Universidade Federal do Maranhão
Departamento de Informática da UFMA - DEINF
Coordenação do curso de Ciência da Computação - COCOM

Análise método Apriori

Relatório feito para analisar os resultados obtidos após aplicação do método Apriori de dados fornecidos pelo professor da disciplina de Engenharia da Informação.

São Luís - MA
2021

1. Introdução

As Regras de associação tem como objetivo principal encontrar determinados elementos que impliquem em outros elementos em uma estipulada transação. São frequentemente utilizadas em campanhas de marketing, controle de estoque de armazém, descobrir o comportamento de compra de clientes em lojas, ações na interface de usuário, etc. Um exemplo de regra de associação é a compra de um produto quando um outro determinado produto é comprado, pode haver aí uma regra de associação forte e por isso é necessário que se faça uma análise dessa informação.

O algoritmo Apriori foi criado por "Agrawal e Srikant" da Índia em 1994. Esse método foi projetado para operar dados de transações onde cada transação é vista como um conjunto de itens. Esse algoritmo fornece um conjunto de regras de associação entre os elementos ou itens a partir dos dados fornecidos onde leva em consideração o **suporte**, **confiança** e **elevação.**

2. Objetivo

Esse relatório tem como objetivo apresentar uma série de resultados para serem examinados utilizando o método Apriori e descrever como foram realizada as etapas de pré-processamento de dados, obtenção dos conjuntos de itens mais frequentes, obtenção das regras de associação, estudo do suporte, confiança e de lifit e descrever um parecer sobre os resultados obtidos e analisados.

3. Materiais e Métodos

O material para análise foi um conjunto de listas de compras de clientes em CSV de um supermercado. A lista contém os nomes de cada item que um determinado cliente comprou. Esse conjunto é composto por 9835 listas de compras.

```
citrus fruit,semi-finished bread,margarine,ready soups,,,,,,,,,,,,,,,,,,,,,,,,,,,
tropical fruit,yogurt,coffee
whole milk
pip fruit,yogurt,cream cheese,meat spreads
other vegetables,whole milk,condensed milk,long life bakery product
whole milk,butter,yogurt,rice,abrasive cleaner
rolls/buns
other vegetables,UHT-milk,rolls/buns,bottled beer,liquor (appetizer)
potted plants
whole milk,cereals
tropical fruit,other vegetables,white bread,bottled water,chocolate
citrus fruit,tropical fruit,whole milk,butter,curd,yogurt,flour,bottled water,dishes
beef
frankfurter,rolls/buns,soda
chicken,tropical fruit
butter,sugar,fruit/vegetable juice,newspapers
fruit/vegetable juice
packaged fruit/vegetables
chocolate
specialty bar
other vegetables
butter milk,pastry
whole milk
tropical fruit,cream cheese,processed cheese,detergent,newspapers
tropical fruit,root vegetables,other vegetables,frozen dessert,rolls/buns,flour,sweet spreads,salty
snack,waffles,candy,bathroom cleaner
bottled water,canned beer
yogurt
sausage,rolls/buns,soda,chocolate
other vegetables
```

conjunto de listas de compras no formato .csv

Para ajudar na extração e análise das regras de associação, utilizaremos as seguintes linguagens, bibliotecas e softwares:

- Python
- Pandas
- Mlxtend
- Numpy
- Jupyter Notebook

4. Análise descritiva

O conjunto analisado consiste em 9865 listas de compras de produtos em um supermercado, contudo, essa lista precisou passar por um pré-processamento e remoção de dados não importantes.

