Clusterização e Agrupamento

Rodrigo Rusa
208592
rodrigorusa@gmail.com

Thales Eduardo Nazatto
074388
tenazatto@gmail.com

I. Introdução

Em Aprendizado de máquina, clusterização ou agrupamento são técnicas de aprendizado não-supervisionado utilizadas para descobrir padrões não descobertos anteriormente sem depender de dados ou padrões anteriores. Dentre elas, a mais comum é o *K-Means*, que consiste em achar agrupamentos em áreas definindo um número *k* de centróides. Um algoritmo *K-Means* se define da seguinte maneira:

1) Definir quais os *k* centróides.
2) Encontrar o centróide mais próximo e atualizar os agrupamentos com base nesses centróides.
3) Mover os centróides para o centro dos seus agrupamentos.
4) Repetir os passos 2 e 3 até a convergência do algoritmo.

Apesar de ser a mais comum, o *K-Means* não é a única técnica existente em clusterização. Ela é considerada uma técnica de clusterização particional, ou seja, que divide o todo dos dados em pequenos agrupamentos. Existem também as técnicas de clusterização hierarquicas, que consistem no inverso: Junta todos os agrupamentos de dados em um único todo, dividindo-os em uma hierarquia.

Existem dois tipos de técnicas hierárquicas: *Agglomerative Clustering* e *Divisive Clustering*. Enquanto a segunda faz uma abordagem *top-down*, começando de um único agrupamento e, com isso, se dividindo e formando toda a hierarquia, a primeira começa com vários pequenos agrupamentos que vão se aglomerando e formando a hierarquia dessa forma. Um algoritmo de *Agglomerative Clustering* se define da seguinte maneira:

1) Calcular a matriz de proximidade, se necessário.
2) Unir os dois agrupamentos mais próximos.
3) Atualizar a matriz de proximidade para refletir a proximidade entre o novo agrupamento e os agrupamentos originais.
4) Repetir os passos 2 e 3 até sobrar apenas um agrupamento.

No caso da proximidade, ela pode ser calculada pelos seguintes critérios:

- **Link simples/MIN:** Proximidade entre os dois pontos mais próximos entre os agrupamentos.
- **Link completo/MAX:** Proximidade entre os dois pontos mais distantes entre os agrupamentos.
- **Média:** Proximidade média entre todos os pontos dos agrupamentos.
- **Centróides:** Proximidade entre os centróides dos agrupamentos.
- **Critério de Ward:** O aumento na variância para o agrupamento que está sendo aglomerado.

Para reduzir o número de *features* e diminuir o tempo computacional necessário para processar um algoritmo de aprendizado sem afetar muito sua precisão, o PCA (*Principal Component Analysis*) é um dos algoritmos a serem utilizados. Nele, depois de utilizarmos o *feature scaling* para normalização dos dados, é calculado a matriz de co-variância presente em (1), calculados os auto-vetores dessa matriz utilizando decomposição em valores singulares presentes em (2), e logo após utilizando os valores de S para encontrar as *features* mais determinantes desconsiderando a menor variância possível (geralmente tolerância de 0,01) presentes em (3).

$$\Sigma = \frac{1}{m}\sum_{i=1}^{n}(x^{(i)})(x^{(i)})^T \tag{1}$$

$$[U, S, V] = svd(\Sigma) \tag{2}$$

$$\frac{\frac{1}{m}\sum_{i=1}^{m}||x^{(i)} - x_{approx}^{(i)}||^2}{\frac{1}{m}\sum_{i=1}^{m}||x^{(i)}||^2} = 1 - \frac{\sum_{i=1}^{k} s_{ii}}{\sum_{i=1}^{n} s_{ii}} \leqslant 0,01 \tag{3}$$

Neste relatório, iremos apresentar na prática as técnicas para clusterização e agrupamento *K-Means* e *Agglomerative Clustering* e verificar o comportamento da aplicação de redução de dimensionalidade utilizando *PCA* com o objetivo de agrupar e extrair informações de dados do *Twitter* sobre saúde [1] usando o *framework Scikit-Learn* em Python.

II. Metodologia

A fim de agrupar e extrair informações do conjunto de dados em questão foram realizados alguns experimentos. Para a representação dos dados do *Twitter* em formato texto para vetores de *features* foi utilizada a técnica *Bag of Words*, que transforma textos em vetores de frequência de palavras. Nessa abordagem é definido um conjunto de palavras (*bag*) como vocabulário. Esse conjunto de palavras pode ser definido pelas palavras extraídas dos textos de entrada ou *a priori* pelo contexto dos dados. Preposições, artigos e acentos ortográficos são removidos, e apenas as palavras no formato radical são consideradas.

