Implementação do algoritmo de Clustering K-means aplicado em dados de câncer de mama em Wisconsin

Maria Eduarda Franklin da Costa de Paula, *Member, Instituto internacional de Neurociências Edmond e Lily Safra (IIN-ELS),*

***Abstract*— Este projeto tem por objetivo implementar um algoritmo de machine Learning (ML) não supervisionado, a qual não recebe nenhum rótulo, e com isso, tem que encontrar as classes que os dados pertecem, e o escolhido para essa aplicação é o K-means, o qual separa os dados e os agrupa em grupos de variância igual, e assim, minimizando a soma dos quadrados dentro de um cluster, ou o conhecido critério de inércia. E além disso, alia-lo à análise de componentes principais (PCA), qual é um procedimento matemático que cria novas métricas, as quais são combinações linaer das váriaveis originais, os quais serão aplicados na base de dados de Wisconsin sobre câncer de mama, disponibilizado pela sckit learn, para predizer a qual tipo do câncer cada dado pertence. O teste do algoritmo foi realizado com 2 centróides, o qual foi inicializado aleatoriamente, pelo parâmetro "Kmeans++" e também foi inicializado após ter sido aplicada a PCA, os quais apresentaram os valores de Silhoutte respectivamente 03.49,0.354, e 0.348 o que indica que a amostra está no ou muito próxima do limite de decisão entre dois clusters vizinhos. Além disso, apresentaram um tempo de processamento consideralvelmente baixo, e não apresentou um grande gasto computacional. Conclui-se que este algoritmo representa um boa solução para predizer o agrupamento dos dados, quando não se tem rotúlos e nem se quer sabe-se o número de grupos que existirão.**

***Index Terms*—Machine Learning, não supervisionada, Kmeans, análise de componentes principais (PCA).**

I. Introdução

A Teoria por trás do machine Learning (ML) surge em 1959 com a necessidade de criar sistemas mais inteligentes, isto é, capazes de aprender sem terem sido programados previamente. Entretanto, o marco inicial mais siginificativo foi o estudo de Samuel, que construiu um sistema que joga dama, o qual é capaz de aprender por generalização, e com a prática prolongada ele consegue sair do nível de iniciante para o de um jogador experiente (MICHIE,1968).

Com a avanço das tecnologias na atualidade o machine learning é uma ferramenta muito importante utilizada na ciência de dados para fazer predições e testar hipotéses (GUTIRREZ,2015). Esse invento pode ser aplicado de duas maneiras sendo eles o supervisionado que é a forma preditiva, e o não supervisionado, a qual diferente da outra maneira não recebe nenhum rótulo, e com isso, tem que encontrar as classes que os dados pertecem (GUTIRREZ,2015).

M. Franklin pertence ao Instituto internacional de Neurociências Edmond e Lily Safra (IIN-ELS), programa de pós-graduação em neuroengenharia, RN, 59280-000 BRA e-mail: (maria.paula@edu.isd.org.br).

Artigo enviado 21 de Setembro, 2020; Recebido em 22 de Setembro, 2020.

O algoritmo de ML não supervisionado se utiliza de um classificador, que avalia um conjunto de dados, e então particiona-os criando rótulos de classes (SANTOS,2020). Com isso, a necessidade de agrupar grandes quantidades foi se fazendo mais presente na ciência de dados, então, foi criado um dos algoritmos mais populares dessa aplicação que é o K-means, o qual separa os dados e agrupa-os em grupos de variância igual, e assim, minimizando a soma dos quadrados dentro de um cluster, ou o conhecido critério de inércia (GUTIRREZ,2015).

Munido dessas ferramentas este projeto tem por objetivo implementar o algoritmo Kmeans na base dados de Wisconsin sobre câncer de mama, disponibilizado pela sckit learn, para predizer a qual tipo do câncer cada dado pertence.

