K-Means
Trabalho 2 - INF01017 - 2020/2

Cassiano Bartz[1], **Douglas Lázaro Silva**[1], **Nícolas Vincent D. Pessutto**[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

{cmbartz, dlspsilva, nvdpessutto}@inf.ufrgs.br

1. Introdução

Este documento consiste no relatório do Trabalho 2 da disciplina de Aprendizado de Máquina, contendo uma descrição das características gerais da implementação do algoritmo K-Means, bem como uma análise sobre os experimentos realizados pelo grupo, a partir dos dataset disponibilizados para a realização do trabalho.

2. Implementação K-Means

A implementação do algoritmo foi feita em python 3 usando jupyter notebook, seguimos a definição teórica do algoritmo separando cada etapa descrita em funções auxiliares utilizando como estrutura de dados pandas dataframe e numpy para as operações matemáticas em arrays, detalhados logo abaixo:

Algoritmo 1: Algoritmo K-means

Entrada: instances, kParam, distMode, loopLimit, bestCentroids
Resultado: bestCentroids, instances clusters
inicialização;
hadUpdate = Verdadeiro;
count = 0;
se ***tem*** *bestCentroids* **então**
  centroids = *bestCentroids*;
senão
  centroids = **initCentroids**();
fim
enquanto *hadUpdate* ***E*** *count* ***menor que*** *loopLimit* **faça**
  para cada *instancia* ***em*** *instancias* **faça**
    centerIndex = **minCentroid**();
    instance $\leftarrow centerIndex$ **fim**
    updatedCentroids = [];
    para cada *centroid* ***em*** *centroids* **faça**
      newCenter = **updateCentroid**();
      updatedCentroids $\leftarrow newCenter$
    fim
    hadUpdate, centroids = **isUpdated**();
    count += 1;
  fim
  retorna *centroids*;

2.1. Entradas

Os parâmetros de entrada conforme descrito no pseudo-código do algoritmo são:

- **instances**
  contendo o dataset das instances em pandas dataframe
- **kParam**
  inteiro que defina a quantidade de clusters.
- **distMode**
  string utilizada para especificar a modo que é calculada a distancia, podendo ser 'euc', 'ave', 'man' para distância Euclideana, Average-Euclideana e Manhattan respectivamente.
- **loopLimit** parametro resonsável por interromper as atualizações dos centroids conforme o número de repetições do laço principal.
- **bestCentroids** lista de centroids para inicialização não aleatória e atribuição direta dos valores iniciais de centroids

2.2. Inicialização Centroids

A função *InitCentroid* é responsável pela inicialização aleatória dos centroids que consiste em para cada feature atribuir um valor aleatório entre o valor mínimo e máximo daquela feature contido no dataset, dessa forma a há uma variação nas inicializações sem gerar pontos fora do intervalo das features do dataset.

2.3. Cálculo da distância

A função *minCentroid* é responsável por aplicar o cálculo de distancia de cada centroid atual e cada instancia do dataset. Para o calculo da distancia foi utilizado o parâmetro *distMode* para definir o método e foram implementados 3 metódos diferentes utilizados no conjunto de testes inicial *USArrests* para observar as implicações de mudar a formúla para obter a medida de distância entre pontos no espaço.

Após ser calculada a distância da instancia para todos os centroids é retornado o label do centroid com menor distância do conjunto.

Parâmetro	Descrição	Fórmula
euc	Distância Euclideana	$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
ave	Distância Média Euclideana	$\sqrt{1/n \sum_{i=1}^{n}(x_i - y_i)^2}$
man	Distância Manhattan	$\sum_{i=1}^{n} \lvert(x_i - y_i)\rvert$

Tabela 1. Cálculo de Distância

2.4. Atualização de centroid

A atualização dos valores dos centrois foi divida em duas funções *updateCentroid, isUpdated* respectivamentes, *updateCentroid* seguindo o fluxo de execução do algoritmo é encarregada de após atribuir labels que identificam a qual cluster pertence a instâcia calcula a média de cada feature do conjunto do cluster e atribui esse valor obtido ao respectivo centroid da label. Depois de ter executado o calculo das médias das features das instancias do cluster para cada centroid a segunda função (*isUpdated*) é chamada para conferir se houve mudança de valores dos centroids e em caso negativo encerra o laço principal do algoritmo.