Após a definição do conjunto, as sentenças ou textos são codificados como vetores de ocorrência das palavras do mesmo, fazendo com que o vetor de *features* codificado tenha dimensionalidade igual ao número de palavras do conjunto definido, garantindo o aproveitamento de todos os dados da base. Contudo, tal abordagem não leva em consideração o

contexto gramatical das sentenças nem a ordem das palavras, que podem ter grande influência na interpretação de um texto feita por uma pessoa.

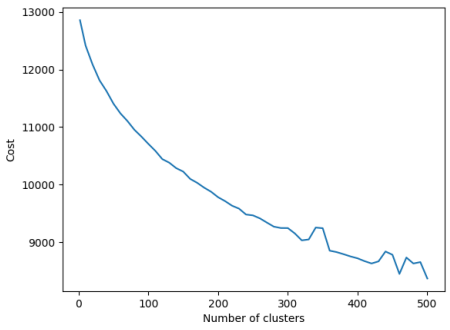
Devido ao fato da base de dados escolhida não possuir um número expressivo de exemplos, 13,229, em todos os experimentos realizados foram utilizados todos os dados da base de dados para o agrupamento sem divisão do dados em conjunto de treinamento e teste. Dentre os experimentos realizados nesse trabalho, foi avaliado duas técnicas distintas de agrupamento, o *K-Means*, técnica de agrupamento de propósito geral muito utilizada na literatura e o *Agglomerative Clustering*, técnica de agrupamento hierárquico muito utilizado em casos de dados com alto número de grupos (*clusters*). Além disso, foram feitos experimentos para encontrar o valor ideal de grupos que melhor agrupam os dados da base de *Tweets* e a influência da aplicação de uma redução de dimensionalidade nos dados utilizando a técnica *PCA*.

III. Experimentos e Discussão

Com o objetivo de agrupar os dados do *Twitter* relacionados a saúde, inicialmente foi aplicada a técnica de agrupamento *K-Means*. Essa técnica de agrupamento tem como parâmetro de entrada o número de grupos a serem utilizados. Como esse valor é desconhecido foi utilizado uma técnica denominada *Elbow* para encontrar o valor ou um intervalo de valores possíveis de k (número de grupos) que agrupe os dados com boa qualidade. Essa técnica consiste em aplicar o algoritmo *K-Means* para diferentes valores de k e encontrar o menor valor ou valores de k em que a função custo *Inertia*, ou soma dos quadrados dentro dos grupos, possua pequena variação.

Para a demonstração da efetividade da técnica *Elbow* foi realizado um experimento variando o valor de k de 2 a 500 com passos de 10 aplicando-se a técnica *K-Means*. Nesse experimento foi utilizado inicialização aleatória, 100 inicializações com diferentes sementes e número máximo de iterações $= 500$. A Figura 1 a seguir ilustra o resultado do experimento.

Figure 1: Função custo *Inertia* para diferentes valores de k



Com base no gráfico podemos observar que a função custo apresenta maior variação para valores de k menores que 200, para valores maiores, podemos ver que a função custo apresenta menor variação. Desse modo, a fim de encontrar um intervalo de valores ideais de k foi realizada uma análise da variação entre o intervalo (200, 300) como ilustra a Tabela I a seguir:

Table I: Análise da variação da função custo pelo número de grupos

Intervalo de k	Variação
(200, 210)	67
(210, 220)	80
(220, 230)	51
(230, 240)	102
(240, 250)	**26**
(250, 260)	52
(260, 270)	73
(270, 280)	71
(280, 290)	23
(290, 300)	1

Analisando a Tabela podemos ver que a variação da função custo pode aumentar ou diminuir dependendo do número de k escolhido, e um intervalo pequeno de variação apenas pode não ser ideal devido ao poder computacional necessário para rodar o algoritmo, que aumenta conforme o número de k também aumenta. Dada a análise, podemos concluir que o menor intervalo de valores de k com menor variação da função custo é o intervalo (240, 250), e portanto o mesmo foi escolhido como intervalo confiável para k.

A fim de averiguar a qualidade e o valor semântico dos grupos formados pelos valores de k definidos, foi realizado um experimento onde foram escolhidos 3 grupos aleatórios dentre o número total de grupos (k) e posteriormente extraído os seus *medoids* e vizinhos mais próximos. Para esse experimento foi utilizado valor de $k = 245$ por ter apresentado grupos com mais coerência semântica. Os resultados são ilustrados na Tabela II.

