Resultados do uso do algoritmo K-médias

Augusto Ribas[1], Bruno Nazário[1] e Doglas Sorgatto[1]

[1]Faculdade de Computação - Universidade Federal de Mato Grosso do Sul

Resumo

fhdfdjfdjfdjfhd

1 Introdução

jdhjdfhjdfhdhf

1.1 Problema

fjdhfjhfdjfd

1.2 Objetivos

Implementar o algoritmo K-médias sem utilizar as bibliotecas prontas disponíveis no Scikit-Learn [1] e aplicá-lo nos conjuntos de dados fornecidos pelo professor avaliando o desempenho

2 Material e Métodos

fhdfhdfdjfhdjfd

2.1 Algoritmos de Agrupamento

Clustering, ou Agrupamento, é uma técnica de *Data Mining* para fazer agrupamentos automáticos de dados segundo seu grau de semelhança. O critério de semelhança faz parte da definição do problema e, dependendo, do algoritmo conforme se lê em [3].

Normalmente o usuário do sistema deve escolher *a priori* o número de grupos a serem detectados. Estes grupos são formados por equações que calculam a "semelhança" entre os dados através de funções de distância.

Os tipos de algoritmos de agrupamento de dados mais comuns são os: Particionais e os Hierárquicos. Os *particionais* procuram criar grupos de semelhança.

De acordo com [2], "Nos *Hierárquico* o processo de identificação de grupos (clusters) é geralmente realimentado recursivamente, utilizando tanto objetos quanto grupos já identificados previamente como entrada para o processamento. Deste modo, constrói-se uma hierarquia de grupos de objetos, no estilo de uma árvore".

2.1.1 K-médias

O algoritmo de agrupamento de dados K-Médias agrupa um conjunto de instâncias em k partições, sendo k um número pré-estabelecido. O arranjo dos elementos é feito de maneira que um elemento pertença a um cluster, cujo centro o elemento é mais próximo. Dessa maneira, o algoritmo K-Médias consegue encontrar k partições disjuntas, buscando sempre minimizar a variância intra-cluster e maximizar a variância inter-cluster. Como critérios de convergência usuais, pode-se citar o número de iterações que o algoritmo executa e o número de realocações de clusters.

2.1.2 Arvore de Decisão

fdfhdjfhdfdhj

2.2 Procedimentos gerais

fjdhfjdfdjfdjhf

2.3 Conjuntos de dados

Para este trabalho, 9 conjuntos de dados foram utilizados. Todos estes conjuntos são bi-dimensionais (isto é, têm 2 atributos), com o número de clusters variando de 2 a 10 (o primeiro conjunto tem 2 clusters, o segundo tem 3 clusters, e assim por diante). Todos estes conjuntos de dados apresentam uma partição de referência (grupo de cada um dos pontos), sendo perfeitamente balanceados (30 exemplos por grupo), como se observa na tabela 1.

Tabela 1: Características gerais dos conjuntos de dados

Nome do conjunto	Número de instâncias	Número de grupos
artificial_2.data	60	2
artificial_3.data	90	3
artificial_4.data	120	4
artificial_5.data	150	5
artificial_6.data	180	6
artificial_7.data	210	7
artificial_8.data	240	8
artificial_9.data	270	9
artificial_10.data	300	10

Cada um dos conjuntos de dados está disposto em um arquivo no formato CSV (*"comma separated values"*). Em cada linha há um exemplo (instância) da base, no formato:

$$ValorAtributo\_1, ValorAtributo\_2, Particao\_de\_Referencia$$

Sendo utilizado para o processamento apenas os dois primeiros atributos.

