Regras de associação e itens frequentes

Jacques Wainer

IC – Unicamp

Outubro 2015

O mito das faldas e cerveja

Uma rede de supermercados (normalmente WallMart) usando mineração de dados descobre que há uma estranha correlação entre compra de cerveja e compra de fraldas. Em algumas versões a rede coloca um estande de cerveja ao lado das fraldas. ▸ link

Em outra versão, a rede coloca cerveja e faldas juntos.

As técnicas de mineração de regras de associação e de conjunto de intens (*itemset*) frequentes é que permitem tirar este tipo de conclusão.

Carrinho de compras

- neste tipo de problema há um conjunto de itens (itens num supermercado) e há transações que contem um subconjunto dos itens (uma compra).
- os itens podem ser paginas num site, a a transação as paginas visitadas em diferentes interações com o site.
- o conceito de transação pode não ser localizado no tempo. Pode ser uma pessoa, e os itens aplicativos que ela instalou no seu celular (não necessariamente ao mesmo tempo).
- ou pode ser proteinas ativas em diferentes tecidos de diferentes individuos (uma transação é a combinação de tecido e individuo).

Itemsets frequentes

Itemsets frequentes são conjunto de itens que aparecem juntos em pelo menos $s\%$ das transações. O número s, que precisa ser fornecido para o algoritmo é chamado de **suporte**

Vamos assumir as seguintes transações

- A B C
- A C
- C D
- A B
- B D
- D

Se o suporte é 1/3, ou seja queremos conjuntos de itens que aparecem em pelo menos 2 das 6 transações, então A B é **um** dos itemsets frequentes. A B aparece como parte da primeira transação (mas A B não é a transação completa) e aparece como "parte" da 4 transação (e neste caso é a transação completa).

Itemsets frequentes

Há vários itemsets frequentes:

- Se A B é um itemset frequente, então tanto A quanto B são também!
- se um itemset tem n itens, então todas as $2^n - 2$ combinações dos itens também são itemsets frequentes.
- no caso das transações do slide anterior , A C é um itemset frequente (e portanto C também é) e D.
- no exemplo anterior

Itemsets frequentes

Portanto é necessario algumas restrições nos itemsets que serão retornados pelos algoritmos

- um itemset *i* é **maximal** se ele tem suporte maior que *s* e todos os itemsets que incluem *i* tem suporte menor que *s*.
- ou seja, não da para incluir mais nenhum item num itemset maximal e ainda ter um itemset com suporte maior que *s*
- No exemplo anterior voltar A B, A C, e D são maximais.
- Um itemset *i* é **fechado** (*closed*) se todos os itemsets que o incluem tem suporte menor que *i*
- no exemplo anterior voltar todos os itemsets são fechados.

Regras de associação

Regras do tipo

Regras de associação

A B => C D

onde A B C D são itens. A **confiança** da regra (c) é a proporção das transações que incluem A B e que também incluem C D. Assim se a confiança da regra acima é 60% a regra pode ser lida como

Regras de associação

60% das pessoas que compraram A e B também compraram C e D

Regras de associação

A confiança de uma regra é formalmente

$$\text{confianca}(\alpha => \beta) = \frac{\#(\alpha\beta)}{\#(alpha)}$$

onde $\#(\alpha)$ é o numero de transações que incluem o itemset α. portanto

$$\text{confianca}(\alpha => \beta) = \frac{\text{suporte}(\alpha\beta)}{\text{suporte}(alpha)}$$

Regras de associação

Como usar um regra de associação A B => C D?

- automaticamente: quando o cliente comprar A e B sugerir C e D. Neste caso queremos regras com alta confiança e regras do tipo A B => C
- KDD: como uma regra para entender o problema. Neste caso queremos tanto alta confiança como alto suporte (não gastar o seu tempo entendendo um fenomeno que só acontece com 1 em dez milhões das transações!)
- para KDD há outras considerações: quão interessante é a regra?

Exemplo

	café	não café	total
chá	150	50	200
não chá	650	150	800
total	800	200	1000

- a regra Chá => Café tem suporte $150/1000 = 0.15$ (alto)
- a confiança da regra é $0.15/0.2 = 75\%$ também alto.
- mas é regra não é interessante e é mesmo enganadora. a regra $\{\} =>$ Café tem confiança de 80%, isto é 80% das pessoas já bebem café. O fato delas tomarem chá *diminue* a probabilidade delas tomarem café!.

Medidas de interesse de regras

- o *lift* mede o quanto não independentes são os 2 lados da regra $\alpha => \beta$.

$$\text{lift}(\alpha => \beta) = \frac{P(\alpha\beta)}{P(\alpha)P(\beta)} \quad (1)$$

- onde $P(\alpha)$ é a probabilidade de α que é $\#(\alpha)/n$. lift =1 indica que α e β são independentes (o que usualmente não é nada interessante)
- lift > 1 indica uma correlação positiva entre α e β.

Medidas de interesse de regras

- Voce quer regras com lift $>>$ 1 (ou muito perto do 0).
- qual um valor minimo de lift?
- há outras medidas de "quão interessante" é uma regra (um artigo que lista e compara 21 diferentes medidas ▸ link)

Descobrindo regras de associação

- Normalmente, para aplicações não automáticas de regras de associação, exige-se um suporte mínimo (para o itemset), uma confiança minima e talvez um lift mínimo.
- Normalmente usa-se um algoritmo para achar itemsets frequentes (portanto que tem suporte mínimo) e destes itemsets gera-se as regras que tem confiança minima e talvez lift minimo.
- há vários algoritmos para minerar itens frequentes: a priori, eclat, fp-grow

Regras e itemsets no R

- Pacote arules link
- implementa o algoritmo apriori link para encontrar regras e itemsets frequentes
- e o algoritmo eclat link apenas para itemsets
- ja contem alguns bancos de dados de transações

Regras e itemsets no R

```
> data("Groceries")
> Groceries
> as(Groceries,"list")[1:5]
> it1=eclat(Groceries,parameter=list(supp=0.04,target="maximal
> it1
> inspect(it1)
> it2=eclat(Groceries,parameter=list(supp=0.04,target="maximal
```

Regras e itemsets no R

```
> r1=apriori(Groceries,parameter=list(supp=0.04,target="rules"
> r1=apriori(Groceries,parameter=list(supp=0.04,target="rules"
> inspect(r1)
```

Regras e itemsets no R

```
> data("AdultUCI")
> head(AdultUCI)
> data("Adult")
> r2=apriori(Adult,parameter=list(supp=0.6))
> inspect(r2)
```

Tarefa

- Leia o dataset em `http://fimi.ua.ac.be/data/retail.dat`que é um dataset real de compras no varejo em uma loja na Belgica. Utilize a função `read.transactions` ▸ link do arules.
- descruba as regras de associação que tenham suporte mínimo de 0.005 e confiança minima de 0.9

Outras idéias em itens frequentes

Mineração de sequências frequentes

- Cada transação, além dos itens possui informação do cliente e da data de compra.
- Voce quer obter regras de associação do tipo 70% dos clientes (confiança) que compram A, acabam comprando B em 2 meses.
- O não-mito do Target e da filha grávida link. Isso poderia ter descoberto de mineração de sequências frequentes (mas o artigo dá a entender que o Target usou outras técnicas).
- o pacote `aruleSequences` link implementa um algoritmo de mineração de sequencias

Outras idéias em itens frequentes

Itemsets minimalmente infrequentes:

- O itemset α tem suporte menos que s mas todos os subconjuntos de α tem suporte maior que s.
- Coisas infrequentes podem estar associadas a fraudes, ou podem gerar quebra de privacidade