Com base nos resultados do experimento realizado podemos observar que os 3 grupos analisados apresentam sentido semântico, ou seja, apresentam similaridade de conteúdo. Por exemplo, o grupo 243 apresenta *Tweets* sobre alergia, já o grupo 115 apresenta *Tweets* sobre a doença Ebola e por fim o grupo 3 apresenta *Tweets* sobre Exercitar-se (*workout*).

Também podemos notar que o agrupamento ocorreu baseado apenas em palavras. Por exemplo, nos grupos 115 e 3 os *Tweets* foram agrupados pelas palavras *Ebola* e *workout*, respectivamente. Isso pode ser uma desvantagem dessa abordagem, pois a mesma não considera contexto semântico nem a ordem das palavras. Desse modo, é possível que ocorram agrupamentos de *Tweets* com ocorrência de uma mesma palavra porém com significado semântico diferentes.

Devido a impossibilidade de averiguar todos os *Tweets* de cada grupo formado e a necessidade de medi-los quantitativamente foram utilizadas 3 métricas de qualidade de agrupamento implementadas no *Scikit-Learn* que não requerem os rótulos reais dos dados (*ground truth*). A primeira métrica, o Coeficiente de Silhueta [2], mais comumente utilizado na literatura, é composto por duas pontuações:

- **a:** A distância média entre uma amostra e todos os outros pontos da mesma classe.
- **b:** A distância média entre uma amostra e todos os outros pontos no próximo grupo mais próximo.

O Coeficiente de Silhueta é então dado como:

$$s = \frac{b - a}{max(a, b)} \tag{4}$$

Table II: Análise dos *Tweets* dos *medoids* e vizinhos mais próximos de cada grupo

Grupo	Elemento	*Tweet*
243	*Medoid*	**Allergies getting you down? Join our #CNNAllergies chat on 4/23 at 1:30 p.m. EST with @AllergyReliefNY**
	Vizinho 1	*What can you do to get allergy relief? Join our #CNNAllergies chat on 4/23 at 1:30 p.m. EST with @AllergyReliefNY*
	Vizinho 2	*Allergies getting you down? Join our #CNNAllergies chat TOMORROW at 1:30pm EST with @AllergyReliefNY*
	Vizinho 3	*TODAY: Allergies getting you down? Join our #CNNAllergies Twitter chat at 1:30pm EST with @AllergyReliefNY*
	Vizinho 4	*Starting soon: Allergies getting you down? Join our #CNNAllergies Twitter chat at 1:30pm EST with @AllergyReliefNY*
	Vizinho 5	*Join our #CNNAllergies chat TOMORROW at 1:30pm EST with @AllergyReliefNY*
115	*Medoid*	**.@PeaceCorps pulling volunteers from #Ebola region**
	Vizinho 1	*How not to catch Ebola*
	Vizinho 2	*Journey through the Ebola heartland*
	Vizinho 3	*All nations 'have a stake' in Ebola*
	Vizinho 4	*WHO vows reform after Ebola 'shocks'*
	Vizinho 5	*Dozens monitored for Ebola in Mali*
3	*Medoid*	**Excuse-proof your workout**
	Vizinho 1	*A trick to boosting your workout:*
	Vizinho 2	*Do you have the workout blues? Here, 10 ways to get pumped to workout:*
	Vizinho 3	*What to eat after a workout*
	Vizinho 4	*Check out @Spotify's top #workout playlists*
	Vizinho 5	6 *workout wreckers to avoid:*

Valores de s próximos a 1, indicam que os grupos foram bem separados. Para valores próximos a -1, indicam que os grupos não foram bem separados e valores próximos a 0 indicam grupos sobrepostos.

A segunda métrica, *Davies-Bouldin Index* [3], é definida como a semelhança média entre cada grupo (*cluster*) e o mais similar. Valores pequenos, próximos de zero, indicam um agrupamento melhor dos dados.

Por fim, a terceira métrica utilizada, *Calinski-Harabaz Index* [4], é dada como a razão entre a média da dispersão entre os grupos e a dispersão dentro do grupo. Quanto maior o valor da métrica, mais densos e melhor separados os grupos são.

A Tabela III a seguir ilustra os resultados obtidos das métricas computadas para agrupamentos variando k de 240 a 250.

