Calibração de Distância em Métodos de Acesso Métrico por meio de Realimentação de Relevância

Renato Gomes Marcacini[1], **Willian Dener de Oliveira**[1], **Agma Juci Machado Traina**[1]

[1]Instituo de Ciências Matemáticas e de Computação – Universidade São Paulo (USP)
Caixa Postal 668 – 13.560-970 – São Carlos – SP – Brazil

renato.gomes.marcacini@usp.br, {willian, agma}@icmc.usp.br

Abstract. *Traditionally Metric Access Methods (MAM) use fixed distance functions to build the metric trees, which in turn prevents a MAM from being able to index elements using two or more distance functions in the same index. A vector of weights correctly learned by Relevance Feedback (RR), allows weighting distance functions and improving data semantics, improving the query accuracy. This work introduces the Tuning Metrics Relevance Feedback (TMRF) method, which shows that when using weighted distance functions in MAMs, the retrieval efficiency of these structures becomes up to 70% more efficient than sequential strategies, in addition to obtaining a gain of 42% through RR learning.*

Resumo. *Usualmente, os Métodos de Acesso Métrico (MAM) utilizam funções de distância fixas para realizar a construção da árvore métrica, o que por sua vez impede que um MAM consiga indexar os elementos utilizando duas ou mais funções de distância na mesma indexação. Um vetor de pesos corretamente aprendido por Realimentação de Relevância (RR), permite ponderar funções de distâncias e aprimorar a semântica dos dados, trazendo maior precisão ao processamento de consultas. Este trabalho apresenta a abordagem denominada Tuning Metrics Relevance Feedback (TMRF), que incluem funções de distâncias ponderadas no MAM Slim-Tree, mantendo-a eficiente, sendo 70% mais rápida em relação a estratégias sequenciais, com ganhos expressivos em termos de acurácia de até 42% através de aprendizado por RR.*

1. Introdução

Os Métodos de Acesso Métrico (MAM) foram desenvolvidos para otimizar o processamento de consultas por similaridade a dados complexos, principalmente para grandes conjuntos de dados. Por exemplo, em uma aplicação de imagens médicas, a agilidade no processamento das mesmas é essencial, já que é necessário analisar uma grande quantidade de imagens em tempo hábil. Com a ajuda de um MAM, é possível comparar rapidamente as características de uma base de dados de imagens médicas, permitindo que o sistema ou especialista identifique rapidamente possíveis diagnósticos e tratamentos previamente armazenados, proporcionando uma referência para diagnósticos atuais. Ou seja, um MAM acelera o processo de análise trazendo suporte à decisão clínica, e fornecendo *insights* valiosos para o tratamento do paciente. Os MAMs envolvem a definição de uma métrica de similaridade para construção da indexação, a qual é utilizada para determinar quais elementos estão mais próximos de um determinado elemento de consulta, o que por sua vez impede que, ordinariamente, um MAM consiga indexar os elementos utilizando duas ou mais funções de distância em uma mesma indexação [Silva 2009, Li et al. 2018].

O uso de um MAM que permita a recuperação otimizada de dados complexos que contemple a semântica de similaridade real dos dados em um espaço métrico e que possa adequar a função de comparação (distância) em tempo real, são fundamentais para a aceitação de sistemas de recuperação de imagens baseada em conteúdo (*content-based image retrieval* - CBIR) pelos especialistas. Portanto, soluções baseadas na inclusão de funções de distância ponderadas no MAM juntamente com técnicas de Realimentação de Relevância (RR), que aprendem pesos para ajustar medidas de similaridade, são essenciais para obter consultas de similaridade e tempos de processamentos mais eficazes pelos MAMs em relação às estratégias de recuperação por matriz de distância. Este trabalho tem como objetivo mostrar que funções de distância ponderadas podem ser implementadas em MAMs, especificamente no MAM *Slim-Tree*, que é uma estrutura de árvore otimizada para consultas em bancos de dados com alta dimensionalidade. A proposta é melhorar os resultados das consultas por similaridade sem afetar o tempo de recuperação dos elementos a longo prazo.

Muitas das pesquisas da literatura ressaltam e demonstram a eficácia das técnicas de RR para aumentar o desempenho geral das consultas aumentando a semântica dos resultados [Kim and Chung 2003, Mohanan and Raju 2017, Tian 2018, Ahmed 2020]. Porém, as formas de recuperação de dados apresentados na literatura ocorrem através de cálculos entre pares de elementos utilizando matrizes de distâncias, que são estratégias menos otimizadas para lidar com a recuperação de grandes volumes de dados complexos.