A. algoritmo K-means

O algoritmo Kmeans foi criado em 1967 por MacQueen, e atualmente é aplamente utilizado em muitos campos científicos devido ao seu algoritmo conciso e aplicação eficiente (SoOLTANI et al,2020). Além disso, ele visa agrupar os dados em grupos de variância igual, e minimizando o critério de inércia (Eq.1).Al(SoOLTANI et al,2020). Além disso, vale ressaltar que essa técnica minimiza a dissimilaridade intracluster e maximiza a dissimilaridade intercluster (SoOLTANI et al,2020). A.

Eq.1- Critério da inérca, $\sum_{k=1}^{N} \min_{\mu j \in C_j} (\|X_i - \mu_i\|^2)$

Para isso, o algoritmo primeiramente escolhe dentro de N amostras uma quantidade K de centróides, então ele calcula a distância Euclidiana (Eq.2) de cada ponto da amostra até o centróide, para escolher o mais proxímo (GUTIRREZ,2015). Ao fim desse processo é calculado a média do grupo e escolher um novo centróide, a partir desse ponto esse mecanismo é repetido até que o modelo convirja (GUTIRREZ,2015).

Eq.2-Distância Euclidiana, d=$\sqrt{\sum_{k=1}^{N}(X_{i,k} - P_{j,k})^2}$

Esse algoritmo também pode ser explicado por intermédio do diagrama de Voronoi, que tem seus grupos calculados a partir dos centróides inicialmente escolhidos, então é calculada uma nova média e escolhido um novo centróide (NICOLETTI,2017). Esse processo é repetido até uma condição de parada ser obedecida, que comumente é a diminuição relativa

na função objetivo entre as interações é menor que o valor de tolerância fornecido (NICOLETTI,2017).

É importante frisar que o número de clusters, isso é, de grupos pode ser escolhido de acordo com a quantidade de classes presentes no estudo. Entretanto, não é sempre que se tem essa informação e por isso foi criado o metódo do cotovelo que é capaz de calcular o número de Clusters (LEAL,et al,2017).

1) Método do cotovelo: O método do cotovelo, ou do inglês Elbow Method, é utilizado para encontrar a melhor divisão do grupo de amostras, ou seja, o número ideal de clusters (LEAL,et al,2017). Para isso essa técnica simula diversas divisões em número crescente de grupos e calcula as variâncias internas de cada grupo, buscando o ponto de equilílibrio (LEAL,et al,2017).

B. Análise de Componentes Principais

A análise de componentes principais (PCA), é um procedimento matemático que cria novas métricas, as quais são combinações linear das váriaveis originais (GUTIRREZ,2015). Dessa forma, é utilizado para a reduzir a dimensionalidade dos dados, sendo assim, sendo aplicado como uma etapa de pre-processamento antes de métodos de clusterização (LEAL,et al,2017). É muito utilizado aliado ao algoritmo K-means, pois, quando espaços de dimensões muito altas, as distâncias euclidianas tendem a se tornar infladas, então, é utilizado esse mecanismo para evitar essa situação.

II. Resultados

O algoritmo Kmeans foi aplicado na base dados de Wisconsin sobre câncer de mama, disponibilizado pela sckit learn, o qual possui quinhetos e sessenta e nove (569) dados e a quantidade de centróides ficou a critério do usuário. Vale ressaltar que o algoritmo da biblioteca sckit learn foi inicializado com o parâmetro ”Kmeans++” que tem oobjetivo de gerar os centróides mais afastados um do outro, o que gera resultados comprovadamente melhores. Mas, também foi inicializado aleatoriamente.

Além disso, para reduzir a dimensionalidade dos dados foi utilizada a PCA, e então criar resultados mais consistentes. Essa aplicação tinha o objetivo de predizer em que tipo de câncer um determinado dado se encaixa. O teste foi realizado com 2 centróides, o qual gerou o seguinte distribuição (Fig.1), vale especificar, que os pontos marcados por um X da cor branca representa os centróides.