Table III: Resultados das métricas de qualidade

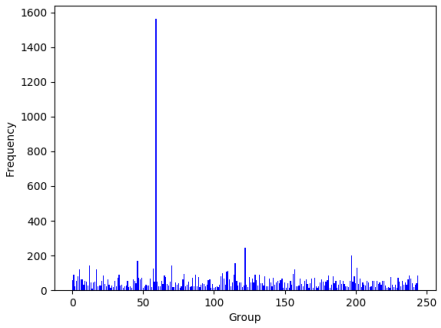
Número de grupos (k)	Silhueta	Davies-Bouldin	Calinski-Harabaz	Time (s)
240	0,075	3,069	19,512	340,235
241	0,076	3,110	19,628	348,619
242	0,076	3,083	19,310	335,653
243	0,075	3,106	19,397	346,789
244	0,076	3,076	19,538	328,633
245	**0,077**	**3,067**	**19,529**	**352,148**
246	0,076	3,083	19,588	332,73
247	0,076	3,107	19,343	344,473
248	0,076	3,021	19,331	345,548
249	0,076	3,082	19,380	347,208
250	0,077	3,145	19,147	334,53

Analisando os resultados podemos observar que as métricas apresentaram pequena variação para os valores de k entre 240 e 250. Além disso, podemos notar pelos valores das métricas computadas que os agrupamentos não apresentaram boa qualidade, pois apresentaram valores baixos de Coeficiente de Silhueta, distantes do valor 1 desejável, e valores altos, distantes de 0, para a métrica *Davies-Bouldin Index*.

A fim de entender o motivo das métricas de qualidade de agrupamento apresentarem resultados não satisfatório foi analisado a distribuição dos grupos construindo um histograma da contagem de exemplos agrupado em cada grupo. Para $k = 245$ a distribuição dos grupos é ilustrada na Figura 2.

Como podemos observar, o agrupamento utilizando *K-Means* não apresentou uma distribuição uniforme como esperado. Existe um grupo com um número muito maior de

Figure 2: Distribuição dos grupos para $k = 245$



exemplos comparado com os outros grupos. Isso foi observado para vários valores de k e diferentes tipos de inicialização.

A fim de entender esse efeito, foi realizado um experimento utilizando outra técnica de agrupamento que agrupa os dados hierarquicamente, o *Agglomerative Clustering*. Para esse experimento foi utilizado a distância euclidiana para computar as distâncias entre os dados, similar ao *K-Means*, e foi utilizado a média como critério de proximidade (*linkage criterion*). A Tabela IV ilustra os resultados obtidos com a execução desse experimento variando o valor de k de 240 a 250.

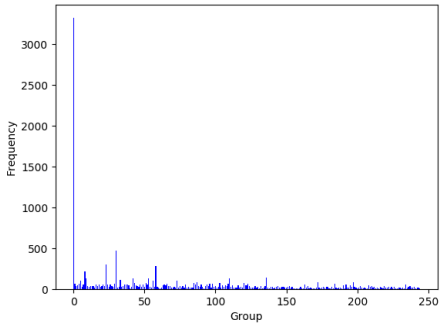
Table IV: Resultados da técnica *Agglomerative Clustering*

Número de grupos (k)	Silhueta	Davies-Bouldin	Calinski-Harabaz	Time (s)
240	0,027	3,579	10,385	60,371
241	0,027	3,573	10,379	60,451
242	0,027	3,565	10,366	60,542
243	0,028	3,564	10,395	60,267
244	0,028	3,560	10,402	60,500
245	0,028	3,552	10,390	60,414
246	0,028	3,552	10,379	61,061
247	0,028	3,544	10,389	59,849
248	0,028	3,538	10,372	60,327
249	0,028	3,535	10,360	60,237
250	0,028	3,527	10,351	60,257

Analisando os resultados podemos observar que os agrupamentos utilizando a técnica *Agglomerative Clustering* também não apresentaram boa qualidade para valores de k utilizados. Comparando os resultados da técnica com os resultados do algoritmo *K-Means*, ilustrados na Tabela III, podemos notar que o *K-Means* apresentou melhor qualidade nos agrupamen-

tos. Para a técnica *Agglomerative Clustering* os agrupamentos também não apresentaram uma distribuição uniforme dos grupos como ilustra a Figura 3 para $k = 245$. Através destas verificações, podemos concluir que esse comportamento não está associado a técnica de agrupamento ou parâmetros utilizados, mas sim aos dados utilizados.