Por meio de experimentos com RR que aprendem vetores de pesos, neste artigo é apresentada uma nova abordagem denominada *Tuning Metrics Relevance Feedback* (TMRF) que traz uma metodologia de inclusão de funções de distância ponderadas no MAM *Slim-Tree*, na qual os vetores de pesos são aprendidos por uma RR revisada de [Rui et al. 1998], que utiliza o inverso do desvio padrão para ressaltar os atributos relevantes, que são então introduzidos como ponderação da medida de similaridade no MAM de forma dinâmica. Além disso, a abordagem define uma metodologia de reindexação do MAM *Slim-Tree* para manter a otimização da estrutura a longo prazo. A inserção de funções ponderadas no MAM *Slim-Tree* inclui uma melhoria na eficiência do tempo de processamento em comparação com as recuperações que utilizam matrizes de distâncias, sendo até 70% mais rápido na recuperação de até 10% da base de imagens. Além disso, a abordagem consegue obter altos níveis de precisão com ganhos de até 42% após o aprendizado por RR.

2. Fundamentação Teórica

2.1. Funções de Distâncias Fixas e Ponderadas

Os sistemas CBIR demandam mecanismos para responder consultas por similaridade e retornar elementos que mais se assemelham a um elemento de consulta [Guo et al. 2020]. As funções de distância são utilizadas para medir a dissimilaridade (ou similaridade) entre dois elementos de um conjunto, ou seja, quão menor for a distância entre dois elementos mais similares eles são. Os cálculos são realizados em um espaço métrico onde o conceito de distância entre elementos pode ser obtido. Há uma grande variedade de funções de distância fornecidas na literatura, as mais utilizadas são as funções da família Minkowski (L_p), entre elas a distância Manhattan (L_1), Euclidiana (L_2) e Chebychev (L_∞), essas funções são representadas pela Equação 1, onde x_i e y_i são os elementos complexos a

serem comparados pela distância d.

$$d((x_1, ..., x_n), (y_1, ..., y_n)) = \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p} \quad (1)$$

As funções de distância ponderada são uma ferramenta poderosa no aprendizado de métricas. Este tipo de métrica de similaridade dinâmica permite calcular a distância entre dois elementos, levando em consideração a importância de cada atributo, ou seja, são atribuídos diferentes pesos às diferentes dimensões dos dados. A ponderação permite ao usuário adaptar a função de distância às necessidades e preferências específicas de uma determinada tarefa, atribuindo pesos maiores às dimensões mais importantes, fazendo com que melhore os resultados de uma consulta [Tian 2018]. As funções de distâncias ponderadas de Minkowski são representadas pela Equação 2, onde w_i é o peso aplicado sobre a dimensão, considerando-se que a soma de todos os pesos deve sempre manter o valor igual a um, ou seja, $\sum wi = 1$.

$$d((x_1, ..., x_n), (y_1, ..., y_n)) = \sqrt[p]{\sum_{i=1}^{n} w_i |x_i - y_i|^p} \quad (2)$$

Ao utilizar uma função de distância ponderada, podemos dar mais peso às dimensões mais importantes, alterando o formato de abrangência da região de busca, desta forma melhorando a precisão de uma consulta por similaridade. A Figura 1 apresenta o efeito de ponderação de uma função de distância euclidiana por pesos, "esticando" o formato de busca passando a incluir elementos relevantes no raio de busca de uma consulta por similaridade.

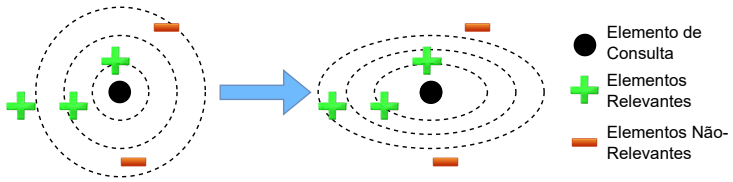

Figura 1. Ajuste de uma consulta sob uma função de distância euclidiana

2.2. Métodos de Acesso Métrico

Os Métodos de Acesso Métrico (MAM) usam funções de distância para organizar os elementos em um banco de dados, e para isso, o MAM assume que os elementos estão em um espaço métrico e usa suas relações de similaridade para organizá-los. Com foco na velocidade de processamento das consultas, o MAM busca reduzir o número de cálculos de distância e acessos a disco necessários para a execução das consultas. Uma medida de similaridade é usada para quantificar o quão similares são dois elementos, permitindo que as consultas sejam expressas com base na similaridade.

