Ficha 6 - DC

João Nunes (A82300)
Luís Braga (A82088)

20/03/2020

Conteúdo

1 Parte I

1.1 O que são regras de associação? Para que servem?

As regras de associação, um pouco como os coeficientes de correlação, servem para tentar encontrar ligações entre os atributos de um *dataset*. Ora, esta metedologia de *Data Mining* é bastante interessante e importante para situações como, por exemplo, farmácias onde é feita a análise do cesto de compra, descobrindo dessa maneira relações entre os dados permitindo descrobrir o que é comprado *"junto"*, o que permite que depois se façam ajustes no modo de operação da farmácia consoante as descobertas.

1.2 Quais são as duas principais métricas calculadas nas regras de associação e como são calculadas?

As duas principais métricas das regras de associação são a percentagem de confiança e a percentagem de suporte. A percentagem de confiança é relativa ao quão verdadeiro é que quando um atributo é sinalizado como verdadeiro o outro atriuto associado também assim o seja. Em título de exemplo:

1) Anti-inflamatório => Antibiótico
2) 4 sacos de antibióticos, 3 sacos de Anti-inflamatórios (total = 7)
3) => antibióticos e anti-inflamatórios só coincidiram em 2 sacos
4) => Podiam ter conincidido em 3 sacos mas apenas coincidiram em 2 (2/3) = 0.67

Portanto, seguindo o exemplo anterior a confiança é de 0.67 aproximadamente. O mesmo processo de cálculo da confiança poderá ser aplicado para o processo contrário (*Antibiótico => Anti-inflamatório*). O suporte é calculado dividindo o número de instâncias que satisfazem uma condição pelo número de instâncias totais.

1.3 De que tipo de dados devem ser os atributos de um dataset para usar os operadores Frequent Pattern no RapidMiner?

Os atributos de um dataset devem ser do tipo nominais para usar os operadores Frequent Pattern.

1.4 Como é que os resultados das regras de associação são interpretados? No exemplo dos slides desta aula, qual foi a regra mais forte e como se sabe?

Tendo em conta o exemplo apresentado os *slides* disponibilizados é necessário conjugar a confiança com o suporte. Desta maneira, sem dúvida que a melhor regra produzida é *Hobbies => Religious* com quase 80% de confiança e 24% de suporte. O que indica que sendo da organização hobby poderá também pertencer a organização religiosa. Desta maneira, é possível estabelecer uma ligação entre estas duas organizações indicando que existem membros em comum em ambas as organizações.

2 Parte II

2.1 Faça download do dataset order.csv, importe-o para o RapidMiner e arraste-o para a janela de processo.Proceda à etapa de Data Understanding.

Começando pela primeira tarefa de visualização dos dados, primeiramente foi verificado se existe *"missing data"* nos dados pelo que se verificou que não existe dados em falta no *dataset*.

Figura 1: Missing data no dataset.

Se existisse dados em falta no *dataset* tal seria possível verificar através do filto descendente do *Missing*. De seguida foram visualizados certos atributos de modo a melhor entender o *dataset* disponibilizado. Por exemplo, o atributo *order_hour_of_day* indica a a hora do dia em que foi feita a compra, onde é possível verificar que a hora de ponta é as 15 horas, sendo que o período em que há menos compras é entre as [2.5,5]. Para além disso é possível também verificar que a média da hora de compra é as 13 horas e 36 minutos, sendo o desvio padrão de 4 horas e aproximadamente 16 minutos. O que é possível verificar então que existe alguns *outliers* neste atributo, ou seja, valores fora da gama de $13.6 \mathrel{+}= 4.263$.

Figura 2: Countplot do atributo orders_hour_of_day.

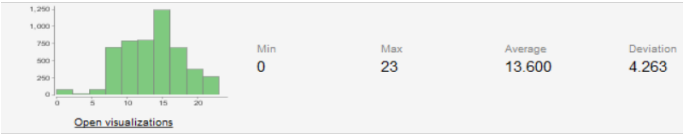

Figura 3: Dados estatísticos do atributo orders_hour_of_day.