3 Resultados e Discussão

Tabela 2: Comparação de eficiência - Menor SSE

K	Seed	SSE	Convergir	Min	Max	Fora	Acurácia
2	2	1.17178486551	5	29	31	1	0.98333
3	2	1.82592591238	5	30	30	0	1.00000
4	0	2.40978786268	9	28	32	2	0.98333
5	6	2.47892858032	8	29	31	2	0.98666
6	10	3.38661981209	8	28	34	5	0.97222
7	4	3.64855687389	8	29	31		0.98571
8	10	4.62062310715	5	25	35		0.96250
9	16	4.6656708739	13	28	32		0.98518
10	10	5.70866779082	11	28	32		0.98333

Tabela 3: Comparação de eficiência - Maior SSE

K	Seed	SSE	Convergir	Min	Max	Fora	Acurácia
2	0	1.18583121013	4	27	33	3	0.95000
3	8	4.18779991225	7	30	30	0	1.00000
4	12	7.00024804537	4	28	32	2	0.98333
5	8	7.47642260367	50	0	61	60	0.60000
6	0	18.7881874706	50	0	123	120	0.33333
7							
8							
9							
10							

Figura 1: Visão geral dos dados antes do agrupamento

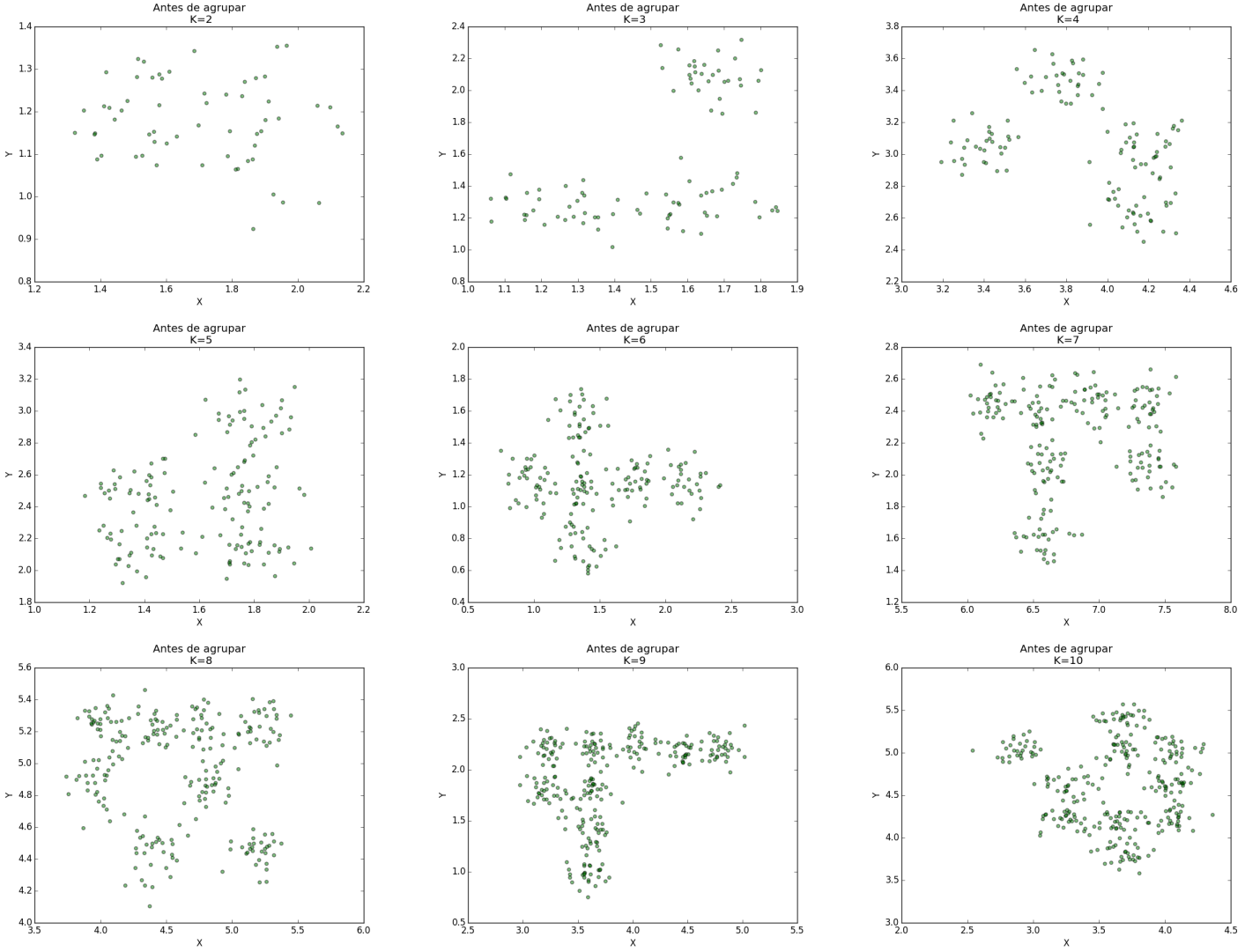