Dados complexos como imagens são indexados e recuperados com base em seus descritores. Um descritor consiste em um par de <*vetor de características* e *função de distância*>, que deve representar o conteúdo de dados complexos, permitindo fazer a comparação mais adequada para um determinado contexto [Tyagi 2018]. Usualmente os MAMs não podem usar duas ou mais funções de distância ao mesmo tempo, portanto a

escolha das funções de distância é extremamente importante para organizar os elementos. Ao longo dos anos, os MAMs evoluíram para lidar com grandes dimensões e serem estruturas dinâmicas, ou seja, permitir que elementos sejam inseridos e removidos após a construção da estrutura, característica fundamental utilizada em sistemas gerenciadores de banco de dados.

2.3. Slim-Tree

A *Slim-Tree* é um MAM projetado para minimizar os acessos a disco durante a recuperação de dados complexos, como imagens. A *Slim-Tree* visa melhorar a eficiência e a precisão da recuperação dos dados usando uma estrutura de árvore balanceada para indexá-los. A estrutura deste MAM permite encontrar rapidamente imagens semelhantes à imagem de consulta, economizando tempo e esforço dos usuários na análise dos resultados. O algoritmo *Slim-Down* da *Slim-Tree* oferece uma vantagem significativa ao reduzir a complexidade da estrutura de indexação de dados, otimizando o processo de busca e organização em espaços métricos. Este MAM pode ser utilizado para suportar vários tipos de consultas, sendo as principais as consultas por similaridade: consultas por abrangência e aos vizinhos mais próximos (*k-Nearest Neighbors* - kNN).

A *Slim-Tree* é um MAM que utiliza funções de distâncias fixas na construção da árvore. Isso significa que a escolha da função de distância usada para calcular a similaridade entre os elementos não muda durante a construção da árvore ou após a inserção de novos elementos. Isto é, as distâncias são pré calculadas para realizar a indexação, e se a função de distância escolhida não for adequada para o tipo de consulta que está sendo realizada, a *Slim-Tree* pode não ser capaz de encontrar os resultados desejados com eficácia. Este trabalho lida justamente com o aspecto inovador de alteração da função de distância que passa a ser incluído em um MAM, e a *Slim-Tree* foi usada como plataforma de trabalho para validar a abordagem proposta, sem perda de generalidade.

2.4. Realimentação de Relevância

A Realimentação de Relevância (RR) é uma técnica usada em sistemas CBIR para melhorar a precisão dos resultados da pesquisa. O objetivo da RR é aprender um conjunto de pesos que podem ser aplicados às características das imagens no banco de dados para melhor corresponder aos critérios de pesquisa do usuário, podendo incluir estratégias que ponderam a medida de similaridade, considerando indicações da semântica dos dados, conforme a visão do usuário. Essa abordagem permite ao usuário, especialista no domínio de dados, ajustar a medida de similaridade, modificando a região de busca no espaço métrico. Em geral, um algoritmo de uma RR atribui diferentes níveis de relevância aos atributos presentes no vetor de características do conjunto de dados. Esses atributos representam a semântica dos dados complexos, em que cada atributo é um valor numérico que representa uma característica específica. Quando uma abordagem demonstra alta eficácia semântica, significa que ela é capaz de fornecer respostas precisas e relevantes, levando em consideração o contexto e a intenção do usuário.

Técnicas que visam alterar a forma da medida de similaridade devem fornecer pesos para cada dimensão no espaço de função L_p. No trabalho de [Rui et al. 1998], a utilização de amostras relevantes para RR demonstram um alto aumento no primeiro ciclo de realimentação em relação a outras abordagens, na qual são utilizados métodos estatísticos como desvio padrão e variância inversa para aprender os pesos com base no

julgamento do usuário. A Figura 2 exemplifica os passos do aprendizado de RR. No processo, o usuário envia uma imagem de consulta e recebe de volta as *top k* imagens mais similares. O usuário então seleciona as imagens relevantes dentre as *top k* de acordo com a imagem de consulta. Essas imagens selecionadas trazem *feedback* positivo e negativo, e tal informação é passada para o algoritmo de RR, cujo o resultado é o *top k* imagens de maior precisão.

