Análise do Resultado do SimpleKMeans

1. Configuração do Algoritmo

a. Foi utilizado SimpleKMeans com N=10, mas na execução foram gerados 5 clusters.

RGM: 41295714

- b. Distância usada: Euclidiana.
- c. Número máximo de iterações: 500.

2. Características dos Dados

- a. Total de instâncias: 4.773
- b. 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

- a. Cluster 0 → 663 instâncias (14%)
- b. Cluster 1 → 834 instâncias (17%)
- c. Cluster 2 → 269 instâncias (6%)
- d. Cluster 3 → 873 instâncias (18%)
- e. **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

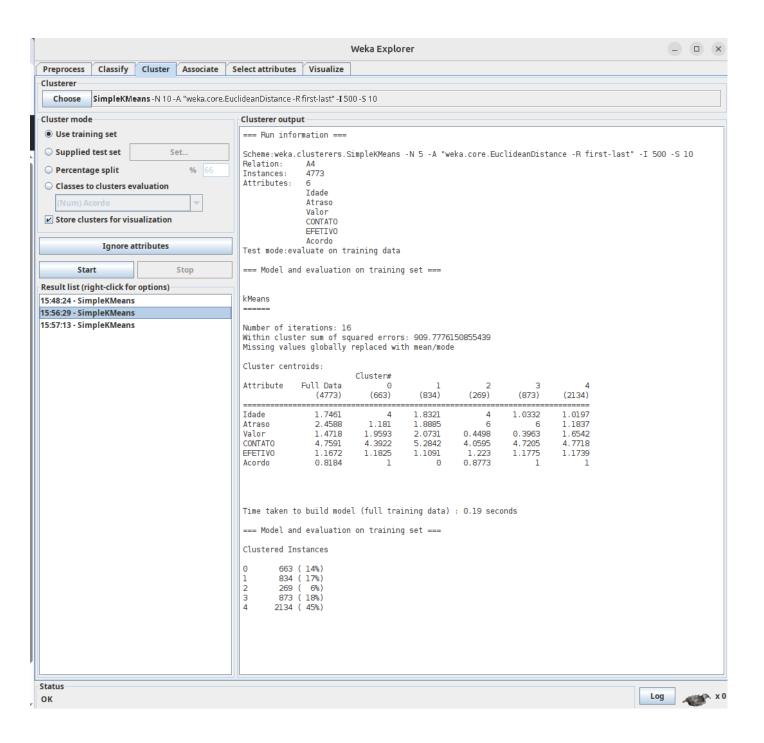
a. Exemplo: Pode ter um cluster de clientes que pagam em dia, outro de inadimplentes, outro de novos clientes, etc.

2. A priori descobrimos padrões de comportamento dentro desses grupos

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

3. Interpretação Complementar

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



1. Carregamento da base de dados

- Abrimos o Weka Explorer e carregamos o arquivo "iris.csv".
- Selecionamos 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:
 - o Cluster 0: 50 instâncias (Setosa)
 - Cluster 1: 50 instâncias (Versicolor)
 - Cluster 2: 50 instâncias (Virginica)

• Centroides dos clusters:

- o 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.