Outro atributo essencial para o *Business Understanding* é o *days_since_prior_order* uma vez que este indica o número de dias que passaram entre a última compra no estabelecimento. O que poderá indicar a fidelização ou a falta de fidelização dos clientes por parte do estabelecimento. Através da análise do *count plot* é possível verificar que é bastante comum os clientes passarem 27 a 30 dias sem fazer mais nenhuma compra com cerca de 1741 instâncias registadas nesse intervalo. Em média os clientes passam 17 dias sem efetuar mais nenhuma compra no estabelecimento e com uma diferença padrão de cerca de 11 dias.

Figura 4: Countplot do atributo days_since_prior_order.

Figura 5: Dados estatísticos do atributo days_since_prior_order.

A maior parte dos restantes atributos são relativos aos produtos disponibilizados pelo estabelecimento e que são comprados pelos clientes.

Para além disso, e passando o foco para o *dataset* em geral, é possível verificar que existem 138 atributos distintos e 0 atributos especiais.

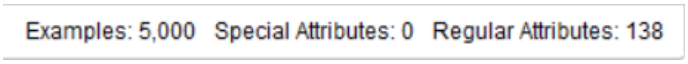

Figura 6: Dados do dataset.

2.2 Conforme necessário, execute os passos referentes à etapa de Data Preparation no seu dataset. Certifique-se de que todas as suas variáveis tenham dados consistentes e que os seus tipos de dados sejam apropriados para o operador FP-Growth.

Tal como foi possível verificar na pergunta anterior, não existem dados em falta no *dataset* e para além disso todos os dados são numéricos. Contudo, é possível proceder à seleção dos atributos mais importantes para aplicar os métodos das regras de associação. Como se pretente associar os produtos, é apenas necessário selecionar os atributos referentes aos produtos.

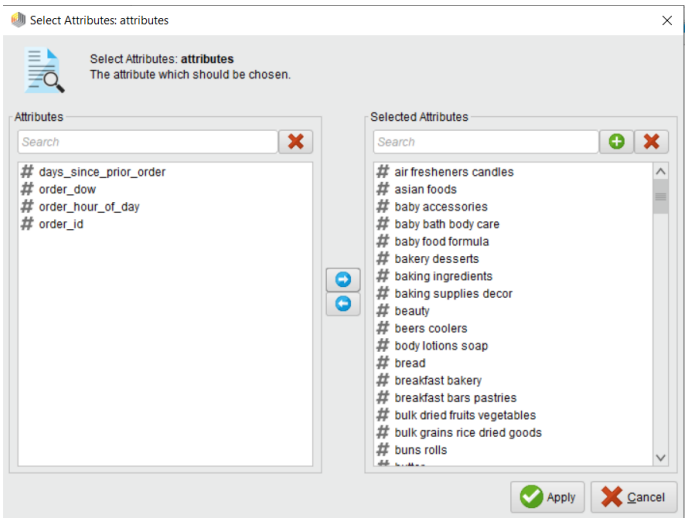

Figura 7: Atributos selecionados do dataset original.

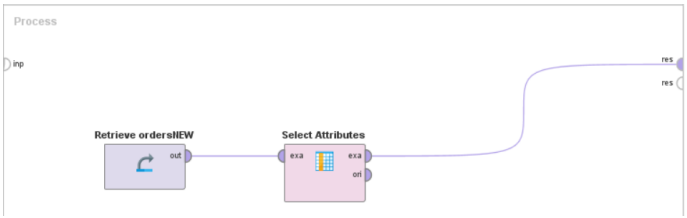

Figura 8: Processo aplicado.

Após a aplicação anterior do processo de seleção de atributos foi possível verificar que no novo dataset já não se encontram os atributos apresentados na figura 7, sendo que o número total de atributos desceu 134 atributos.

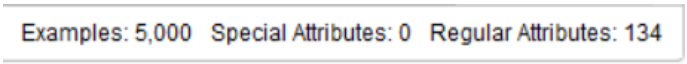

Figura 9: Dados do novo dataset.