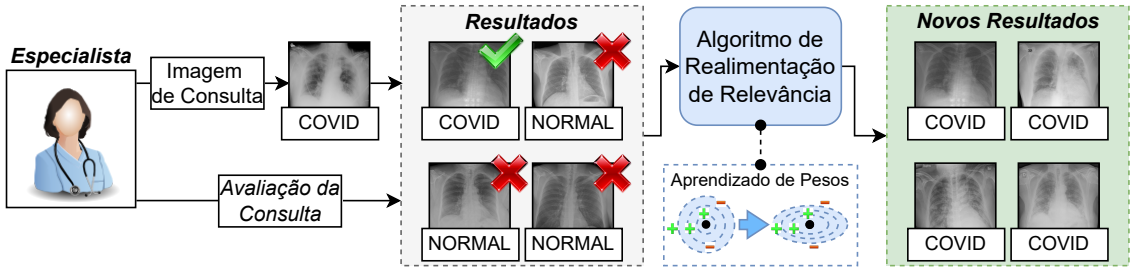

Figura 2. Processo de realimentação de relevância com pesos na função de distância, em um ciclo.

3. Trabalhos Relacionados

A precisão e o tempo de processamento de consultas por similaridade são importantes para a tomada de decisão por especialistas de um domínio. Os trabalhos da literatura detalham algumas técnicas para ampliar a precisão de consultas por similaridade. No trabalho de [Rui et al. 1998], os autores definiram uma heurística com atualização de parâmetros baseada em otimização da matriz de distâncias, e descobriram que o método baseado em otimização com elementos relevantes alcança maior precisão. Em [Chang et al. 2009], foi desenvolvida uma heurística que deforma o espaço métrico afastando elementos irrelevantes de relevantes percorrendo toda a base de elementos. Em [Bressan et al. 2019], a utilização de *Support Vector Machine Active* com RR define um hiperplano separador entre os elementos rotulados como relevantes em imagens de mamografias. No trabalho de [Kumaran et al. 2021], o uso de aprendizado de métrica consistiu em utilizar a aprendizagem ativa para selecionar exemplos relevantes para atualizar a matriz de Mahalanobis durante o treinamento e calcular a distância entre as amostras para classificação.

Segundo a literatura anteriormente listada, as estratégias apoiam-se no desenvolvimento de técnicas eficientes para aumentar a precisão de consultas a fim de evitar falsos positivos. Entretanto, a maioria dessas estratégias demandam alto custo computacional por não fazerem uso de estruturas de indexação que otimizam a recuperação de dados complexos. Com foco neste cenário, este artigo apresenta uma metodologia que inclui a técnica de RR para ponderar a função de distância no MAM Slim-Tree, analisando a eficiência da RR em melhorar a precisão da consulta e também o custo pago relativo ao tempo de processamento de MAMs ponderados em comparação à recuperação sequencial. A Tabela 1 apresenta as principais abordagens de RR que modificam a métrica de similaridade aprimorando o espaço métrico e as características. Nas características abordadas, a **“Métrica de Similaridade Ponderada"** indica se os trabalhos utilizam da abordagem de refinamento da medida de similaridade. A **“Melhora Semântica"** indica se os trabalhos têm sucesso em aprimorar a acurácia de conjuntos de dados. Em **“Utiliza poucas amostras"** indica se o aprendizado da técnica é eficaz com poucos dados de *feedback* ou

amostras. Em **"Utiliza MAM"** indica se os trabalhos utilizam estruturas de índices para otimizar a consultas por similaridade.

Tabela 1. Revisão da Literatura de abordagens de realimentação de relevância que aprendem métricas de similaridade em matriz de distâncias

Literatura	Métrica de Similaridade Ponderada	Melhora a Semântica	Utiliza Poucas Amostras	Utiliza MAM
[Rui et al. 1998]	✓	✓	✓	✗
[Chang et al. 2009]	✓	✓	✗	✗
[Bressan et al. 2019]	✓	✓	✗	✗
[Kumaran et al. 2021]	✓	✓	✗	✗
Abordagem TMRF	✓	✓	✓	✓

4. Abordagem Proposta: *Tuning Metrics Relevance Feedback*

Este trabalho propõe a *Tuning Metrics Relevance Feedback* (TMRF), que inclui a técnica *RR Desvio Padrão* aqui proposta, que aprende vetores de pesos para ponderar métricas de similaridade diretamente no MAM *Slim-Tree*. Nas abordagens convencionais, o processo de RR pondera toda a matriz de características, tanto do *feedback* informado pelo usuário, quanto a matriz de características da base de dados, gerando um "gargalo" no tempo de processamento das operações de consultas em dados complexos.