Out[3]:

	0	1	2	3	4	5	6	7	8	9	...	22	23	24	25	26	27	28	29	30	31
0	citrus fruit	semi-finished bread	margarine	ready soups	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	tropical fruit	yogurt	coffee	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	whole milk	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	pip fruit	yogurt	cream cheese	meat spreads	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	other vegetables	whole milk	condensed milk	long life bakery product	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
5	whole milk	butter	yogurt	rice	abrasive cleaner	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
6	rolls/buns	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
7	other vegetables	UHT-milk	rolls/buns	bottled beer	liquor (appetizer)	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
8	potted plants	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
9	whole milk	cereals	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

10 rows × 32 columns

conjunto antes da remoção de dados não relevantes

Podemos ver na figura de um dataframe acima que o mesmo possui vários valores zerados, e isso seria irrelevante para nossa aplicação, logo a remoção desses valores se fez necessária.

```
Out[14]: [['citrus fruit', 'semi-finished bread', 'margarine', 'ready soups'],
          ['tropical fruit', 'yogurt', 'coffee'],
          ['whole milk'],
          ['pip fruit', 'yogurt', 'cream cheese', 'meat spreads'],
          ['other vegetables',
           'whole milk',
           'condensed milk',
           'long life bakery product'],
          ['whole milk', 'butter', 'yogurt', 'rice', 'abrasive cleaner'],
          ['rolls/buns'],
          ['other vegetables',
           'UHT-milk',
           'rolls/buns',
           'bottled beer',
           'liquor (appetizer)'],
          ['potted plants'],
          ['whole milk', 'cereals'],
          ['tropical fruit',
           'other vegetables',
```

conjunto de listas sem os valores zerados

Após isso, passamos por um próximo pré-processamento que consiste em criar um dataframe onde os produtos são as colunas e os valores das linhas são booleanos, que nos mostra que se um determinados produto está na lista, ele vai apresentar True como valor, caso contrário ele vai apresentar False.

Out[15]:

	Instant food products	UHT-milk	abrasive cleaner	artif. sweetener	baby cosmetics	baby food	bags	baking powder	bathroom cleaner	beef	...	turkey	vinegar	waffles	whipped/sour cream	whisky	white bread	wi w
0	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	Fi
1	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	Fi
2	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	Fi
3	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	Fi
4	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	Fi
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
9830	False	False	False	False	False	False	False	False	False	True	...	False	False	False	True	False	False	Fi
9831	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	Fi
9832	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	Fi
9833	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	Fi
9834	False	False	False	False	False	False	False	False	False	False	...	False	True	False	False	False	False	Fi

9835 rows × 169 columns

dataframe booleano

Dados os pré-processamentos necessários, podemos assim criar as nossas regras de associação. Primeiramente utilizaremos a biblioteca *mlxtend.frequent_patterns* com a função *apriori* para extrair a frequência em que cada item aparece nas listas e o seu suporte.

In [27]:
```
from mlxtend.frequent_patterns import apriori, association_rules

# extraindo os itens mais frequentes
frequent_itens = apriori(df, min_support = 0.01, use_colnames = True)

# retorna uma lista de itens mais frequentes ordenada de maneira do maior para o menor support
frequent_itens.sort_values(by=['support'], ascending = False)
```

Out[27]:

	support	itemsets
86	0.255516	(whole milk)
55	0.193493	(other vegetables)
66	0.183935	(rolls/buns)
75	0.174377	(soda)
87	0.139502	(yogurt)
...	...	...
178	0.010066	(frankfurter, sausage)
306	0.010066	(yogurt, curd, whole milk)
160	0.010066	(curd, rolls/buns)
212	0.010066	(napkins, tropical fruit)
199	0.010066	(hard cheese, whole milk)

333 rows × 2 columns

função apriori e resultados de itens frequentes e seu suporte

A função apriori necessita que se passe por parâmetro um dataframe e um suporte mínimo para suas operações que retornar a quantidade de frequência de itens, então foi determinado que os itens que aparecem em no mínimo em 1% das listas poderiam entrar na lista de itens frequentes.

