

Ficha 7 - DC

João Nunes (A82300)
Luís Braga (A82088)

27/03/2020

Conteúdo

1 Parte I	2
1.1 O que significa o 'k' em k-Means clustering?	2
1.2 Como se identificam os clusters? Qual é o processo que o Rapid Miner usa para definir e colocar as observações num determinado cluster?	2
1.3 O que revela a Centroid Table ao utilizador? Como se interpretam os valores dessa tabela?	2
1.4 Como é que a presença de outliers nos atributos de um dataset influencia a utilidade de um modelo de k-Means clustering? O que poderia ser feito para resolver este problema?	2
2 Parte II	2
2.1 Pense num problema que possa ser resolvido agrupando observações em clusters. Procure na internet um dataset que possa ser utilizado e aplicado a um modelo de k-Means.	2
2.1.1 a) Importe os dados para o RapidMiner. Não se esqueça de garantir que estes estejam no formato CSV. Execute a etapa de Data Understanding.	2
2.1.2 b) Efectue a etapa de Data Preparation. Pode incluir componentes de inconsistência de dados, missing values, ou alteração do tipo de dados.	3
2.1.3 c) Adicione um operador de k-means clustering ao dataset no RapidMiner e altere os parâmetros de acordo com a necessidade (sobretudo ovalue, para adequar ao problema em questão).	4
2.1.4 d) Estude a Centroid Table, Folder View, e outras ferramentas de avaliação.	5
2.1.5 e) Reporte todos os passos anteriores e as evidências encontradas, bem como de que forma o que foi encontrado permite responder ao problema inicial.	6
2.2 Experimente o mesmo dataset com diferentes operadores de k-Means como o Kernel ou Fast. Em que medida diferem do modelo original. Estes operadores mudam os clusters originais? Se sim, em que medida?	6

1 Parte I

1.1 O que significa o 'k' em k-Means clustering?

O k em k-Means clustering significa o número de grupos ou *clusters* cujos valores serão agrupados.

1.2 Como se identificam os clusters? Qual é o processo que o Rapid Miner usa para definir e colocar as observações num determinado cluster?

O algoritmo de *K-Means clustering* amostra um conjunto de observações do *dataset*, de seguida calcula as médias de cada atributo para as observações das amostras e compara os outros atributos do dataset com a média dessa amostra. Os valores dos atributos mais semelhantes às médias de cada um dos clusters juntam-se ao próprio cluster. O processo aplicado no *Rapid Miner* é semelhante na medida em que inicialmente o processo começa com k pontos centroid que são sucessivamente recalculados através da média de todos os valores do cluster até que seja satisfeita a condição de paragem (*max optimization steps*).

1.3 O que revela a Centroid Table ao utilizador? Como se interpretam os valores dessa tabela?

Na *Centroid Table* estão disponibilizados todos os clusters definidos pelo utilizador possuindo as médias para cada atributo para cada um dos clusters definidos. Desta maneira, e no contexto do exemplo apresentado nos *slides* é possível chegar à conclusão que o *cluster 2* é o que possui o grupo de risco de doenças cardio-vasculares através da análise dos valores do atributo desse cluster.

1.4 Como é que a presença de outliers nos atributos de um dataset influencia a utilidade de um modelo de k-Means clustering? O que poderia ser feito para resolver este problema?

Poderia ser aplicado um processo de *Interquartil* de modo a remover os *outliers* uma vez que estes valores não são representativos dos valores reais do atributo (a sua grande maioria) e podem variar depois o *clustering* que é feito no processo de *K-Means clustering* dando resultados erróneos.

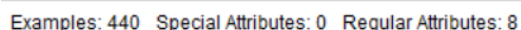
2 Parte II

2.1 Pense num problema que possa ser resolvido agrupando observações em clusters. Procure na internet um dataset que possa ser utilizado e aplicado a um modelo de k-Means.

Foi utilizado o seguinte dataset para aplicar o modelo de k-means <https://archive.ics.uci.edu/ml/datasets/Wholesale+customers>.

2.1.1 a) Importe os dados para o RapidMiner. Não se esqueça de garantir que estes estejam no formato CSV. Execute a etapa de Data Understanding.

Na etapa de visualização dos dados foi possível verificar que existem 440 instâncias, 8 atributos distintos e que não possui dados em falta.