2.5. Validação da Implementação

Para validar e testar se a implementação do algoritmo estava funcionando corretamente utilizamos o dataset USA Arrest, mesmo conjunto de dados utilizado nos exemplos contidos nos slides da disciplina, combinado com funções auxiliares de plotagem para verificar a execução do algoritmo e classificação dos dados utilizando como *Ground Truth* as informações exbidas nos slides da disciplina.

Logo a seguir são apresentados os resultados obtidos da validação do algoritmo:

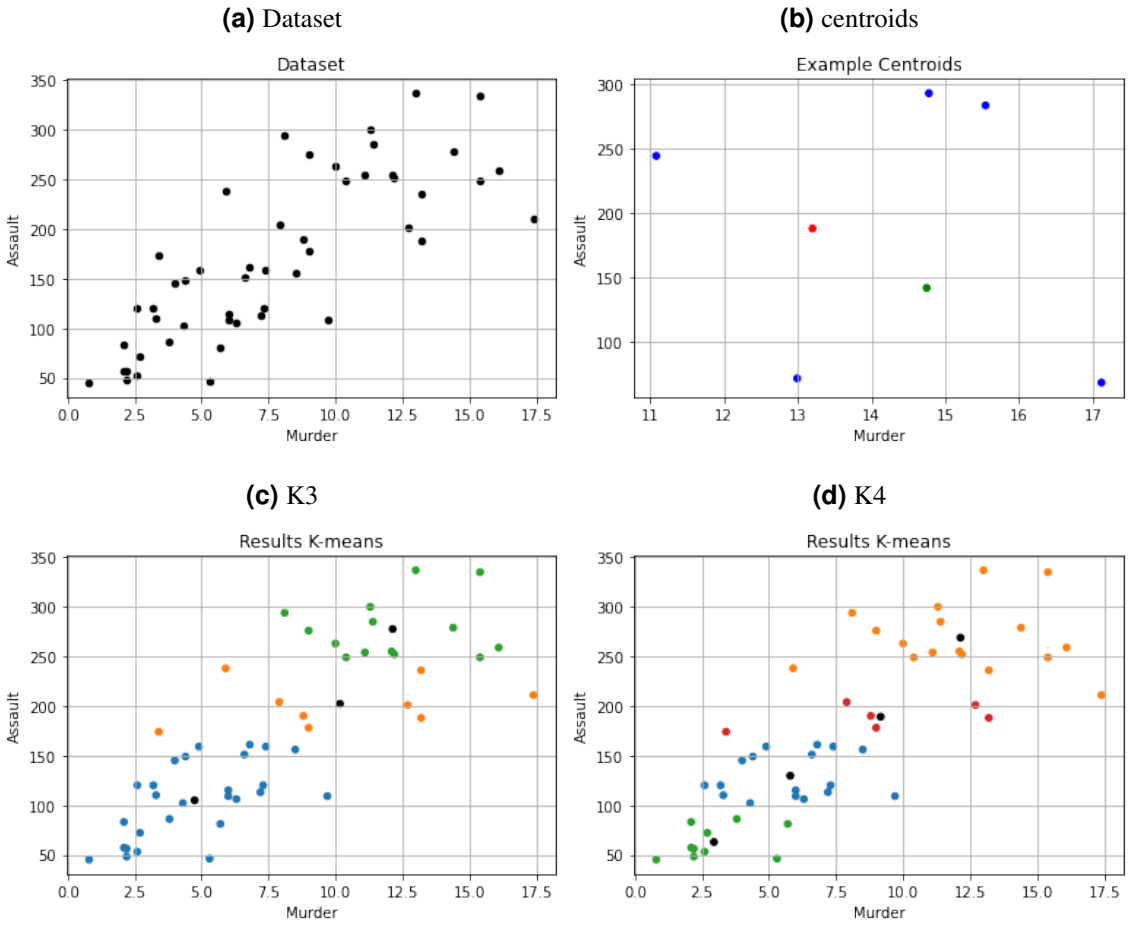

Figura 1. Validação K-means USA arrest dataset

2.6. Dissimilaridade

Para calcularmos a dissimilaridade, executamos 100 vezes o algoritmo K-Means para um determinado K, desse modo, inicializamos aleatoriamete os centroids para cada uma dessas execuções. Ao obtermos os centroids de retorno do K-Means, calculamos a dissmilaridade para cada centroid do retorno em relação às instâncias pertencentes ao cluster do mesmo. Para tal, utilizamos um dos modos de cálculo distância disponibilizados na nossa implementação. Verificadas as dissmilaridades obtidas para cada conjunto de centroids, elas são adicionadas à um array para no final retornarmos o conjunto de centroids, dentre as 100 execuções, que possui a menor dissimilaridade.

2.7. Elbow Method

Nossa função para a implementação do Elbow Method part do princípio de pegar o conjunto de centroids com a melhor dissimilaridade obtida. Fazendo isso para cada K no intervalo entre 1 e 10. Obtidos estes centroids, colocamos eles em um array e demonstramos em um gráfico as dissimilaridades obtidadas para cada K. Entretanto, a escolha do melhor K não é feita pelo algoritmo, ao demonstrar o gráfico, o grupo analisou visualmente e escolheu o melhor K para cada hipótese.

2.8. Execução da implementação

Conforme mencionado anteriormente a aplicação foi desenvolvida utilizando jupter notebooks, e os parâmetros de execução estão sendo passados / modificados direteamente nas celulas conforme os comentários entre as células e funções do código.

Acesso ao repositório
`https://github.com/dlazarosps/Kmeans-algorithm/`

3. Análise e Experimentos