Figura 2: Evolução dos Somatórios de Erro Quadrático - SSE

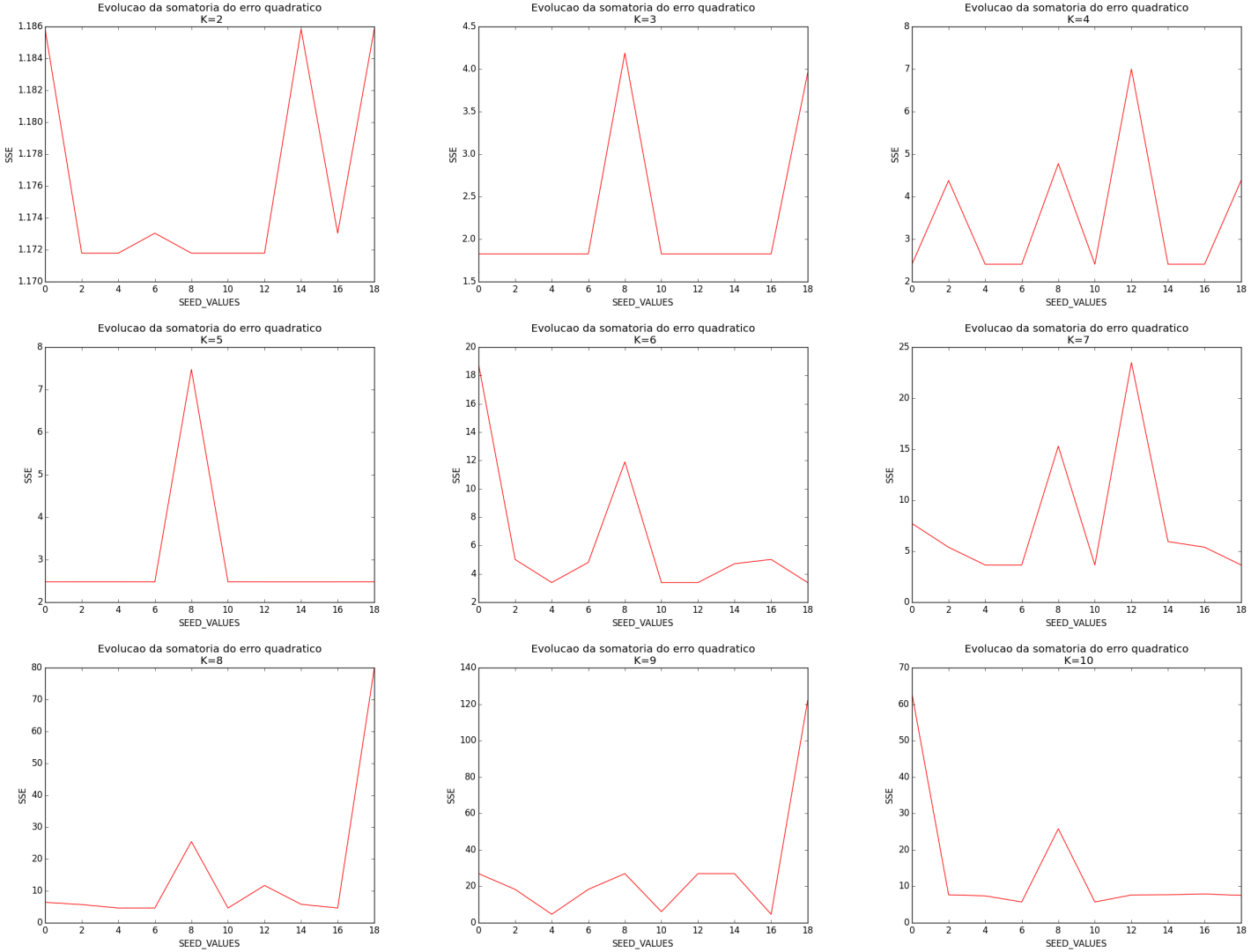
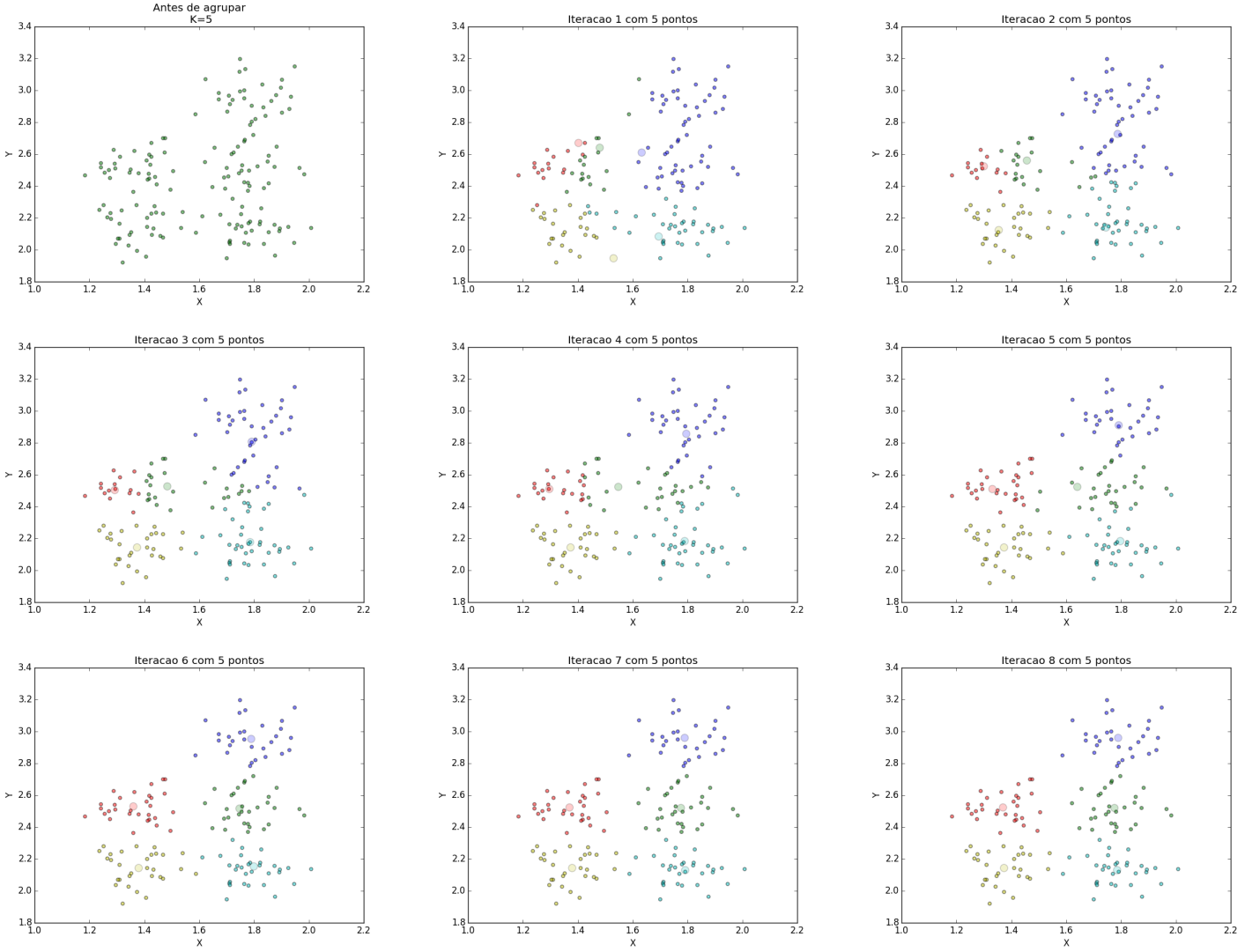


Figura 3: Exemplo de evolução do processo de agrupamento para k = 5



Referências

[1] Scikit-learn: Machine Learning in Python.

[2] Sopa de Letrinhas. Algoritmos de análise de agrupamentos: Método hierárquico – agnes e diana. 2012.

[3] Wikipedia. Clustering. 2015.