Análise de clusters do dataset Iris

Guilherme Barreto Boscaro

Um cluster

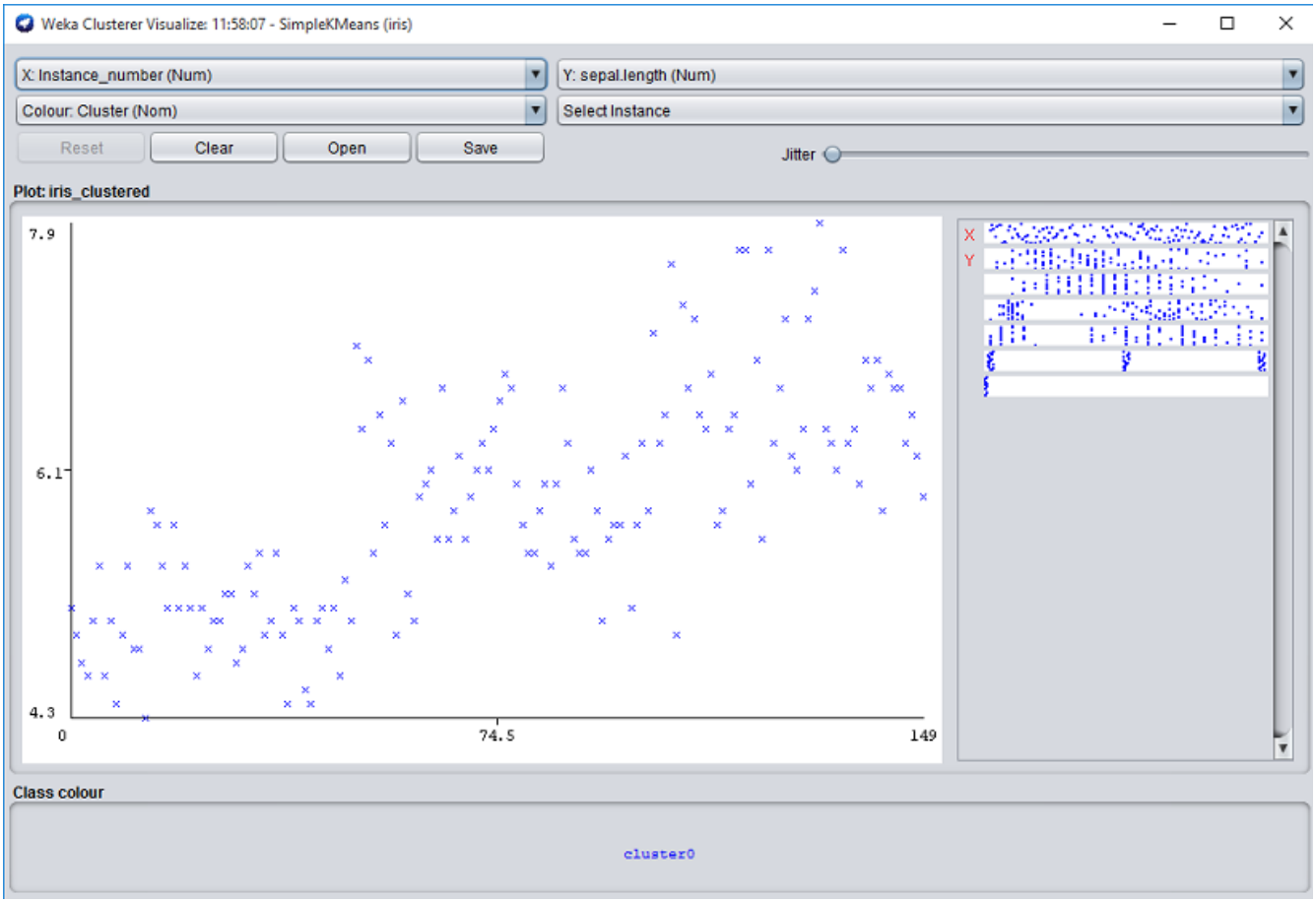
Utilizando o software Weka em conjunto com o algoritmo Kmeans para definir qual é o melhor número de clusters, primeiro defini a quantidade de clusters do algoritmo para 1, para ter como base o valor de todas as instâncias dentro de apenas um cluster.