Foram selecionadas três combinações de áreas dos datasets para a confecção das hipóteses. Hobbies e Personalidade, Musicas e Filmes e Personalidades com Hábitos de Gastos. Para cada combinação, foram elaboradas as seguintes hipóteses que intrigam a sociedade atual:

- Personalidade e Hábitos de Consumo
  - Pessoas são workaholics pensando nas suas finanças?
  - Pessoas que pensam no futuro costumam gastar em roupas?
- Músicas e Filmes
  - Quem gosta de Metal também gosta de filmes românticos?
  - Quem escuta Punk assiste filmes de Fantasia?
  - Quem escuta Reggae gosta de Ficção Científica?
- Hobbies e Personalidades
  - Pessoas que gostam de Matemática são pessoas que tentam ser engraçadas?

Estes foram as hipóteses executadas e analisadas pelo grupo. Pórem, neste relatório apresentaremos apenas uma hipótese de cada conjunto, por acreditarmos que houve uma repetição de conceitos muito frequente entre eles.

3.1. Personalidades e Hábitos de Consumo

3.1.1. Pessoas são workaholics pensando nas suas finanças?

Para analisar a hipótese de quem é Workaholic (pessoas que estudam e trabalham mesmo no tempo livre) tende a guardar mais dinheiro, selecionamos esses atributos no dataset.
Como podemos ver na Figura 2, o gráfico de dissimilaridade para o número de clusters não possui um elbow saliente. Portanto, para analisarmos melhor essa hipótese, decidimos analisar os clusters referentes à K=3, K=4 e K=5. A nossa escolha para fazermos a análise dessa maneira vem do raciocício de que um desses deve ser o nosso elbow, mesmo que visualmente não seja tão fácil identificar.

Olhando para o gráfico do K-Means para K=5 na Figura 3, podemos analisar os dados da seguinte maneira:

- Pessoas que não são workaholics também não tem a tendência de guardar seu dinheiro, isso é observável olhando para o centróide do cluster de cor azul. Como pode ser difícil olhar pela cor, o centróide tem coordenadas aproximadas de (1.4,1.9). O que pode ser visto é que pessoas que tendem a não concordar com a frase de que trabalham e estudam no tempo livre, tendem também a não concordar com a frase de que guardam seu dinheiro.

