

◇ Análise do Resultado do SimpleKMeans

1. Configuração do Algoritmo

- Foi utilizado **SimpleKMeans** com **N=10** , mas na execução foram gerados **5 clusters**.
- Distância usada: **Euclidiana**.
- Número máximo de iterações: **500** .

2. Características dos Dados

- Total de instâncias:** 4.773
- Atributos:** Idade, Atraso, Valor, Contato, Efetivo, Acordo.

3. Centroides dos Clusters

Os valores médios de cada atributo dentro de cada cluster apresentam padrões diferenciados nos dados.

4. Distribuição dos Clusters

- Cluster 0** → 663 instâncias (14%)
- Cluster 1** → 834 instâncias (17%)
- Cluster 2** → 269 instâncias (6%)
- Cluster 3** → 873 instâncias (18%)
- Cluster 4** → 2.134 instâncias (45%)

O cluster 4 é o maior, representando **quase metade dos dados** .

◇ Relação com o Apriori

Agora, para comparar com as **regras de associação geradas pelo Apriori** , podemos refletir:

1. Clusterização identifica grupos de perfis semelhantes

- Exemplo: Pode ter um cluster de clientes que pagam em dia, outro de inadimplentes, outro de novos clientes, etc.

2. A priori descobrimos padrões de comportamento dentro desses grupos

- Se aplicarmos ou a priori dentro de cada cluster, podemos identificar padrões mais específicos.
- Exemplo: "Clientes no cluster 4 têm maior probabilidade de fechar um acordo se tiverem um contato telefônico ativo".

3. Interpretação Complementar

- A clusterização mostra **quem são os grupos**.
- A priori revela **quais padrões existem dentro desses grupos** .

Clusterer

Choose SimpleKMeans -N 10 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10

Cluster mode

☒ Use training set☐ Supplied test set

Set...

☐ Percentage split

% 66

☐ Classes to clusters evaluation

(Num) Acordo

☒ Store clusters for visualization

Ignore attributes

Start

Stop

Result list (right-click for options)

15:48:24 - SimpleKMeans

15:56:29 - SimpleKMeans

15:57:13 - SimpleKMeans

Clusterer output

=== Run information ===

Scheme:weka.clusterers.SimpleKMeans -N 5 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10

Relation: A4

Instances: 4773

Attributes: 6

Idade

Atraso

Valor

CONTATO

EFETIVO

Acordo

Test mode:evaluate on training data

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

kMeans

=====

Number of iterations: 16

Within cluster sum of squared errors: 909.7776150855439

Missing values globally replaced with mean/mode

Cluster centroids:

Attribute	Full Data (4773)	Cluster#				
		0 (663)	1 (834)	2 (269)	3 (873)	4 (2134)
Idade	1.7461	4	1.8321	4	1.0332	1.0197
Atraso	2.4588	1.181	1.8885	6	6	1.1837
Valor	1.4718	1.9593	2.0731	0.4498	0.3963	1.6542
CONTATO	4.7591	4.3922	5.2842	4.0595	4.7205	4.7718
EFETIVO	1.1672	1.1825	1.1091	1.223	1.1775	1.1739
Acordo	0.8184	1	0	0.8773	1	1

Time taken to build model (full training data) : 0.19 seconds

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


Clustered Instances

0	663 (14%)
1	834 (17%)
2	269 (6%)
3	873 (18%)
4	2134 (45%)

Status

OK

Log

 x 0

1. Carregamento da base de dados

- Abrimos o Weka Explorer e carregamos o arquivo "iris.csv".
- Seleccionamos a aba "Cluster" e escolhemos o algoritmo **SimpleKMeans**.

2. Execução do experimento

- Inicialmente, executamos o algoritmo com diferentes valores de **k** (número de clusters), variando de 1 a 10.
- Para cada execução, anotamos o erro RMS (**Root Mean Square Error**) para comparar os resultados.

3. Análise dos resultados

- O erro RMS indica o quão bem os dados são agrupados dentro dos clusters. Quanto menor o erro, melhor a segmentação.
- Geramos um **gráfico** para visualizar como o erro RMS varia com o número de clusters e identificar o ponto de "diminuição acentuada" (**elbow method**).

K-Means foi ajustado para **k = 3**, o que faz mais sentido para o conjunto de dados **Iris**, que tem três classes naturais (**Setosa, Versicolor e Virginica**).

Resultados principais:

- **Número de iterações:** 3
- **Soma dos erros ao quadrado dentro dos clusters:** 7.8015 (bem menor do que no experimento anterior, indicando uma melhor separação)
- **Distribuição dos clusters:**
 - **Cluster 0:** 50 instâncias (**Setosa**)
 - **Cluster 1:** 50 instâncias (**Versicolor**)
 - **Cluster 2:** 50 instâncias (**Virginica**)
- **Centroides dos clusters:**
 - Cluster 0 (Setosa) → menor comprimento de pétala e sépala
 - Cluster 1 (Versicolor) → características intermediárias
 - Cluster 2 (Virginica) → maior comprimento de pétala e sépala

4. Determinação do melhor número de clusters

- No experimento realizado, verificamos que **k = 3** é a melhor escolha, pois corresponde às três classes naturais do conjunto de dados (**Setosa, Versicolor e Virginica**), e o erro RMS se estabiliza a partir desse ponto.

5. Conclusão

- O experimento confirmou que **o número ideal de clusters para o conjunto de dados "IrisDataSet" é 3**.
- O uso do gráfico de erro RMS foi essencial para validar essa escolha, aplicando o **método do cotovelo**.

Weka Explorer

PreprocessClassifyClusterAssociateSelect attributesVisualize

Clusterer

ChooseSimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Cluster mode

Use training set

Supplied test set

Percentage split

Classes to clusters evaluation

(Nom) variety

Store clusters for visualization

Set...

% 66

▼

Ignore attributes

Start

Stop

Result list (right-click for options)

14:19:46 - SimpleKMeans

14:26:31 - SimpleKMeans

Clusterer output

sepal.length
sepal.width
petal.length
petal.width
variety
Test mode: evaluate on training data

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

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,Versicolour
Cluster 1: 6.2,2.9,4.3,1.3,Versicolour
Cluster 2: 6.9,3.1,5.1,2.3,Virginica

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute	Full Data	Cluster#		
		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 Versicolour	Setosa	Virginica	

Time taken to build model (full training data) : 0 seconds

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

Clustered Instances

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

Status

OK

Log

Principais Notícias
Advogada de D...

Pesquisar

POR
INTL

14:26
28/03/2025