2.3 Gere regras de associação para o dataset. Modifique os valores de confiança (min confidence) e suporte (min support) para identificar os níveis ideais, de modo a obter regras interessantes com valores de confiança e suporte razoáveis. Analise as outras medidas de força das regras, como LaPlace ou Conviction. Documente as suas descobertas. Que regras encontrou? Que atributos estão mais fortemente associados? Existem produtos frequentemente conectados que o surpreendam? Quantas vezes tentou diferentes valores de suporte e confiança antes de encontrar algumas regras de associação? Alguma das suas regras de associação é boa o suficiente ao ponto de se basear nela para tomar decisões? Porquê?

Para gerar as regras de associação será utilizado o operador *FP-Growth*.

Ao usar o *FP-Growth* foi possível verificar uma falha, uma vez que para usar este operador é necessário que os dados sejam nominais, pelo que foi necessário alterar o *pipeline* de modo a transformar os dados todos em dados nominais, seleccionar apenas os atributos correspondentes ao produto e depois aplicar o operador.

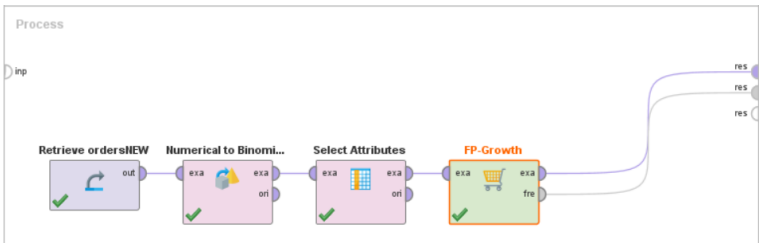

Figura 10: Pipeline de operações de modo a aplicar o FP-Growth.

De seguida, e após executar o processo anterior foi possível obter a seguinte tabela de resultados, de onde é possível encontrar uma relação entre o produto *fresh fruits* e *fresh vegetables*, *fresh fruits* e *packaged vegetable fruits*, *fresh vegetables* e *packaged vegetable fruits* e por fim *fresh fruits* com *fresh vegetables* e *packaged vegetable fruits*. Indicando alguma relação entre estes três atributos mencionados sendo que existe maior suporte para *fresh fruit* com *fresh vegetables* (*fresh fruit => fresh vegetables*). Tal como foi testado e com os parâmetros standard não é possível indicar com todas as certezas que a melhor regra de associação é a regra anteriormente identificada, contudo é lógico que comprando fruta fresca existe maior probabilidade de comprar também vegetais frescos.

Size	Support	Item 1	Item 2	Item 3
1	0.546	fresh fruits		
1	0.458	fresh vegetables		
1	0.385	packaged vegetables fruits		
1	0.257	yogurt		
1	0.230	packaged cheese		
1	0.227	milk		
1	0.208	water seltzer sparkling w...		
2	0.328	fresh fruits	fresh vegetables	
2	0.284	fresh fruits	packaged vegetables fruits	
2	0.253	fresh vegetables	packaged vegetables fruits	
3	0.205	fresh fruits	fresh vegetables	packaged vegetables fruits

Figura 11: Resultado de execução do FP-Growth.

De seguida, foi criada uma regra de associação, sendo que esta foi adicionada de seguida ao processo anterior, tendo sido utilizado um *min_confidence* de 0.8 com *theta* 2 e *laplace* 1.

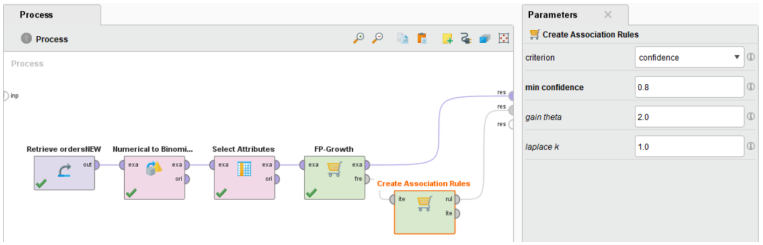

Figura 12: Parâmetros e processo com regra de associação.