Após a implementação do algoritmo, o qual foi inicializado de duas formas aleatório e ”Kmeans++” e o que foi aplicado o PCA, os quais apresentaram os valores de Silhoutte respectivamente 03.49,0.354, e 0.348 o que indica que a amostra está no ou muito próxima do limite de decisão entre dois clusters vizinhos. Além disso, os três apresentaram o mesmo valor para o critério de inércia, que foi 1595. Porém, quando inicializado aleatoriamente o algoritmo levou 0.05s para processar, 0.53s para o caso ”Kmeans” e 0.01 para o que foi aplicado o PCA.

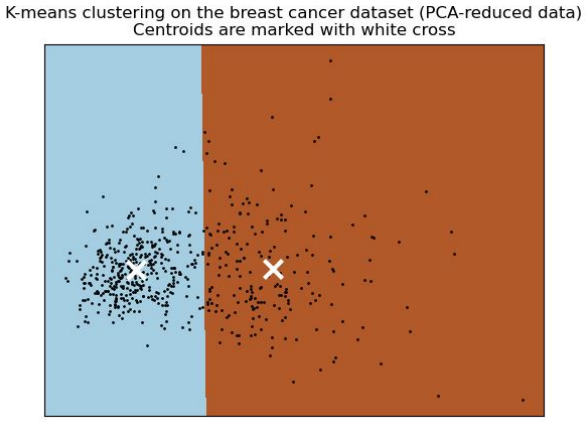

Fig. 1. Distribuição dos dados em 2 clusters.

III. Conclusão

A partir das análises pode-se concluir que este algoritmo representa um boa solução para predizer o agrupamento dos dados, quando não se tem rotúlos e nem se quer sabe-se o número de grupos que existirão. Podendo ser aplicada em várias, além disso, não é de díficil implementação e possui uma comunidade grande e muito ativa, também não tem um gasto computacional exorbitante. Ademais, pode ser aliada com outros mecanismos de Machine Learning o que a torna uma ferramenta muito importante. Com isso, futuramente esse mesmo algoritimo será aplicado a projeto alinhado neuroengenharia, especificamente, em uma base de dados de ângulos que um determinado segmento do corpo pode realizar. E assim, prevendo de qual intervalo o ângulo deve ser escolhido para realizar um eletroestimulação magnética, com intuito de evitar fadiga muscular os distenções tanto nas articulações quanto nos ligamentos, isto é, respeitando os limites dos usuários. Ademais, a escolha correta do ângulo gera movimentos mais precisos, aumenta a autonomia da bateria do sistema Internet Of Things (IoT).

References

[1] D. Michie, *”Memo” Functions and Machine Learning* , v.218. Nature, 1968.

[2] D. Gutierrez, *Machine Learning and Data Science: an introduction to statical learning methods with R* Basking Ridge, USA: Technics Publications, 2015.

[3] G. Santos, *Plataforma Para Segmentação de Clientes na Área Financeira* São Paulo, Brasil:Pós-Graduação em Data Science – Laureate International Universities - Faculdades Metropolitanas Unidas, 2020.

[4] S. E. Santos and S. . Aluisio and E. S. Rodrigues and J. M. M. Vieira and E. N. Texeira, *Metodos de Clusterização para a Criaçãode Corpus para Rastreamento Ocular durante a Leitura de Paragrafos em Português* São Paulo, Brasil:Repositório da Universidade de São Paulo, 2017.

[5] A. Soltani, and A.-Bermad, and H Boutaghane, and A. Oukiu, and O Abdalla, and M. Hasbaia, and R Oulebsir, and S. Zeroual, and A. Lefkir *Uma abordagem integrada para avaliar a qualidade das águas superficiais: Caso da barragem Beni Haroun (Nordeste da Argélia)* Argélia: Springer.Environ Monit Assess 192, 630 (2020).