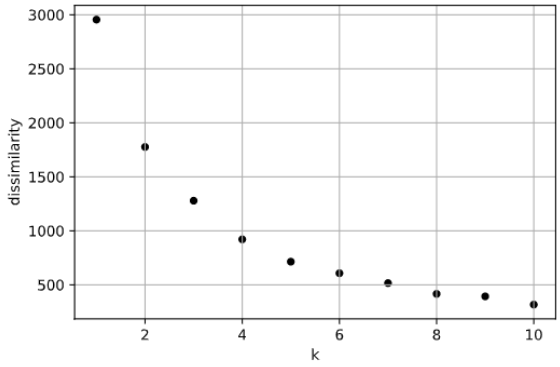

Figura 2. Elbow Method para Workaholism x Finances

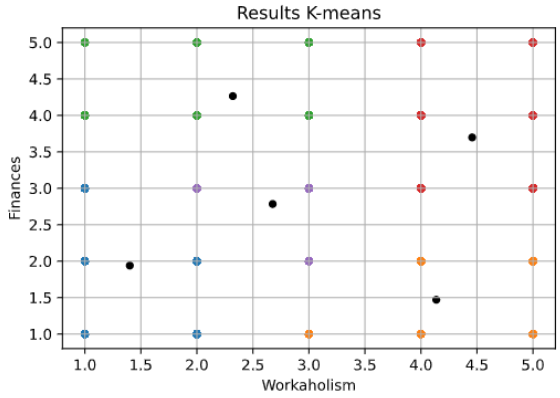

Figura 3. K-means para K=5 Workaholism X Finance

- Não foi possível encontrar um elo entre pessoas que são workaholics e a tendência de guardar seu dinheiro. Excluíndo os centróides aproximados em (2.3,4.2) e (2.6,2.7) por estarem muito perto de 2.5, ou seja, neutros em relação à pelo menos uma das frases, ficamos apenas com mais dois centróides, (4.1,1.4) e (4.4,3.7). Como podemos ver por estes dois centróides restantes, temos grupos de pessoas que concordam bastante com a frase de trabalhar e estudar no seu tempo livre, mas ao mesmo tempo, um dos grupos não tem a tendência alta de guardar o dinheiro, enquanto o outro tem uma tendência considerável de fazer isso.

Ao analisarmos o gráfico para K=3 na Figura 4, fica mais difícil chegar a uma conclusão. Isso pois os centróides aproximados em (2.8,4.2), (2.2,2.3) e (4.6,2.8) possuem um ou ambos os eixos com valores muito próximos de 2.5, o que seria considerado pessoas neutras. Baseado nisso, com K=3 não tivemos um ganho de informação útil.

Já ao analisarmos o gráfico para K=4 na Figura 5, podemos observar um comportamento semelhante ao comportamento visto no K=5, onde não conseguimos encontrar um elo entre pessoas workaholics e a tendência de guardar dinheiro. Excluindo da análise os centróides (2.1,1.6) e (2.3,3.6), pelo mesmo motivo de neutralidade referente a um dos atributos, ficamos com os centróides (4.5,1.6) e (4.4,3.7). Esses centróides restantes, demonstram que pessoas workaholics podem tanto ter a tendência de guardar seu dinheiro, quanto não ter. Em comparação com K=5, com 4 clusters não foi possível identificar a informação de que pessoas não workaholics não guardam seu dinheiro, motivo que nos