Uma associação com maior suporte significa que ela aparece mais vezes dentro de uma base de dados, o que significa que caso ela apareça muito mais vezes a probabilidade dela ser relevante pro nosso cálculo é muito maior, por isso é definido um valor de suporte mínimo para que se evite a inserção de itens não relevantes em nossa análise. No caso do nosso conjunto, os itens que mais aparece na lista é o (Whole Milk), que aparece em 25% dos itens comprados, seguido por (Other vegetables) com 19% e (rolls/buns) com 18%.

Abaixo uma tabela com os 5 conjuntos de itens de maior suporte dependendo da quantidade de itens dentro desse conjunto.

1 item

suporte	itens
0.255516	(Whole Milk)
0.193493	(Other vegetables)
0.183935	(rolls/bund)
0.174377	(soda)
0.139502	(yogurt)

2 itens

suporte	itens
0.074835	(Other vegetables,Whole Milk)
0.056634	(whole milk, rolls/buns)
0.056024	(yogurt, whole milk)
0.048907	(root vegetables, whole milk)
0.047382	(root vegetables, other vegetables)

Tabela 1

3 itens

suporte	itens
0.023183	(root vegetables, other vegetables, whole milk)
0.017895	(other vegetables, whole milk, rolls/buns)
0.017082	(tropical fruit, other vegetables, whole milk)
0.015557	(yogurt, whole milk, rolls/buns)
0.015150	(tropical fruit, yogurt, whole milk)

Tabela 2

Agora utilizaremos diferentes valores para a confiança para assim analisar as associações, e para tal utilizaremos a função da biblioteca *frequent_patterns* do *mlxtend* chamada de *association_rules*, que é uma espécie de função construtora de regras associativas e por parâmetro iremos passar os nossos itens mais frequentes, a métrica e o valor da métrica. A métrica que usaremos é a confiança e os valores que usaremos serão: **0.3, 0.5, 0.7**.

A função *association_rules* irá retornar vários valores como o suporte do antecedente, suporte do consequente, convicção, etc. Porém, usaremos apenas os valores de antecedentes, consequentes, suporte, confiança e lift para a análise, pois os demais não seriam de tão relevância pro momento.

Veja abaixo o ranking das 5 regras associativas mais fortes e as 5 regras associativas mais fracas de acordo com cada valor de confiança de acordo com a confiança.

- **confiança 0.3**
    - Para esse nível de confiança foram mantidas 125 regras de associação

Tabela contendo 5 associações mais fortes com confiança 0.3:

antecedentes	consequentes	suporte	confiança
(root vegetables, citrus fruit)	(other vegetables)	0.010371	0.586207
(root vegetables, tropical fruit)	(other vegetables)	0.012303	0.584541
(curd, yogurt)	(whole milk)	0.010066	0.582353
(other vegetables, butter)	(whole milk)	0.011490	0.573604
(root vegetables, tropical fruit)	(whole milk)	0.011998	0.570048

Tabela 3

Tabela contendo 5 associações mais fracas com confiança 0.3:

antecedentes	consequentes	suporte	confiança
(yogurt)	(other vegetables)	0.043416	0.311224
(bottled water)	(whole milk)	0.034367	0.310948
(other vegetables, whole milk)	(root vegetables)	0.023183	0.309783
(berries)	(other vegetables)	0.010269	0.308869
(rolls/buns)	(whole milk)	0.056634	0.307905

Tabela

- **Confiança: 0.5**
    - Para esse nível de confiança apenas 15 regras de associação foram mantidas.