Examples: 440 Special Attributes: 0 Regular Attributes: 8

Figura 1: Dados gerais do dataset.

Para além disso e verificando os atributos individualmente, como por exemplo o atributo *Fresh* onde se apresenta o gasto anual em produtos frescos, é possível verificar que cerca de 3 a 11217.8 unidades monetárias em compras de produto frescos, sendo que este atributo possui a seguinte distribuição de valores.

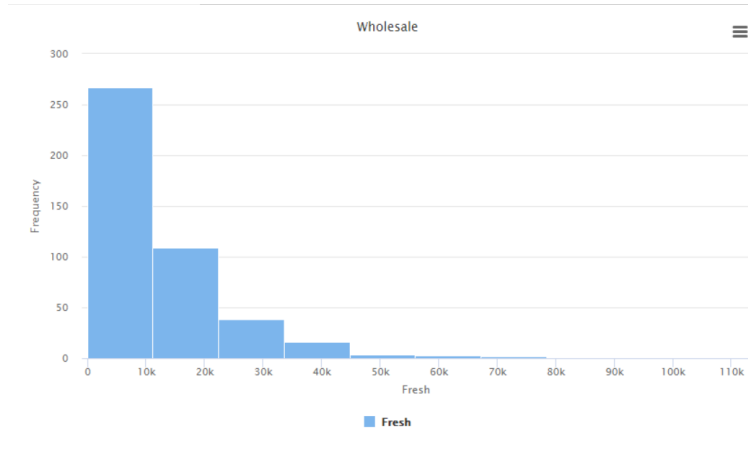


Figura 2: Count plot do atributo Fresh.

O atributo *Fresh* possui também os seguintes dados estatísticos, onde é possível verificar que em média os clientes gastam 12000 unidades monetárias em produtos frescos, sendo o mínimo 3 u.m e o máximo de 112151 u.m. Como tal, é possível também verificar que para além deste atributo existem também outros atributos com *outliers*, sendo evidente que na próxima fase é preciso tratar destes valores *outliers*.

2.1.2 b) Efectue a etapa de Data Preparation. Pode incluir componentes de inconsistência de dados, missing values, ou alteração do tipo de dados.

Para a etapa de *Data Preparation* e como foi verificado anteriormente, foi necessário remover os *outliers* existentes nos atributos. Como tal foi montado o seguinte modelo, onde existe um componente que deteta os *outliers* através do cálculo de uma função de distância euclidiana e de seguida é aplicado um *filter* onde na coluna *outlier* todos os valores a *true* são apagados.

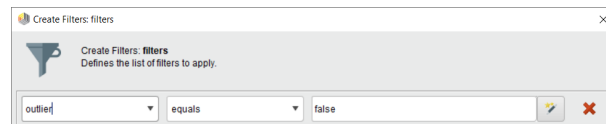


Figura 3: Filter aplicado.

Row No.	outlier	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_...	Delicassen
1	false	2	3	12669	9656	7561	214	2674	1338
2	false	2	3	7057	9810	9568	1762	3293	1776
3	false	1	3	13295	1196	4221	6404	507	1788
4	false	2	3	22615	5410	7198	3915	1777	5185
5	false	2	3	9413	8259	5126	666	1795	1451
6	false	2	3	12126	3199	6975	480	3140	545
7	false	2	3	7579	4956	9426	1669	3321	2566
8	false	1	3	5963	3648	6192	425	1716	750
9	false	2	3	6006	11093	18881	1159	7425	2098
10	false	2	3	3366	5403	12974	4400	5977	1744
11	false	2	3	13146	1124	4523	1420	549	497
12	false	1	3	10253	1114	3821	397	964	412

ExampleSet (340 examples, 1 special attribute, 8 regular attributes)

Figura 4: Dataset depois de aplicar o filtro.

De seguida é aplicado um componente de *feature selection* onde são mantidas todos os atributos originais do *data set* excepto a nova coluna gerada anteriormente a *outlier*.

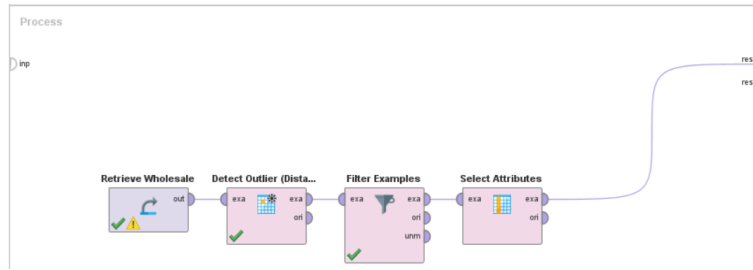


Figura 5: Modelo final aplicado.