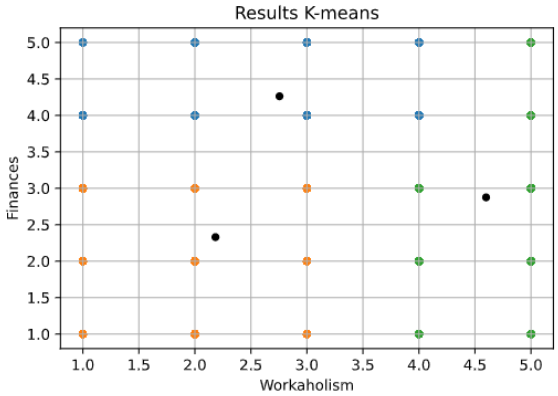

Figura 4. K-means para K=3 Workaholism X Finance

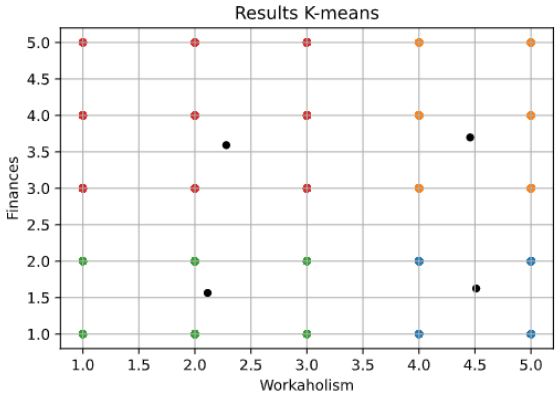

Figura 5. K-means para K=4 Workaholism X Finance

leva a concluir que para esta hipótese, o valor ideal de K é 5.

3.2. Músicas e Filmes

3.2.1. Quem gosta de Metal também gosta de filmes românticos?

Ao considerar a hipótese de quem gosta de Metal também gosta de filmes românticos, foram selecionados os respectivos atributos do dataset e gerado o gráfico do Elbow Method, como a figura 6 mostra a seguir:

Ao analisar o gráfico, não se nota um claro ponto de quebra, onde se escolheria o melhor k para a execução do algoritmo. Neste caso, há uma dúvida entre 2, 3 ou 4 clusters. A figura 7 compara as execuções com 3 e 4 clusters.

Ao optarmos pela execução com 4 clusters como melhor k retornado pelo Elbow, comparamos com a execução de 3 clusters, e percebemos que o k-means com K=3 colocaria no mesmo grupo pessoas que gostam de Metal e que gostam ou não de filmes românticos. Analisando, deste ponto de vista, a escolha por 4 clusters se demonstra mais apropriada, pois ele consegue diferenciar este grupo com o centroid adicional. Refletindo sobre a posição dos clusters, podemos indicar que, apesar do espaçamento simétrico entre eles, o grupo de pessoas que preferem mais filmes românticos do que músicas de Metal