Tabela contendo 5 associações mais fortes com confiança 0.5

antecedentes	consequentes	suporte	confiança
(root vegetables, citrus fruit)	(other vegetables)	0.010371	0.586207
(root vegetables, tropical fruit)	(other vegetables)	0.012303	0.584541
(curd, yogurt)	(whole milk)	0.010066	0.582353
(other vegetables, butter)	(whole milk)	0.011490	0.573604
(root vegetables, tropical fruit)	(whole milk)	0.011998	0.570048

tabela 5

Tabela contendo 5 associações mais fracas com confiança 0.5:

antecedentes	consequentes	suporte	confiança
(yogurt, tropical fruit)	(whole milk)	0.015150	0.517361
(other vegetables, yogurt)	(whole milk)	0.022267	0.512881
(whipped/sour cream, other vegetables)	(whole milk)	0.014642	0.507042
(root vegetables, rolls/buns)	(other vegetables)	0.012201	0.502092
(root vegetables, yogurt)	(other vegetables)	0.012913	0.500000

Tabela 6

- **Confiança: 0.7**
    - Para esse nível de confiança a análise não achou nenhuma associação, ainda foi testado 0.6 e mesmo assim não houve resultado, portanto o nível de confiança dessas regras de associação não vai além de 0.586207.

O lift, também conhecido como medida de surpresa, é uma medida que busca encontrar associações com muito mais frequência. Por exemplo, o lift nos dirá qual a chance de X ser comprado após o Y ser comprado, considerando assim toda a popularidade do item Y.

O lift > 1 é o fator em que a ocorrência de Y potencializa a probabilidade que que X ocorra, então quanto maior o lif, maior a força daquela regra associativa.

Foi feita uma nova análise, agora utilizando o lift para verificar qual a regra associativa com maior força, utilizando os valores “0.3 e 0.5” como parâmetro de confiança, obtivemos os seguintes resultados:

- **Tabela com as 10 regras associativas mais fortes de acordo com o lift com valor de confiança em 0.3**

antecedentes	consequentes	suporte	confiança	lift
(other vegetables, citrus fruit)	(root vegetables)	0.010371	0.359155	3.295045
(other vegetables, tropical fruit)	(root vegetables)	0.012303	0.342776	3.144780
(beef)	(root vegetables)	0.017387	0.331395	3.040367
(root vegetables, citrus fruit)	(other vegetables)	0.010371	0.586207	3.029608
(root vegetables, tropical fruit)	(other vegetables)	0.012303	0.584541	3.020999
(other vegetables, whole milk)	(root vegetables)	0.023183	0.309783	2.842082
(curd, whole milk)	(yogurt)	0.010066	0.385214	2.761356
(root vegetables, rolls/buns)	(other vegetables)	0.012201	0.502092	2.594890

(root vegetables, yogurt)	(other vegetables)	0.012913	0.500000	2.584078
(whole milk, tropical fruit)	(yogurt)	0.015150	0.358173	2.567516

Tabela 7

- **Tabela com as 10 regras associativas mais fortes de acordo com o lift com valor de confiança em 0.5**

antecedentes	consequentes	suporte	confiança	lift
(root vegetables, citrus fruit)	(other vegetables)	0.010371	0.586207	3.029608
(root vegetables, tropical fruit)	(other vegetables)	0.012303	0.584541	3.020999
(root vegetables, rolls/buns)	(other vegetables)	0.012201	0.502092	2.594890
(root vegetables, yogurt)	(other vegetables)	0.012913	0.500000	2.584078
(curd, yogurt)	(whole milk)	0.010066	0.582353	2.279125
(other vegetables, butter)	(whole milk)	0.011490	0.573604	2.244885
(root vegetables, tropical fruit)	(whole milk)	0.011998	0.570048	2.230969
(root vegetables, yogurt)	(whole milk)	0.014540	0.562992	2.203354
(domestic eggs, other vegetables)	(whole milk)	0.012303	0.552511	2.162336
(whipped/sour cream, yogurt)	(whole milk)	0.010880	0.524510	2.052747

Tabela 8

Os valores de confiança acima de 0.3 não nos retornam um lift que seja menor do que 1, então, foi utilizado o valor 0.2 onde retorna 3 valores onde o lift é menor do que um e pode ser considerado com uma confiança alta.