2.1.3 c) Adicione um operador de k-means clustering ao dataset no RapidMiner e altere os parâmetros de acordo com a necessidade (sobretudo ovalork, para adequar ao problema em questão).

Adicionado o operador de k-means, primeiramente foi alterado o *k* do seu valor *standard* para 3, de forma a acomodar os três valores distintos que o atributo da região poderá ter. Os restantes parâmetros permaneceram *standard*.

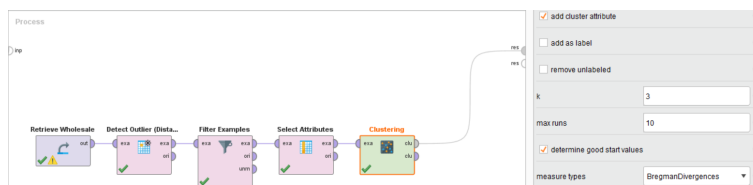


Figura 6: Modelo utilizado para aplicar k-means.

Após correr o modelo foi possível verificar a seguinte distribuição de instâncias pelos *clusters* onde foi possível verificar que o cluster 2 possui um número maior de instâncias que os outros clusters.

Cluster Model

```
Cluster 0: 95 items
Cluster 1: 74 items
Cluster 2: 171 items
Total number of items: 340
```

Figura 7: Distribuição de instâncias com $k=3$.

Contudo tanto o cluster 1 como o cluster 0 possuem um número semelhante de instâncias em cada um deles. Variando o *K* para 4 em vez de três, foi possível obter uma distribuição mais justa de instâncias por cada um dos clusters.

Cluster Model

```
Cluster 0: 110 items
Cluster 1: 76 items
Cluster 2: 62 items
Cluster 3: 92 items
Total number of items: 340
```

Figura 8: Distribuição de instâncias com $k=4$.

Uma vez que com $k=4$ já foi gerada uma distribuição justa de instâncias pelos clusters, ou seja, os clusters são bastante equilibrados dá-se por terminada esta fase de experimentação.

2.1.4 d) Estude a Centroid Table, Folder View, e outras ferramentas de avaliação.

Para cada um dos k anteriores também é possível verificar outras ferramentas de avaliação como a *Centroid table* e a *Folder view*. Para o $k=3$ é possível verificar a seguinte tabela.

Attribute	cluster_0	cluster_1	cluster_2
Channel	1.200	1.865	1.064
Region	2.568	2.635	2.485
Fresh	18396.537	3970.676	5550.953
Milk	2946.926	8536.068	2585.550
Grocery	4130.263	13153.432	3133.579
Frozen	2824.463	1220.041	2307.275
Detergents_Paper	934.947	5785.297	845.819
Delicassen	1351.495	1266.405	838.772

Figura 9: Centroid table $k=3$.

Para esta *Centroid table* é possível verificar as médias de cada um dos atributos do cluster. Ou seja, no cluster 1 o *channel* é o mais próximo do 2, sendo que os clientes que mais gastam em produtos *fresh* encontram-se agrupados no cluster 0, gastando em média 18397 u.m. No cluster 1 por sua vez em gasta-se em média 13153 u.m em *Grocery*.

Por sua vez com o $k=4$ foi gerada a seguinte tabela.

Attribute	cluster_0	cluster_1	cluster_2	cluster_3
Channel	1.027	1.842	1.226	1.141
Region	2.500	2.645	2.581	2.478
Fresh	3367.282	4007.789	20717.177	11209.304
Milk	2570.891	8494.842	3052.468	2566.272
Grocery	3010.700	13005.579	4377.419	3375.761
Frozen	1893.682	1249.829	3217.435	2721.500
Detergents_Paper	797.973	5705.829	1028.935	829.924
Delicassen	852.436	1274.461	1439.758	930.913

Figura 10: Centroid table $k=4$.