Figure 3: Distribuição dos grupos para a técnica *Agglomerative Clustering* com $k = 245$



Por fim, com o objetivo de investigar os efeitos da aplicação de uma redução de dimensionalidade foi realizado um experimento considerando o melhor modelo de agrupamento encontrado, utilizando *K-Means* com $k = 245$, e foi aplicado a técnica PCA de redução de dimensionalidade mantendo diferentes valores de variação dos dados para observar o efeito dessa redução na qualidade e *performance* dos agrupamentos. Os resultados obtidos são ilustrados na Tabela V a seguir:

Table V: Resultados do PCA + *K-Means* para diferentes valores de variação dos dados

Comp.	Variação	Silhueta	Davies-Bouldin	Calinski-Harabaz	Time (s)
500	**0,7128**	**0,087**	**2,355**	**31,452**	**174,465**
600	0,7745	0,085	2,566	27,516	207,156
700	0,8283	0,085	2,702	25,545	227,065
800	0,8752	0,082	2,772	23,622	262,379
900	0,9162	0,079	2,830	21,969	274,719
1000	0,9537	0,077	2,988	20,792	295,713
1100	0,9813	0,073	3,054	19,580	334,494
1200	0,9997	0,076	3,045	19,665	365,65

Analisando os resultados obtidos podemos observar que a redução de 1203 dimensões dos dados originais para 500 dimensões mantendo cerca de 71% da variância dos dados apresentou melhor qualidade nos agrupamentos formados atingindo $0,087$ de Coeficiente de Silhueta, superando o valor atingido utilizando apenas o *K-Means*. Além disso, devido a redução de dimensionalidade o tempo de execução do algoritmo *K-Means* diminui de $352,148$ segundos para $174,465$ segundos, obtendo um *SpeedUp* de 2,01.

O fato da redução de dimensionalidade mesmo reduzindo a variância dos dados em 29% apresentar melhora na qualidade dos agrupamentos formados pode ser explicada pelo fato da redução ter removido possível ruído existente nos dados utilizados, formando grupos separados de maneira mais uniforme e coesa.

IV. Conclusões

Nesse trabalho foram exploradas técnicas de clusterização com o objetivo de agrupar e extrair informações de uma base de dados do *Twitter* sobre o tema saúde. Para isso foram utilizadas duas técnicas de clusterização, o *K-Means* e o *Agglomerative Clustering*. Diversos experimentos foram realizados para encontrar o melhor valor de k e o melhor modelo que agrupasse os dados avaliando os grupos formados visualmente pela análise dos *medoids* e seus vizinhos e através de métricas de qualidade como o Coeficiente de Silhueta. Além disso, foi avaliado o impacto da utilização de uma redução de dimensionalidade dos dados com a utilização do PCA.

Com base nos resultados dos experimentos realizados, concluímos que o algoritmo *K-Means* com valor de $k = 245$ apresentou melhores resultados embora as métricas de qualidade, como o Coeficiente de Silhueta, tenham apresentado valores baixos de qualidade. Porém, foi possível observar na análise dos *medoids* e seus vizinhos mais próximos que os grupos formados apresentaram certa semelhança e foram agrupados pela ocorrência de determinadas palavras. Foi verificado que essa abordagem tem suas desvantagens: Uma vez que sentenças de *Twitter* possuem um número pequeno de palavras, um modelo de agrupamento baseado em ocorrência de palavras torna-se muito sensível estatísticamente, pois determinada palavra não ocorrerá com tanta frequência em um *Tweet* como em um texto completo ou um artigo científico, por exemplo. Uma alternativa possível para esse caso seria o uso de um modelo *Bag of Words* binário ao invés de frequência. Outra melhoria que poderia ser incorporada ao modelo *Bag of Words* seria a construção de um vocabulário específico ao tema em questão, saúde, inibindo o agrupamento de sentenças por palavras comuns sem sentido semântico ao tema.

Por fim, foi possível concluir também que a utilização de uma redução da dimensionalidade com o PCA mantendo cerca de 71% da variância dos dados apresentou melhora na qualidade dos grupos formados e na performance da clusterização atingindo um *SpeedUp* de 2,01. Essa melhora na qualidade dos grupos deve-se ao fato da redução da dimensionalidade ter removido um possível ruído existente nos dados, aumentando a separação dos mesmos e melhorando a qualidade dos grupos formados.

References

[1] Health News in *Twitter* https://archive.ics.uci.edu/ml/datasets/Health+News+in+Twitter#

[2] Peter J. Rousseeuw (1987). "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". Computational and Applied Mathematics 20: 53–65.

[3] Davies, David L.; Bouldin, Donald W. (1979). "A Cluster Separation Measure" IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI-1 (2): 224-227.

[4] Caliński, T., & Harabasz, J. (1974). "A dendrite method for cluster analysis". Communications in Statistics-theory and Methods 3: 1-27.