Tabela com regras associativas com lift menor do que 1 e confiança alta

antecedentes	consequentes	suporte	confiança	lift
(bottled beer)	(whole milk)	0.020437	0.253788	0.993237
(shopping bags)	(whole milk)	0.024504	0.248710	0.973364
(soda)	(whole milk)	0.040061	0.229738	0.899112

Tabela 9

5. Conclusão

Podemos verificar que o método Apriori é realmente uma excelente escolha para se prever possíveis comportamentos, e assim sendo, somos capazes de verificar que os itens que mais aparecem com frequência dentro dos dados de compras que nos foi fornecido, muito provavelmente, estarão em destaque dentro das regras associativas com maior força.

Em destaque, podemos observar que os conjuntos de itens que mais se parecem, ou que tem algo haver com seu conteúdo, tem a probabilidade de estarem associados com outros produtos parecido, por exemplo, na tabela 8 nós temos dados onde os valores estão ordenados do maior para o menor lift, os primeiros itens da tabela, além de possuírem um valor consideravelmente alto para a confiança, possui um lift alto, que reforça ainda mais de que essa associação é realmente forte.

Dentro da tabela 8 nós temos a seguinte associação (root vegetables, citrus fruits) (other vegetables), que quer dizer que quem compra esse combo de root vegetable e citrus fruits provavelmente irá comprar outros vegetais, e o seu lift alto de 3.029608 e confiança de 0.586207 só reforçam a tese de que isso realmente irá ocorrer. Outro exemplo na mesma tabela é o item “curd” que é uma coalhada, que é um produto de origem animal proveniente do leite bovino, e o conjunto (curd, yogurt) está associado à compra de whole milk, que é um laticínio, tal qual o conjunto antecedente.

Analisando as tabelas, vemos também a ocorrência de determinadas regras com possíveis confianças altas porém lift abaixo de 1. O lift abaixo de 1 determina que muito provavelmente essa regra não irá ocorrer, porém a confiança um pouco elevada pode ser um indício que na verdade essa correlação entre associações pode ter ocorrido, mas ocorreu tão pouco que mesmo com a confiança elevada ela pode ser considerada uma informação não relevante para a análise.

6. Referência

Sistemas de Recomendação com Apriori (Prática com Python) - Machine Learning 23.2. Publicado pelo canal Universo Discreto. Disponível em: https://www.youtube.com/watch?v=Mq5HPAFXrOI. Acesso em: 12 de setembro de 2021.

Sistemas de Recomendação e Regras Associativas com Apriori (Teoria) - Machine Learning 23.1. Publicado pelo canal Universo Discreto. DIsponível em: https://www.youtube.com/watch?v=YGEYty0xYc0. Acesso em: 12 de setembro de 2021.

Apriori Algorithm Explained | Association Rule Mining | Finding Frequent Itemset | Edureka. Publicado pelo canal edureka!. Disponível em: https://www.youtube.com/watch?v=guVvtZ7ZClw. Acesso em: 13 de setembro de 2021.

FRANCYLES, Italo. Aula16 - sr baseados em descoberta de conhecimento. 26-26 de aug de 2021. 15 p. Notas de Aula.

ROMÃO, Wesley; NIEDERAUER, Carlos; MARTINS, Alejandro, TCHOLAKIAN, Aran; PACHECO, Roberto e BARCIA, Ricardo. EXTRAÇÃO DE REGRAS DE ASSOCIAÇÃO EM C&T: O ALGORITMO APRIORI. Florianópolis, SC, Brasil. Disponível em: http://www.abepro.org.br/biblioteca/ENEGEP1999_A0901.PDF. Acessadoo em: 15 de setembro de 2021.

NOGARE, Diego. Algoritmo Apriori para sistemas de recomendação. DIEGO NOGARE INTELIGÊNCIA ARTIFICIAL & MACHINE LEARNING, 2020. Disponível em: https://diegonogare.net/2020/05/algoritmo-apriori-para-sistemas-de-recomendacao/. Acessado em: 13 de setembro de 2021.