Algoritmos de Clusterização

FILGUEIRAS, Lucas Parreira; **COSTA**, Ramon Elias; **AMARAL**,
Victor Paiva da Silva; **VERNECK**, Felipe Pereira; **BAIA**, Joas Weslei

CENTRO UNIVERSITÁRIO GOVERNADOR OZANAM COELHO.

Ciência da Computação

UNIFAGOC
CENTRO UNIVERSITÁRIO GOVERNADOR OZANAM COELHO

INTRODUÇÃO

A clusterização é uma tecnologia usada no nosso dia a dia. Podemos encontra-lá em sistemas meteorológicos, em aplicativos de mapeamento genético e até em programas de renderização de imagens, no geral é uma ótima ferramenta para agrupar dados, seja de um cliente, um computador, ou o que for necessário para encontrar semelhanças entre esses dados, e assim trazer uma maior eficiência e satisfação caso a clusterização tenha sido utilizada para um cliente específico. Para realizar esse processo são necessários algoritmos que necessitam de fórmulas, e dependendo dos dados que você deseja trabalhar um algoritmo pode ser mais eficiente que o outro.
A clusterização é uma ferramenta que facilita a vida do usuário, não importa a ocasião, podendo até ser usada em âmbito empresarial, afinal, Michael Porter, criador do conceito “cluster industrial” dizia: “O impacto da tecnologia da informação é tão difuso que os executivos se defrontam com um problema difícil: excesso de informação.”, Porter diz nessa citação que nos dias atuais há excesso de informações, ou seja, de dados, porém a clusterização existe para facilitar este processo organizando estes dados.

OBJETIVO

O objetivo da clusterização é muito diverso, podemos usa-lá no meio empresarial, educacional, artístico, e em diversas outras áreas. Mas essencialmente ela é usada para agrupar dados, não importa qual seja a origem deste dado, por exemplo, vamos supor que você queira saber qual o gosto musical predominante dos alunos de uma escola, primeiramente deverá ser coletado os dados dos alunos, e assim ser criado um cluster dividindo entre todos os ritmos musicais escolhidos e ver qual a maior quantidade de alunos que gostam de um certo rítmo de música, dependendo, você poderá exibir graficamente este resultado, vendo de maneira mais clara qual o ritmo predominante dos alunos desta escola.

Uma pergunta muito frequente sobre este assunto é: “Tudo pode ser clusterizado?” A resposta para esta pergunta é simples, independente do tipo de dado, se ele for quantificável, sim, ele pode ser clusterizado. Afinal o objetivo do cluster é facilitar a realização das estratégias, seja da sua empresa ou uma situação que você está passando, agrupando dados e fazendo uma análise detalhada.

Neste projeto foi criado um cluster com o objetivo de analisar dados e mostra-los graficamente, usando vários tipos de dados reais como teste. O gráfico demonstra, em um plano cartesiano, os dados agrupados considerando o seu grau de semelhança, dividindo por cores para facilitar a visualização. Porém, um objetivo “secundário” era também mostrar a diferença de resultados entre diferentes algoritmos de clusterização, como o k-means e dbscan que foram usados neste projeto.

MATERIAIS E MÉTODOS

Os materiais usados neste projeto foi: a IDE PyCharm, usada para construir programas na linguagem Python. A plataforma GitHub para construir um repositório e compartilhar entre os demais membros do grupo. Foi também usado o programa Microsoft Visio para a construção do Canvas para ajudar no desenvolvimento do projeto, e sites e artigos sobre clusterização para o aprendizado dos integrantes sobre o assunto tratado.

Os métodos para a construção deste cluster foram os algoritmos k-means e dbscan, juntamente com uma extensão visual destes resultados graficamente em um plano cartesiano. A utilização da linguagem Python foi escolhida pois se mostrou a melhor para construção de programas de clusterização.

RESULTADOS

Observando os resultados dos algoritmos usados neste projeto podemos extrair algumas informações. No algoritmo k-means a visualização gráfica ficou melhor devido aos centroides, eles possibilitam que o usuário enxergue de forma mais clara a divisão dos dados através de suas semelhanças. É um ótimo algoritmo para o uso de mercado em geral, como por exemplo, os perfis de clientes em servidores de streaming como a NetFlix.

O outro algoritmo usado foi o dbscan. Neste algoritmo, a facilidade de visualização gráfica é inferior ao k-means, pois ele não possui centroides. Ele divide os grupos através de uma certa distância entre cada cluster, portanto é muito utilizado e eficaz em divisões hierárquicas, um bom exemplo é um gráfico da relação da idade da pessoa e o salário que ela recebe, neste suposto caso o algoritmo dbsan traz melhores resultados.

CONCLUSÃO

Dentre todos os tópicos discutidos, podemos observar que a clusterização é uma tecnologia extremamente útil para resolver problemas de diversos ambiente em que você a use. Existem vários algoritmos de agrupamento, cada um tendo o seu ponto forte e podendo ser usado da melhor forma dependendo do seu objetivo. Os resultados do uso desta ferramenta traz vantagens competitivas quando se trata de análise e agrupamento de dados.

REFERÊNCIAS

PORTER, Michael. **Estratégia Competitiva**. [s.l]. Campus, 1980.

MEDIA, Dev. **Data Mining na Prática: Algoritmo K-Means**. [s.l.]. DevMedia. 2007. Disponível em: https://www.devmedia.com.br/data-mining-na-pratica-algoritmo-k-means/4584.