Com os parâmetros anteriores foi apenas produzida uma regra com as seguintes duas premissas e uma conclusão.

Fresh Vegetables,Packaged Vegetable Fruits => Fresh Fruits

A regra anterior possui uma confiança de 0.809 com um suporte de 0.205 o que indica que a associação anterior é fidedigna e de confiança, podendo-se utilizar a regra anterior de modo a tomar alguma decisão ao nível do negócio (como por exemplo colocar os três produtos no mesmo corredor do supermercado).

	Conclusion	Support	Confidence	LaPlace
d vegetables fruits	fresh fruits	0.205	0.809	0.961

Figura 13: Resultados da melhor regra de associação.

Diminuindo a confiança é possível verificar as outras regras de associação com menor confiança, por exemplo, diminuindo a confiança mínima para 0.5 é possível obter a seguinte tabela.

Premises	Conclusion	Support	Confidence
fresh vegetables	packaged vegetables fruits	0.253	0.552
fresh fruits	fresh vegetables	0.328	0.602
fresh fruits, fresh vegetables	packaged vegetables fruits	0.205	0.623
packaged vegetables fruits	fresh vegetables	0.253	0.658
fresh vegetables	fresh fruits	0.328	0.716
fresh fruits, packaged vegetables fruits	fresh vegetables	0.205	0.719
packaged vegetables fruits	fresh fruits	0.284	0.739
fresh vegetables, packaged vegetables fruits	fresh fruits	0.205	0.809

Figura 14: Tabela com as várias regras de associação com confiança superior a 50%.

Analisando a tabela anterior, existem também outras associações satisfatórias como por exemplo *fresh vegetables => fresh fruits* com suporte 0.328 e confiança 0.716, ou *packaged vegetable fruits => fresh fruits* com confiança 0.739 e suporte 0.284. Podendo estas regras de associação ser também usadas de modo a tomar decisões. Para além disso também é possível gerar um gráfico que mostra as relações entre as diferentes regras de associação e os atributos envolvidos nestas regras de associação.

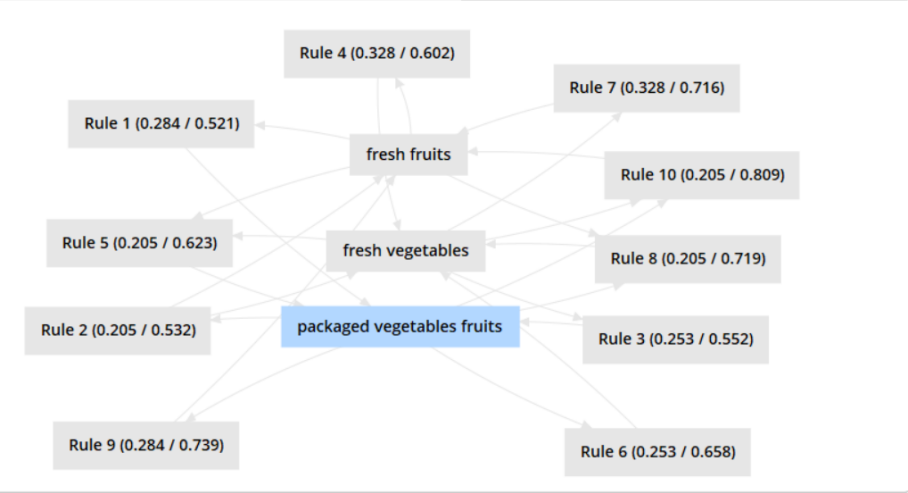

Figura 15: Grafo com as diferentes regras de associação e atributos.

2.4 Crie um novo modelo de regras de associação usando o mesmo dataset, mas desta vez, use o operador WFPGrowth no RapidMiner. Para poder utilizar este operador, instale primeiro a extensão "Weka Extension" em Extensions -> Marketplace (procure Weka). (Dicas para usar o operador W-FPGrowth: (1) Este operador cria as suas próprias regra sem a ajuda de outros operadores; e (2) Os parâmetros de suporte e confiança deste operador são identificados como U e C, respetivamente. Apresente e discuta os resultados obtidos.