O que o software me retornou o seguinte relatório:

```
=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 1
Within cluster sum of squared errors: 141.16611042137328

Initial starting points (random):

Cluster 0: 6.1,2.9,4.7,1.4,Versicolor

Missing values globally replaced with mean/mode

Final cluster centroids:
                           Cluster#
Attribute       Full Data          0
                  (150.0)    (150.0)
====================================
sepal.length       5.8433     5.8433
sepal.width        3.0573     3.0573
petal.length        3.758      3.758
petal.width        1.1993     1.1993
variety            Setosa     Setosa



Time taken to build model (full training data) : 0.03 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      150 (100%)
```

O relatório indica que há um erro RMS de 141,16611042137328 quando somente um cluster contendo todas as instâncias é gerado.

O Weka permite visualização gráfica dos clusters gerados, o que facilita o entendimento das informações, como na imagem abaixo, onde foi utilizado o cluster gerado anteriormente com as informações de instance_number no eixo horizontal X e sepal.length no eixo vertical Y.



Três clusters

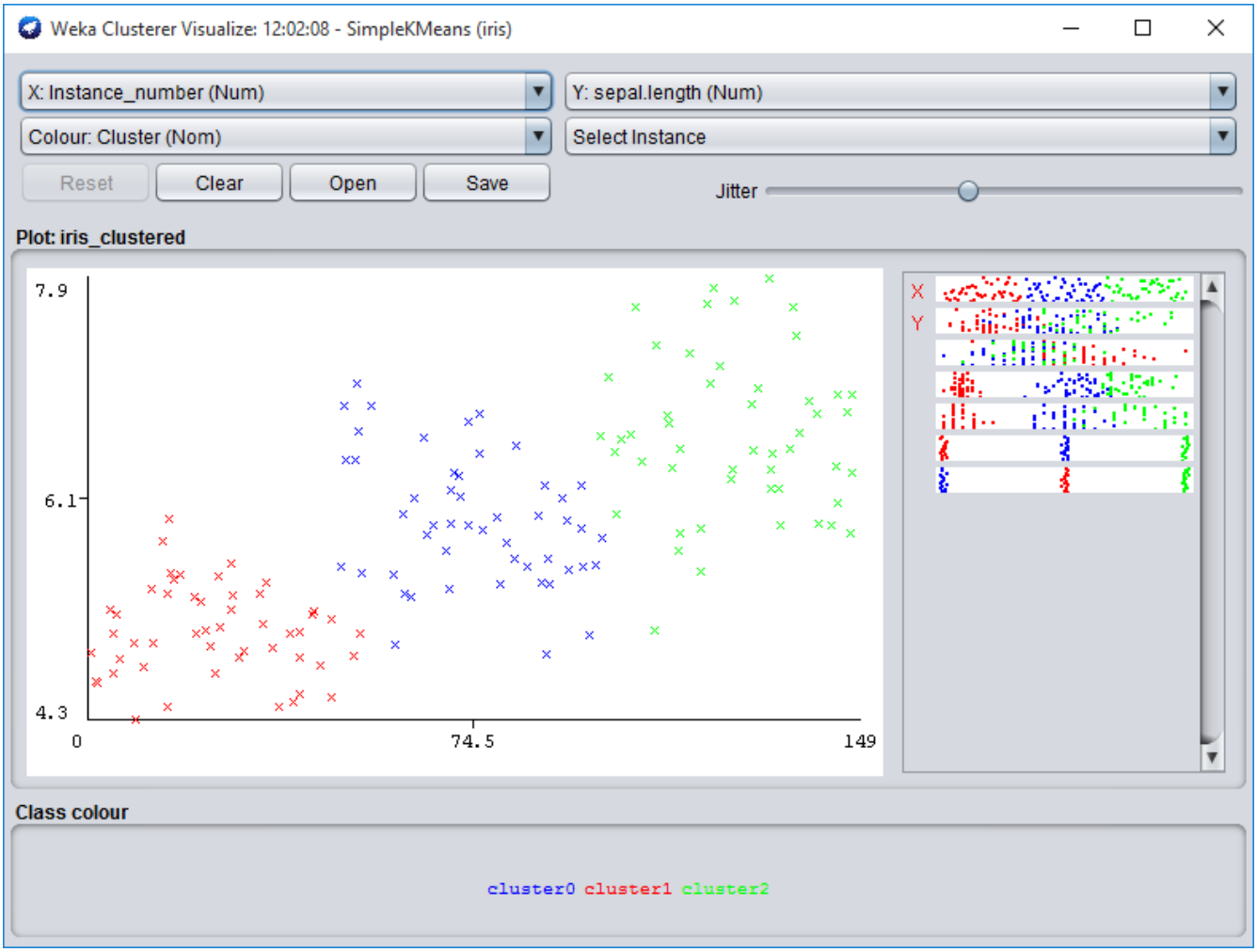
Ao gerar três clusters o relatório do Weka indica um erro RMS de 7,801559361268048, aproximadamente vinte vezes menor do que o o erro RMS de 141 de apenas um cluster.

```
kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 7.801559361268048

Initial starting points (random):

Cluster 0: 6.1,2.9,4.7,1.4,Versicolor
Cluster 1: 6.2,2.9,4.3,1.3,Versicolor
Cluster 2: 6.9,3.1,5.1,2.3,Virginica

Missing values globally replaced with mean/mode

Final cluster centroids:
                           Cluster#
Attribute       Full Data          0          1          2
                  (150.0)     (50.0)     (50.0)     (50.0)
============================================================
sepal.length       5.8433      5.936      5.006      6.588
sepal.width        3.0573       2.77      3.428      2.974
petal.length        3.758       4.26      1.462      5.552
petal.width        1.1993      1.326      0.246      2.026
variety            Setosa Versicolor     Setosa  Virginica



Time taken to build model (full training data) : 0.03 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       50 ( 33%)
1       50 ( 33%)
2       50 ( 33%)
```

