Métodos Probabilísticos para Engenharia Informática

Docentes: Prof. António Teixeira e Prof. Carlos Bastos

Aplicação de processamento do valor da qualidade do ar

Luís Sousa 108583

Simão Almeida 113085

DETI

Universidade de Aveiro

12-2024

Introdução

Este projeto foi motivado pela necessidade de desenvolver e testar uma aplicação que detete itens similares, que detete a pertença a um conjunto e que utilize um classificador de Naïve Bayes em *MATLAB*. Com base nesta necessidade, tomamos a decisão de desenvolver um programa que processe e avalie os valores da qualidade do ar com base nos níveis de CO2.

Para desenvolver este projeto, foram utilizados os conhecimentos obtidos durante todo o semestre em Métodos Probabilísticos para Engenharia Informática.

Objectivo

A qualidade do ar é um dos principais indicadores de saúde ambiental, desempenhando um papel crucial no bem-estar humano, na preservação dos ecossistemas e no desenvolvimento sustentável. Com o aumento da urbanização, das atividades industriais e das emissões de poluentes, monitorar e avaliar a qualidade do ar tornou-se uma prioridade global.

Até em ambientes escolares é importante ter uma boa qualidade do ar para garantir a saúde, o bem-estar e o desempenho acadêmico de alunos, professores e funcionários. Como crianças e jovens passam grande parte de seu tempo em sala de aula, um ar limpo e saudável é crucial para promover um ambiente propício ao aprendizado e ao desenvolvimento integral.

O programa desenvolvido visa utilizar os três módulos desenvolvidos (MinHash, Bloom Filter, Naïve Bayes) para avaliar e classificar a qualidade do ar de acordo com os niveis de CO2.

Dados utilizados

O conjunto de dados utilizados foi obtido através da página web de um projeto que mede a qualidade do ar num espaço interior, Dataset.csv . Este conjunto tem vários dados, sendo que foram utilizados a data e a hora e a concentração média de CO2 nos últimos 15 minutos.

Bloom Filter

O **Bloom Filter** é uma ferramenta probabilística usada para verificar se um elemento pertence a um conjunto. Ele oferece uma abordagem de alta performance em termos de uso de memória e tempo, mas com uma característica peculiar: pode ocorrer falsos positivos (indicando que um elemento pertence ao conjunto quando na verdade não pertence), embora nunca ocorram falsos negativos.

Neste projeto, o programa vai adicionar elementos ao Bloom Filter e vai verificar se esse elemento já existia anteriormente.

MinHash

O **MinHash** é uma técnica probabilística utilizada para estimar de forma eficiente a similaridade de conjuntos. Ele é amplamente usado em aplicações onde é necessário lidar com grandes volumes de dados. O MinHash é baseado na **similaridade de Jaccard**, uma métrica que mede a similaridade entre dois conjuntos através da razão entre o tamanho da interseção e o tamanho da união dos conjuntos. Neste caso, vai criar uma assinatura MinHash para cada conjunto e vai comparar as assinaturas.

Naïve Bayes

O classificador **Naïve Bayes** é um algoritmo de aprendizado de máquina baseado no teorema de Bayes, que usa probabilidades para classificar dados. Ele assume que as características de entrada são independentes entre si, o que simplifica os cálculos e torna o modelo eficiente, mesmo em grandes volumes de dados. Neste projeto, vai classificar a qualidade do ar em 3 categorias de acordo com os volumes de CO2.

Testes

Para verificar o funcionamento das nossas funções de **Bloom Filter**, de **MinHash** e de **Naive Bayes**, foram realizados alguns testes.

No teste do **Bloom Filter**, configuramos um filtro que utiliza 3 funções de hash. Inicialmente, ele está vazio e é preenchido com os cinco primeiros valores de uma lista de dados. Em seguida, verifica-se se alguns valores pertencem ao conjunto. Se o filtro indicar que o valor está presente, a mensagem "in the set" será exibida; caso contrário, a mensagem será "not in the set". É importante notar que o Bloom Filter pode gerar falsos positivos, mas nunca falsos negativos.

Para o teste do **MinHash** são criados dois subconjuntos. Em seguida, as assinaturas de MinHash para cada subconjunto são geradas com base em um número predefinido de hashes. Por fim, a similaridade entre as assinaturas dos subconjuntos é calculada e exibida.

No teste para a função do classificador de **Naive Bayes**, classificamos o valor da concentração de CO2 em diferentes categorias. Valores menores ou iguais a 600ppm são classificados como "Great", entre 601 e 800ppm como "Good", e acima de 800ppm como "Bad".

Resultados

O modelo Naive Bayes foi inicialmente aplicado para a tarefa de classificação com base em dados de entrada relacionados à qualidade do ar. No entanto, a limitação do Naive Bayes é evidente em cenários com dados altamente correlacionados ou quando o volume de dados aumenta consideravelmente, havendo vários casos falso positivos e falso negativos.

A adição do Bloom Filter ao modelo Naive Bayes, em conjunto com o MinHash, trouxe varios benefícios, principalmente em relação ao uso eficiente da memória e à redução de falsos positivos durante a classificação. O MinHash, que utiliza uma abordagem probabilística para aproximar a similaridade de conjuntos, foi integrado para melhorar a performance na análise de grandes volumes de dados. O Bloom Filter, que permite testar rapidamente a existência de um item em um conjunto sem armazenar explicitamente os dados, ajudou a reduzir o custo computacional de operações repetitivas, como verificações de existência em grandes volumes de dados.

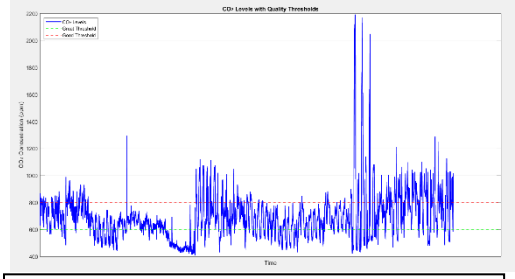

Figura 1: Grafico dos niveis de CO2 sem funções de MinHash e Bloom Filter

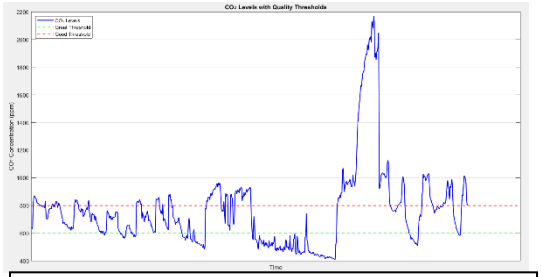

Figura 2: Grafico dos niveis de CO2 com funções de MinHash e Bloom Filter

Conclusão

O projeto alcançou com sucesso os objetivos propostos, integrando as técnicas de MinHash, Bloom Filter e Naïve Bayes em uma aplicação funcional para análise da qualidade do ar com base nos níveis de CO2. Cada módulo contribuiu de forma significativa: o MinHash demonstrou eficiência na análise de similaridade entre conjuntos, o Bloom Filter mostrou-se eficaz na verificação rápida de pertinência a conjuntos, e o Naïve Bayes ofereceu uma abordagem probabilística para classificação dos dados.

A integração dessas ferramentas possibilitou o desenvolvimento de um sistema capaz de lidar com grandes volumes de dados de forma eficiente, mantendo um bom equilíbrio entre precisão, desempenho e uso de recursos. A aplicação demonstrou ser adequada para monitorar e classificar a qualidade do ar, destacando a relevância de utilizar técnicas probabilísticas em cenários práticos que exigem soluções escaláveis e rápidas.