Passando para a *Centroid table* em todos os clusters verifica-se que a área continua semelhante para todos os clusters, contudo o cluster 1 aproxima-se mais do *Channel 2* que qualquer outro cluster. Para além disso, o cluster 2 em média possui um maior número de u.m gastas em produtos frescos. No cluster 1 os clientes em média gastam mais dinheiro em leite, e no cluster 0 e 3 os clientes agrupados nestes clusters não possuem em média valores que permitem distinguir dos outros.

Outra visualização dos resultados é o *folder view* contudo com este *dataset* não foi possível efetuar a visualização individual das instâncias em cada um dos clusters.

2.1.5 e) Reporte todos os passos anteriores e as evidências encontradas, bem como de que forma o que foi encontrado permite responder ao problema inicial.

Portanto, foi possível inferir que o valor de k mais correto é com $k=4$ para aplicar o algoritmo de *k-means*, uma vez que com $k=4$ resulta numa distribuição mais justa pelos clusters. Para além disso para alguns atributos é possível observar uma distribuição distinta de médias pelos cluster, como por exemplo, no cluster 1 em média os clientes gastam mais dinheiro em *grocery* e a partir daí é possível aplicar medidas em concreto para tratar deste segmento de clientes.

2.2 Experimente o mesmo dataset com diferentes operadores de k-Means como o Kernel ou Fast. Em que medida diferem do modelo original. Estes operadores mudam os clusters originais? Se sim, em que medida?

Mudando o operador de *k-means* para o *k-means (fast)* foi possível verificar após, correr o modelo, a seguinte distribuição de instâncias com $k=3$.

Cluster Model

```
Cluster 0: 169 items
Cluster 1: 95 items
Cluster 2: 76 items
Total number of items: 340
```

Figura 11: K-Means Fast distribuição de instâncias com $k=3$.

Ou seja, o cluster com maior distribuição de instâncias (quando comparado com o *K-Means* anterior) passou a ser o zero com 169 instâncias, o cluster 1 passou a ter 95 instâncias e o 2 passou a ter 76 instâncias. Na *centroid table* os valores médios dos atributos são semelhantes aos valores apresentados na tabela anterior com $k=3$ apenas distribuídos por clusters diferentes.

Attribute	cluster_0	cluster_1	cluster_2
Channel	1.065	1.200	1.842
Region	2.491	2.568	2.618
Fresh	5569.651	18396.537	3970.684
Milk	2526.320	2946.926	8511.184
Grocery	3088.467	4130.263	12990.066
Frozen	2299.645	2824.463	1265.618
Detergents_Paper	817.911	934.947	5717.368
Delicassen	832.385	1351.495	1269.355

Figura 12: K-Means Fast centroid table $k=3$.

No que toca à distribuição pelos clusters com $k=4$ utilizando o *K-Means Fast* pode-se verificar que o cluster 0 e 3 possuem exatamente o mesmo número de instâncias distribuídas quando comparado com o original. O cluster 1 e 2 quando comparado com a distribuição original possuem um número de instâncias trocado.

Cluster Model

```
Cluster 0: 110 items
Cluster 1: 62 items
Cluster 2: 76 items
Cluster 3: 92 items
Total number of items: 340
```

Figura 13: K-Means Fast distribuição de instâncias com $k=4$.

No que toca à *centroid table* neste algoritmo os valores médios de cada um dos clusters, no cluster 0 e 3 são iguais ao original (*k-means* com $k=4$), sendo que os valores médios do cluster 1 e 2 os valores encontram-se trocados um com o outro quando comparado com o original.

Attribute	cluster_0	cluster_1	cluster_2	cluster_3
Channel	1.027	1.226	1.842	1.141
Region	2.500	2.581	2.645	2.478
Fresh	3367.282	20717.177	4007.789	11209.304
Milk	2570.891	3052.468	8494.842	2566.272
Grocery	3010.700	4377.419	13005.579	3375.761
Frozen	1893.682	3217.435	1249.829	2721.500
Detergents_Paper	797.973	1028.935	5705.829	829.924
Delicassen	852.436	1439.758	1274.461	930.913

Figura 14: K-Means Fast centroid table $k=4$.

Portanto, repetindo os passos efetuados no exercício anterior, ou seja, correr o modelo com $k=3$ e $k=4$ os resultados obtidos foram iguais (ou bastante semelhantes) aos obtidos com o uso do *k-means standard*, portanto, não existe comparativamente ao original, nenhuma vantagem em utilizar esta versão do *k-means*.