Também é possível notar que as instâncias foram divididas em três grupos de mesma quantidade:

```
=== Model and evaluation on training set ===

Clustered Instances

0       50 ( 33%)
1       50 ( 33%)
2       50 ( 33%)
```

O que fica bem claro quando representado graficamente, utilizando as mesmas métricas que o gráfico anterior:



Como as informações estão separadas em grupos distintos que não se sobrepoem e possuem um baixo valor para o erro RMS, três clusters parecem ser a quantidade ideal para este modelo, no entanto, para me certificar, aumentarei a quantidade de clusters para 5, 10, 20 e 50.

Cinco clusters

Já com cinco clusters podemos notar que o erro RMS diminuiu para 6.277659330769319.

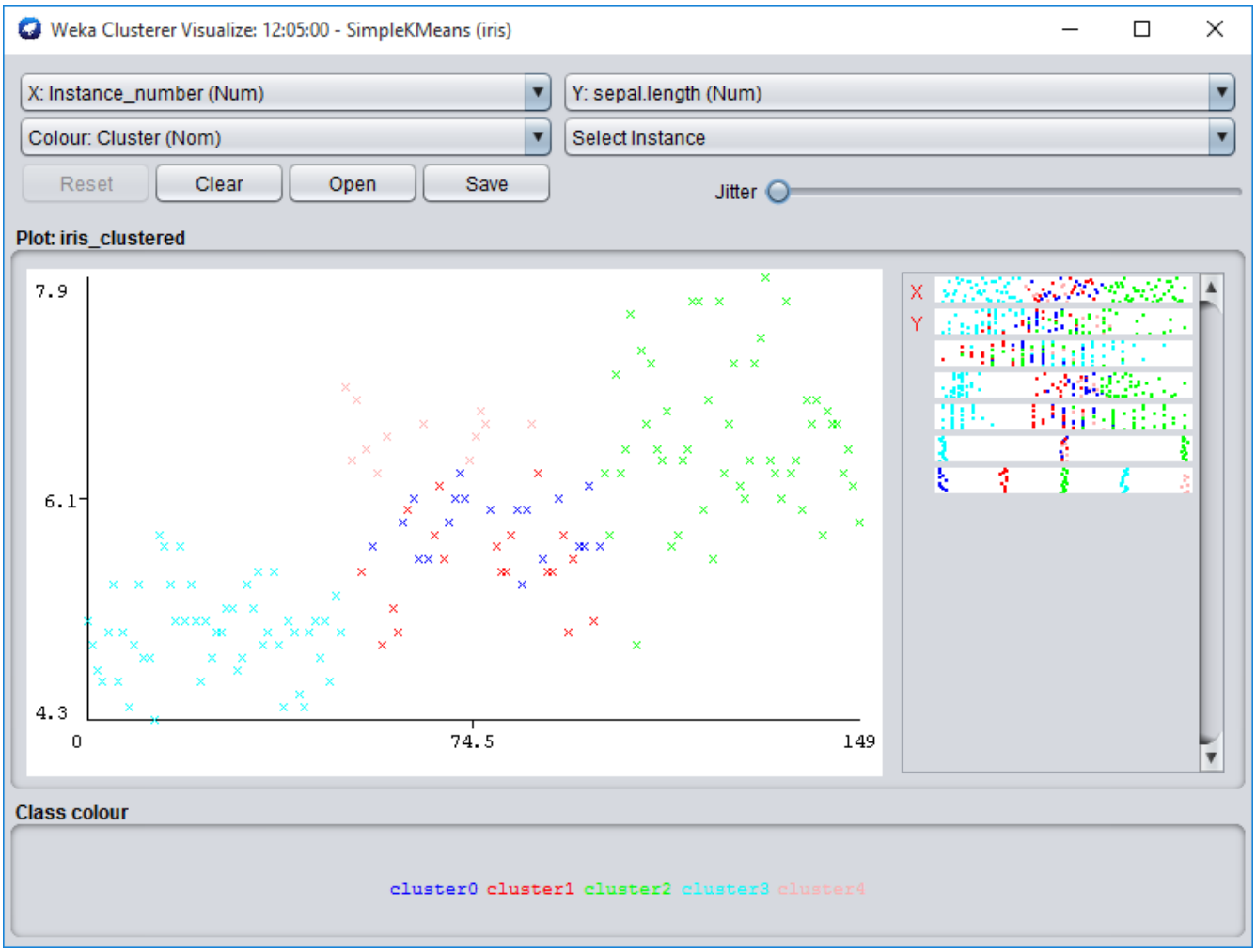
```
kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 6.277659330769319

Initial starting points (random):

Cluster 0: 6.1,2.9,4.7,1.4,Versicolor
Cluster 1: 6.2,2.9,4.3,1.3,Versicolor
Cluster 2: 6.9,3.1,5.1,2.3,Virginica
Cluster 3: 5.5,4.2,1.4,0.2,Setosa
Cluster 4: 6.9,3.1,4.9,1.5,Versicolor

Missing values globally replaced with mean/mode

Final cluster centroids:
                        Cluster#
Attribute       Full Data          0          1          2          3          4
                  (150.0)     (19.0)     (19.0)     (50.0)     (50.0)     (12.0)
================================================================================
sepal.length       5.8433     5.8789     5.5526      6.588      5.006     6.6333
sepal.width        3.0573     2.9211     2.4526      2.974      3.428     3.0333
petal.length        3.758     4.4211     3.8632      5.552      1.462     4.6333