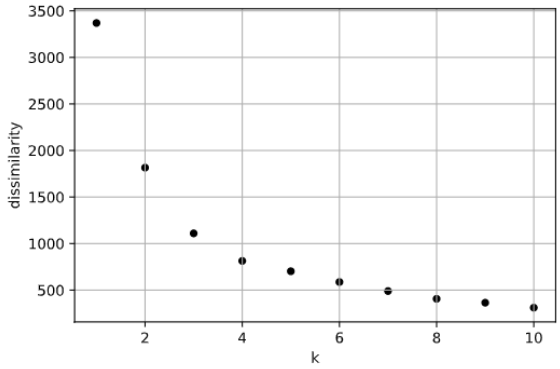

Figura 6. Elbow Method para Metal x Romantic

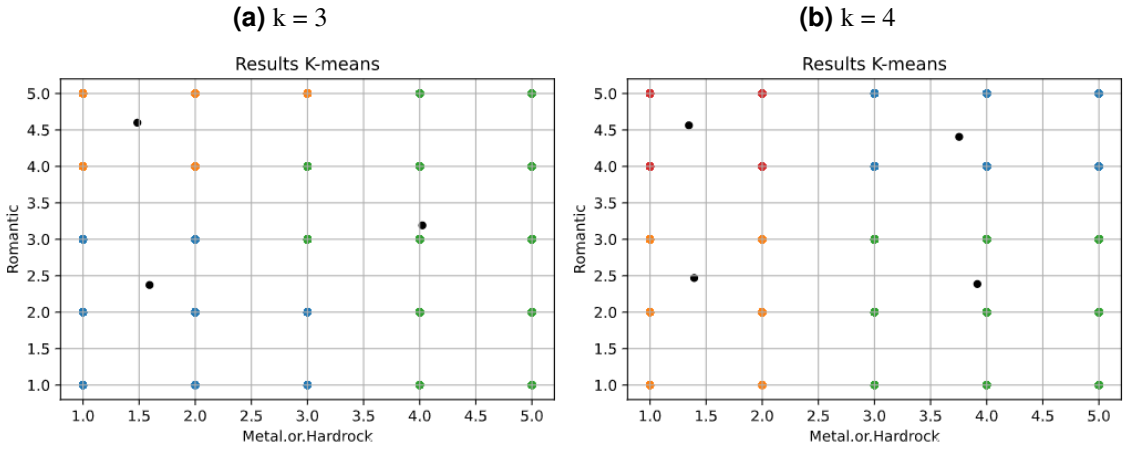

Figura 7. Execuções com 3 e 4 clusters.

é um grupo com opiniões mais fortes, com notas de preferência superiores do que em comparação ao grupo que prefere o gênero musical.

3.3. Hobbies e Personalidades

3.3.1. Pessoas que gostam de Matemática são pessoas que tentam ser engraçadas?

O gráfico do Elbow Method novamente é de certa forma inconclusivo visualmente, nos forçando a analisar as execuções com K mais propensos a serem os melhores, de acordo com o resultado obtido. Podemos observar na figura 8 que os pontos mais prováveis de melhor K são os de 3, 4 ou 5 centroids.

Similar ao comportamento observado no Workaholism x Finances, o valor de K que mais nos dá informação é K=5. Como podemos ver na Figura 9, podemos extrair as seguintes informações:

- Pessoas que não tentam sempre ser as mais engraçadas, tendem também a não gostar de Matemática. Isso é observável no centróide aproximado de (1.6,1.4).
- Para pessoas que tentam ser as mais engraçadas, não foi possível encontrar um elo com o gosto por matemática. Isso acontece porque temos dois centróides excluídos da análise, os centróides aproximados em (2.6,3.6) e (2.9,3.7), por questoes de neutralidade em uma das frases. Nos resta os centróides em (4.4,1.4), onde

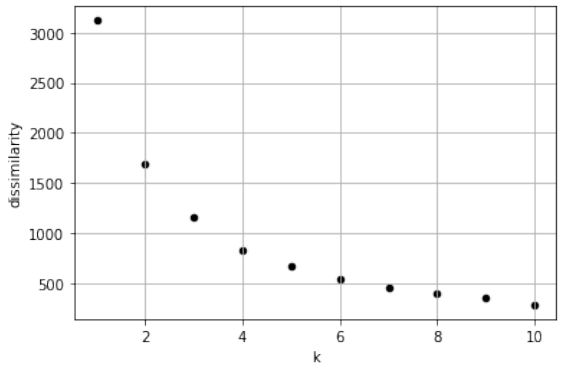

Figura 8. Elbow Method para Math x Fun

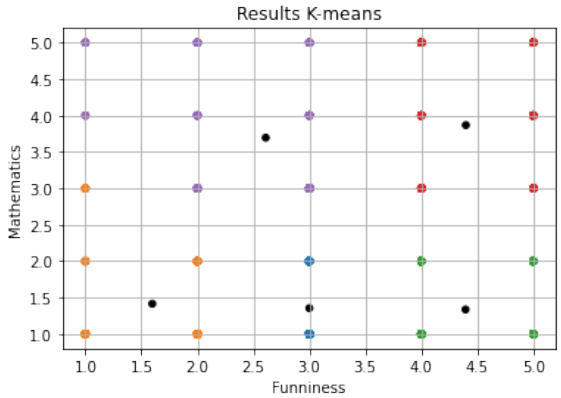

Figura 9. K-means para K=5 Funniness X Mathematics

são pessoas que tentam ser as mais engraçadas, mas, não gostam de matemática, e o centróide (4.4,3.9), onde temos pessoas que tentam ser engraçadas e gostam de matemática também. Por esse motivo, não foi possível concluir em um elo.