No *W-FPGrowth* foi primeiro necessário diminuir o fator de confiança do operador (que estava em 0.9) para 0.5, tendo sido executado o processo com a seguinte configuração do *W-FPGrowth*.

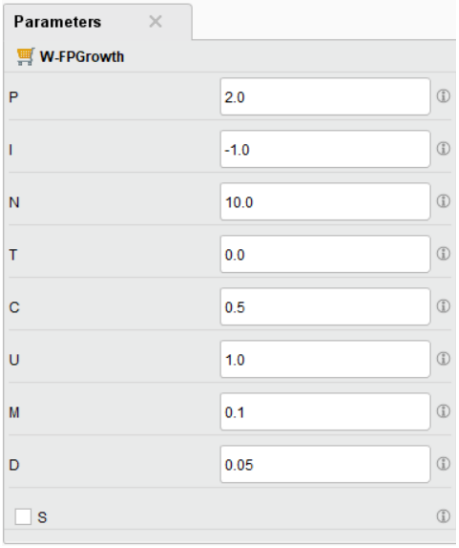

Figura 16: Parâmetros do operador W-FPGrowth.

Para além disso foi também definido o seguinte modelo.

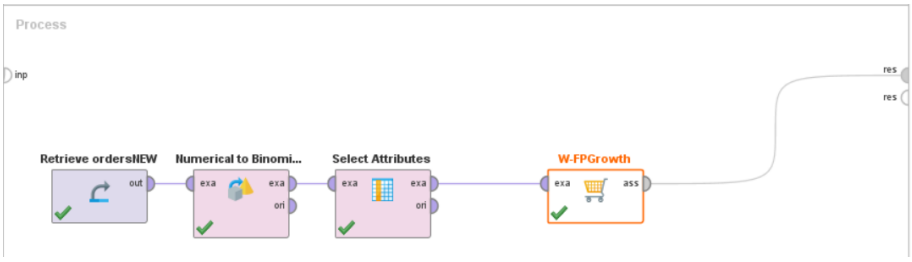

Figura 17: Modelo para o operador W-FPGrowth.

Após executar o modelo, foi possível obter um *log* dos resultados do operador onde se apresenta por ordem decrescente de maior confiança para menor confiança as diferentes regras de associação, de onde é possível retirar que segundo este operador a regra de associação com maior confiança, sendo ela 0.66, dentro de todos os produtos é:

packaged vegetable fruits => fresh vegetables

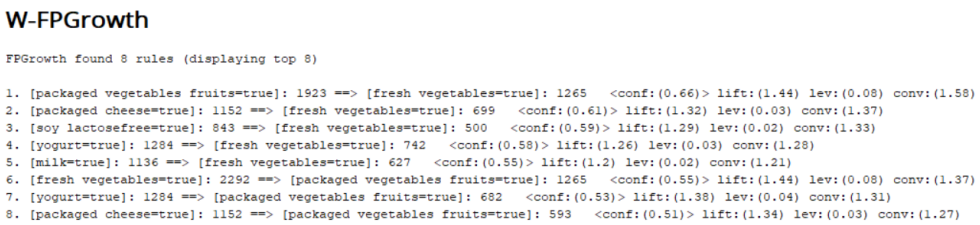

Figura 18: Resultados do operador W-FPGrowth.

2.5 O algoritmo Apriori é frequentemente usado no processo de Data Mining para associações. Pesquise Apriori (W-Apriori) nos operadores do RapidMiner e adicione-o ao seu dataset num novo processo. Use o separador de Ajuda no canto inferior direito do RapidMiner para aprender sobre os parâmetros e funções desse operador. Apresente e discuta os resultados obtidos.

No algoritmo *Apriori* e tal como no algoritmo anterior dos parâmetros mais importantes de testar e mudar são o *C* (confiança mínima da regra de associação) e *U* (limite superior do suporte) e o *M* (limite inferior do suporte). Como tal, foi necessário diminuir a confiança mínima da regra de associação (C) que por defeito é de 0.9, esta foi diminuída para 0.5.

