 방법은 규모가 큰 사용자 데이터베이스인 경우 계산량이 많아질 수 있다.

↳ 군집 방법론을 적용하여 사용자들의 선호에 따라 동종의 군집으로 그룹화 하고,

각각의 군집들과 관심대상 사용자 간의 거리를 측정 → 많은 계산량을 군집단계에 집중 (미리 실행 가능)

↳ 해당 사용자와 각각의 군집을 동시에 비교함으로써 효율적이고 빠름

↳ 단점은 가장 가까운 군집의 멤버들이 해당 사용자와 가장 비슷한 것은 아니기에 비교적 덜 정확한 추천일 수가 있음

- 항목 기반 협업필터링

사용자들의 수가 항목들의 수보다 훨씬 큰 경우, 비슷한 사용자들보다 비슷한 항목들을 찾는 것이 계산적으로 효율적이고 빠르다.

구체적으로 사용자가 특정 항목에 관심을 표현하면, 항목기반 협업필터링 알고리즘은

① (임의의 사용자가) 관심을 가지는 항목과 공동으로 평가 혹은 구매한 항목들을 찾는다.

② 비슷한 항목들 중에서 가장 대등하거나 상관관계가 높은 항목을 추천.

↳ 이러면 유사도는 사용자들 대신에 항목들 사이에서 계산

모든 항목들 간의 유사도를 미리 계산할 수 있으며 양의 상관계수를 실시간으로 추천 가능

└ 단점은, 항목들간의 다양성이 적어서 추천들이 뻔해질 수 있음

• 협업필터링의 장점과 취약점

협업필터링은 사용자들의 선호도와 주관적인 정보에 의존

└ 만약 DB가 비슷한 사용자들을 충분히 보유하고 있다면 비주류 항목들에 대해서도 (Long Tail)
└ 많을 필요는 없지만 적어도 사용자별로 어느정도
유용한 추천을 제공하여 각각의 사용자가 비슷한 취향의 다른 사용자들을 찾을 수 있다

비슷하게 데이터는 항목별 평가나 구매내역을 충분히 보유해야함.

⇒ 협업필터링의 한 가지 한계점은 새로운 사용자들이나 새로운 항목에 대한 추천 불가능 (이 문제를 해결하기 위한 다양한 방법이 있음)

사용자 기반 협업필터링은 높게 평가되거나 선호하는 항목들의 유사도를 찾는다.

└ 그러나 낮게 평가되거나 원하지 않는 항목들의 데이터는 고려 X. → 원하지 않는 항목 탐색에는 사용 X

사용자 기반 협업필터링은 개인 맞춤형 추천을 제공하기 위한 사람들의 취향 유사도 파악에 도움

└ 사용자 수가 너무 많아지면 계산이 어려워짐. → 항목기반 알고리즘, 사용자 군집화, 차원축소 등으로 해결

'예측'이라는 용어가 사용되기도 하지만 본질적으로 비지도학습 기법. → 실제 결과값 X
사용자들의 피드백으로 개선 가능

• 협업필터링 vs 연관규칙

둘 다 추천을 생성하는 비지도학습 방법이지만 여러 관점에서 차이

- 빈발 아이템세트 vs 개인 맞춤형 추천

연관규칙 ~ 빈발 항목의 조합을 찾으며 오로지 찾은 항목들에 대한 추천 제공

협업필터링 - 모든 항목들에 대해 개인 맞춤형 추천 제공, 특히 선호도 가진 사용자에게까지 제공

⇒ 협업필터링은 선호도의 비주류 (long-tail)을 포착, 연관규칙은 주류(Head)를 찾음

└ 이러한 차이점은 필요한 데이터에 대해 다음과 같은 함축성을 지님

연관규칙은 항목들의 특정 조합을 포함하는 충분한 수의 장바구니(장바구니 거래데이터)를 찾기위해 수많은 거래데이터 필요

협업필터링은 많은 '장바구니'가 필요하진 않지만, 여러 사용자로 부터 최대한 많은 항목의 데이터 필요.

연관규칙은 장바구니 레벨에서 적용 – (포괄적이고 객관적인 규칙 생성. 상점 제품 배치, 행원의 린단검선 순서 결정 등에 응용)

협업필터링은 사용자 레벨에서 적용 – 특정 사용자를 위한 추천을 생성 ⇒ 개인 맞춤형 도구

- 거래 데이터 vs 사용자 데이터

연관규칙 - '여러 거래/장바구니' 안에 있는 다른 항목과의 공동 구매내역을 기반으로 추천

협업필터링 - 특성도 있는 다른 '사용자'와의 공동 구매내역 혹은 평가 점수가 기반

- 이전데이터 및 평가점수 데이터

연관규칙은 각 항목을 이진데이터로 처리, 협업필터링은 이진데이터 및 수치화된 평가점수 데이터 모두 활용

- 두 개 이상의 항목

연관규칙 - 조건부, 결론부 모두 한 개 이상의 항목 포함 가능

⇒ 하나의 추천은 여러 항목으로 이루어진 하나의 묶음

협업필터링 - 두 항목 혹은 두 사용자 간의 유사도가 측정

⇒ 단일 항목이나 각 항목끼리 전혀 관련없을수도 있는 여러 단일 항목이 추천됨

ㄴ 이러한 차이들은 비인기항목의 구매 및 추천에 대해, 연관규칙과 사용자 기반 협업필터링을 비교 시 더 잘 드러남.

3. 요약

연관규칙, 협업필터링 → 거래 데이터베이스에서 구매된 아이템들 사이의 연관성 추론위한 비지도학습

연관규칙 - 'IF x 구매 then y도 구매'와 같은 명확하고 간단한 규칙 생성 → 방법이 매우 명료. 이해하기 쉬움

2단계로 구성 (후보 규칙들의 생성 - 신뢰도, 지지도에 근거한 규칙 평가)

⇒ 생성되는 규칙이 너무 많은게 단점 → 유용하고 강한 규칙들로 이루어진 작은 집합으로 줄이기 위한 방법 필요

정보를 농축하기 위한 중요한 비자동기법은 정보가 없거나 사소한 규칙들뿐 아니라 동일한 지지도를 갖는 규칙 조사

드문 조합은 최소 지지도 조건을 못 맞출 가능성이 큼 → 데이터 상에서 동일한 빈도를 가지는 항목을 갖는게 낫다.

협업필터링 - 항목을 구매 혹은 평가 하는 등의 비슷한 행동을 한 사용자들로부터 형성된 항목 간의 관계에 기반

효과적인 사용을 위해서는 사용자들의 피드백과 사용자들이 각 항목에 대한 충분한 정보 필요

ㄴ 단점은 새로운 사용자나 항목에 대한 추천 불가능