petal.width        1.1993     1.4105     1.1579      2.026      0.246     1.4583
variety            Setosa Versicolor Versicolor  Virginica     Setosa Versicolor



Time taken to build model (full training data) : 0.02 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       19 ( 13%)
1       19 ( 13%)
2       50 ( 33%)
3       50 ( 33%)
4       12 (  8%)
```

No entanto, houve uma maior especialização das informações, o que fez com que o grupo que possui o atributo “variety” igual a “Versicolor” se dividisse em três, respectivamente clusters 0,1 e 4 em comparação a quando foram definidos três clusters:

Final cluster centroids:

		Cluster#				
Attribute	Full Data	0	1	2	3	4
	(150.0)	(19.0)	(19.0)	(50.0)	(50.0)	(12.0)
sepal.length	5.8433	5.8789	5.5526	6.588	5.006	6.6333
sepal.width	3.0573	2.9211	2.4526	2.974	3.428	3.0333
petal.length	3.758	4.4211	3.8632	5.552	1.462	4.6333
petal.width	1.1993	1.4105	1.1579	2.026	0.246	1.4583
variety	Setosa	Versicolor	Versicolor	Virginica	Setosa	Versicolor

Enquanto o cluster 2 “Virginica” e o cluster 3 “Setosa” permaneceram bem definidos, como é possível ver no gráfico:



Dez, vinte, cinquenta clusters e o “joelho” da curva

Com uma maior quantidade de clusters é gerada uma maior especialização das informações, diminuindo o erro RMS quanto mais próximo do número total de elementos no dataset.

Quantidade de Clusters	Erro RMS
1	141,17
3	7,80
5	6,28
10	4,59
20	1,59
50	0,68

Ao gerar um gráfico utilizando as informações acima podemos verificar que o “joelho” da curva se dá em 3 clusters com erro RMS de 7,80, sendo considerado o número ideal de grupos para este modelo.