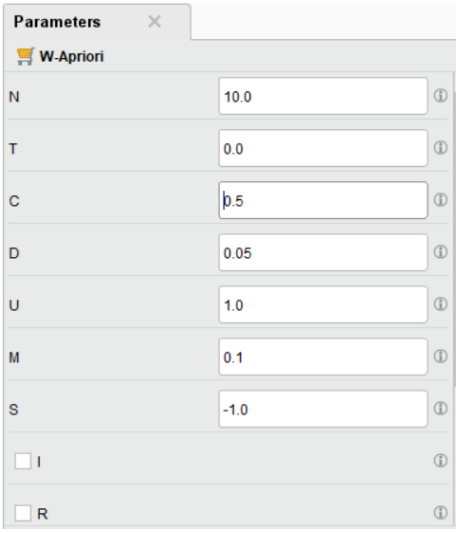

Figura 19: Parâmetros do operador W-Apriori.

Como tal, foi também montado o seguinte modelo para este operador.

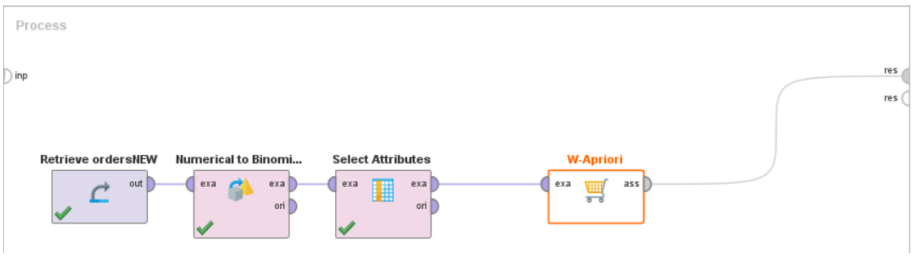

Figura 20: Modelo para o operador W-Apriori.

A execução do algoritmo *W-Apriori* permitiu identificar, tal como no caso anterior, que a regra de associação com maior confiança é:

packaged vegetable fruits => fresh vegetables

Com exatamente o mesmo fator de confiança de 0.66. As restantes regras de associação são também exatamente iguais às do modelo anterior (*W-FPGrowth*) e possuem também a mesma confiança, pelo que é possível inferir que é um bocado redundante utilizar ambos os algoritmos, uma vez que os resultados são iguais, sendo portanto mais aconselhável utilizar um ou outro.

2.6 Apresente uma conclusão global dos exercícios realizados, comparando os resultados obtidos através de cada uma das técnicas utilizadas.

De modo a sumarizar todo o trabalho efectuada, foi feita a seguinte tabela onde se apresenta para cada um dos algoritmos utilizados os valores de confiança e suporte para cada uma das melhores regras de associação.

	FP-Growth	W-FPGrowth	W-Apriori
Confiança	0.809	0.66	0.66
Suporte	0.205	-	-

Sendo que para o algoritmo *FP-Growth* a melhor regra de associação corresponde a:

Fresh Vegetables, Packaged Vegetable Fruits => Fresh Fruits

Para os dois algoritmos seguintes *W-FPGrowth* e *W-Apriori* estes dois algoritmos identificaram a mesma regra de associação como sendo a melhor:

packaged vegetable fruits => fresh vegetables

Ou seja, é portanto possível inferir que o algoritmo *FP-Growth* apresentou melhores resultados, uma vez que apresentou em geral sempre regras de associação com maior confiança, sendo que se fosse necessário escolher um algoritmo para basear decisões seria melhor escolher o *FP-Growth*. Para além disso, nos três algoritmos a melhor regra de associação (maior confiança) é a regra identificada pelo algoritmo *FP-Growth* pelo que é possível indicar com elevada probabilidade se os clientes comprarem vegetais frescos e frutas embaladas então irão também comprar frutas frescas.