- Uma análise extra que pode ser feita, sem excluir nenhum centróide, é que pessoas que gostam de matemática, são ao menos neutras em tentar ser as mais engraçadas, ou concordam fortemente com a frase. O que significa que não é identificado nenhum cluster de pessoas que gostam de matemática, mas que não tentam ser as mais engraçadas.

Como podemos ver na Figura 10, assim como foi com Workaholism X Finances, K=3 e K=4 acaba nos dando menos informações e não é possível extrair a mesma quantidade de informações como no K=5. Por esse motivo consideramos que K=5 é o valor de K ótimo para essa hipótese.

4. Conclusão

A implementação do algoritmo se demonstrou de certa forma simples, o que nos tomou mais esforço foi a análise dos resultados. Os resultados do Elbow Method não foram tão claros, necessitando uma análise mais profunda entre os K que aparentavam ser a melhor escolha. Outra dificuldade foi em relação à sobreposição de instâncias no gráfico, uma vez que a demonstração do mesmo foi em duas dimensões e os datasets analisados possuíam valores fixos e repetitivos, impossibilitando o reconhecimento visual da

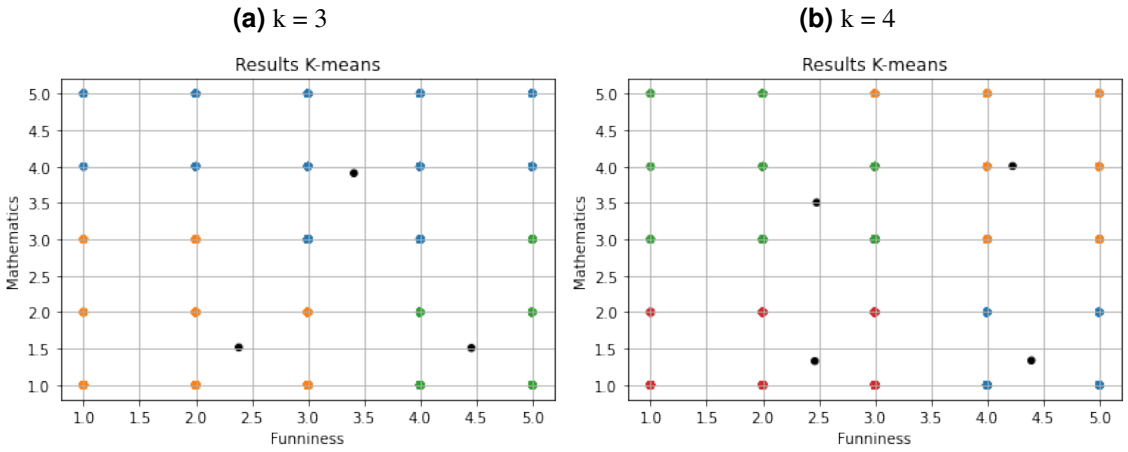

Figura 10. Execuções com 3 e 4 clusters para Mathematics X Funniness.

clara quantidade de instâncias pertencentes à um determinado cluster. O algoritmo desenvolvido suporta a análise de n atributos, não se restringindo à apenas dois, como utilizamos em nossos experimentos. Porém, a visualização para mais do que 2 atributos se mostrou complexa, dificultando os esclarecimentos para nossas hipóteses. Logo, o Elbow Method, apesar de não nos apresentar uma resposta clara visual de melhor k aos nossos datasets, nos indicou a melhor aproximação de K para as nossas execuções, as quais pudemos nos decidir posteriormente ao executá-las. O video de apresentação do trabalho pode ser acessado pelo link `https://drive.google.com/file/d/1B--q5zYY6mki8_Ul9VUn7bTj6OqhtoZj/view?usp=sharing`