Redução da dimensionalidade em bigdata

Lucca de Castro Machado

Prof. Dr. Anderson Borba

Motivação e Objetivo

Com o aumento constante da quantidade de dados gerados em diversas áreas mundialmente (Figura 1), a análise desses dados tornou-se um desafio cada vez maior, pois com mais dados gerados mais processamento será preciso para fazer a análise, tornando o tempo de processamento alto impactando entregas de projetos. A partir desse fato, motiva-se a redução da dimensionalidade em big data para lidar com esse desafio, permitindo a análise de grandes quantidades de dados de forma mais eficiente e precisa.

O objetivo deste projeto é explorar as técnicas de redução de dimensionalidade em big data, destacando suas vantagens e desvantagens, e mostrar na prática a diferença no tempo de processado e na taxa de acerto nas previsões, comparando entre uma mesma base de dados com a dimensionalidade reduzida e a outra normal. Assim podendo provar como elas podem ser aplicadas na prática para lidar com grandes conjuntos de dados.

Metodologia

O projeto busca entender a funcionalidade entre 3 algoritmos de redução de dimensionalidade, o Análise de Componentes Principais (PCA), Kernel PCA e o Análise Discriminante Linear (LDA). Nesta primeira etapa do projeto, os 3 algoritmos foram aplicados em uma base de dados classificação, onde contém 14 variáveis independente e 1 variável dependente, as 14 variáveis vão influenciar a previsão se um indivíduo ganha acima ou abaixo de 50 mil dólares no ano na nossa base de dados.

O objetivo da pesquisa é reduzir a dimensionalidade das variáveis independentes com PCA, Kernel PCA e LDA e fazer a comparação entre os algoritmos e uma base sem redução de dimensionalidade, a comparação ser a partir da precisão preditiva utilizando o Random Forest e o tempo de processamento utilizando a biblioteca time, para o resultado ficar mais assertivo, vão ser rodado 10 vezes a mesma rotina de código, e a cada vez rodada vão ser chamada de teste e em seguida o número ordinal de teste, como por exemplo teste 1, teste 2 e assim por diante, com os resultados de tempo obtido, calcular a média e a mediana.

A pesquisa foi realizada em um computador com processador AMD ryzen 7 5700g, 32gb de RAM, placa mãe B550M-Plus, fonte de 650W, SSD 512gb de Leitura 3100MBs e Gravação 1500MBs, placa de vídeo RTX 3070 Ti. O software utilizado foi o Jupyter Notebook 6.5.2 e Python 3.10.2 no sistema operacional Windows 11 Pro.

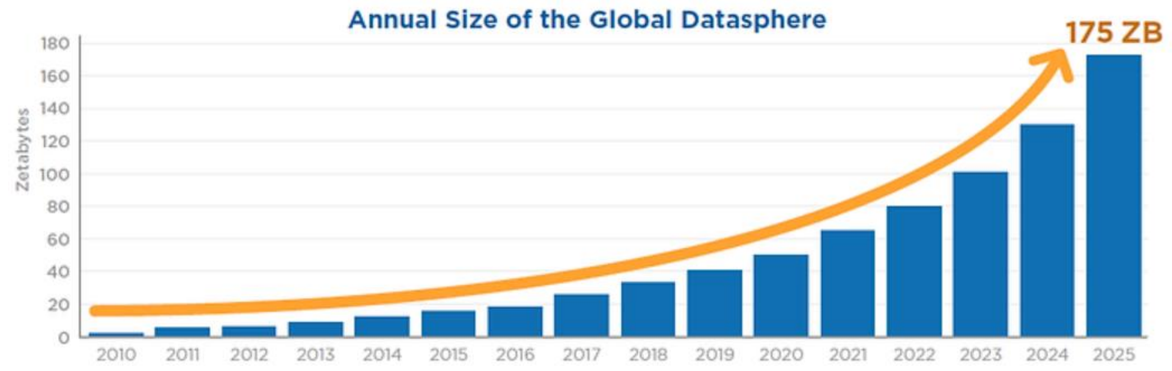

Figura 1 – Tamanho anual da esfera de dados global *(Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018)*

Resultados Preliminares

Com base na metodologia apresentada foram obtidos os seguintes resultados:

Precisão preditiva:

	Base sem Redução	PCA	Kernel PCA	LDA
Precisão	84.33%	83.73%	82.35%	73.34%

Tempo de processamento:

	Base sem redução	PCA	Kernel PCA	LDA
Média	0, 08095	0,09185	0,08712	0,09148
Mediana	0,0755	0,0915	0,0912	0,09805

Analisados os resultados obtidos, podemos observar que com a utilização de algoritmos de redução de dimensionalidade, pode ocasionar queda na precisão preditiva e um acréscimo de tempo de processamento em relação a uma base sem redução.

Conclusões e próximos passos

Concluindo-se, em relação a pesquisa realizada, base de dados de classificação não são muito eficaz utilizando os algoritmos de redução de dimensionalidade PCA, Kernel PCA e LDA.

Os próximos passos são analisar outras base de dados com características diferentes, como uma base de dados de regressão, uma base de dados com mais variáveis independentes e uma base de dados com maior número de registros. E usar outros métodos de redução de dimensionalidade e comparar com as utilizada na pesquisa inicial.

Referências

REINSEL, David; GANTZ, John; RYDNING, John. The Digitization of the World From Edge to Core. Framingham: Idc, 2018. 28 slides.