Nosso objetivo foi desenvolver uma metodologia que inclua funções de distâncias ponderadas no MAM e que consiga ponderar a métrica de similaridade de forma dinâmica, ou seja, que a função de distância possa ser modificada durante os ciclos de RR realizados pelos usuários, além de prover a reindexação da estrutura de MAMs quando necessário. A finalidade é de utilizar a eficiência do tempo de processamento de recuperação deste tipo de estrutura e o aprimoramento do espaço métrico através da RR Desvio Padrão. A Figura 3 ilustra o processo de aprendizado do RR para retornar uma consulta por similaridade, apresentando o "gargalo" que atua nas operações de matrizes de distância nas abordagens convencionais, e a nossa abordagem que acopla o MAM para lidar com o processamento de modo mais eficaz em adição com uma metodologia de reindexação para manter a otimização da estrutura.

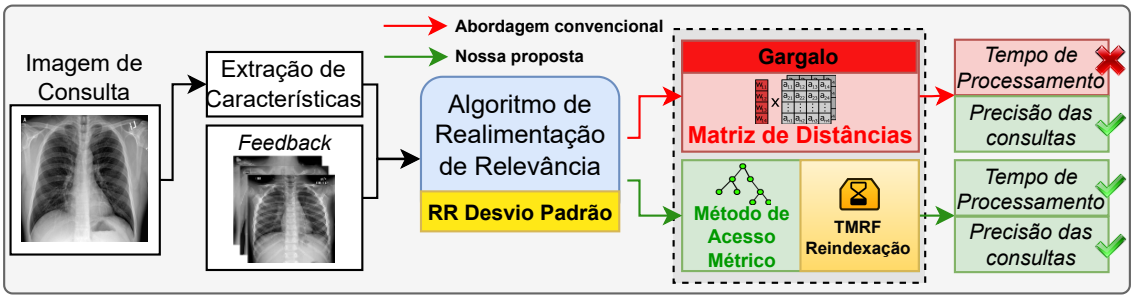

Figura 3. Abordagem proposta de alteração da distância por ponderação.

4.1. Aquisição da Base de Dados

A análise aqui apresentada utiliza três bases representativas de imagens públicas. A primeira base é a vencedora do *COVID-19 Dataset Award da Kaggle Community*, que é cons-

tituída de imagens de Raio-X pulmonar separadas em três classes, imagens de COVID-19, pneumonia e normais[1], com um total de 3000 imagens separando 1000 imagens para cada classe. A segunda base é constituída de imagens de mamografias organizadas em duas classes, benignas e malignas, sobre um total de 200 imagens separando 100 imagens para cada classe[2]. A terceira base, denominada de COREL-1000, é constituída de um total de 1000 imagens generalizadas separadas em dez classes, onde cada classe possui 100 imagens[3].

A extração de características gera um vetor de características que é a representação visual das imagens. Neste estudo, as análises sobre as bases de imagens foram realizadas utilizando cinco extratores indicados por [Giakoumoglou 2021]. A Tabela 2 apresenta a descrição e dimensões dos extratores de características.

Tabela 2. Descrição e dimensões dos extratores de características

Extrator	Descrição	Dimensões
Gray Histogram	Histograma de nível de cinza da imagem	32
Gray Level Co-occurrence Matrix (GLCM)	Calcula as frequências de ocorrência conjunta de pares de níveis de cinza em uma imagem.	28
First Order Statistics (FOS)	Calculadas a partir do histograma da imagem, que é a função de densidade de probabilidade empírica para pixels únicos.	16
Statistical Feature Matrix (SFM)	Mede as propriedades estatísticas de pares de *pixels* em diversas distâncias dentro de uma imagem.	4
Law's Texture Energy Measures (LTE)	Mede a variação dentro de uma janela de tamanho fixo que percorre a imagem.	6

4.2. Abordagem *Tuning Metrics Relevance Feedback* e Aprendizado de Pesos

Neste trabalho foi desenvolvido um novo algoritmo de RR que otimiza a técnica de [Rui et al. 1998] apresentado na Subseção 2.4, realizando um aprimoramento na forma do cálculo original. No algoritmo original, existem 3 vetores de pesos pré-definidos, distribuindo os pesos de cada camada do algoritmo, realizando múltiplos cálculos na matriz de características para ressaltar cada atributo. Isto acaba levando a um aumento considerável no tempo de processamento em grandes conjuntos de dados, o que não é eficiente em termos de desempenho. No entanto, em nosso algoritmo denominado aqui como *RR Desvio Padrão*, temos apenas uma matriz de peso calculada usando a variância da matriz de características das imagens relevantes (*feedback* do usuário), diminuindo assim o número de cálculo de matrizes.

A Figura 4 apresenta o processo da técnica *RR Desvio Padrão* para obter um vetor de pesos sobre uma matriz de características de imagens relevantes. Primeiro, para cada coluna da matriz de características (a) é calculado o inverso do desvio padrão (b). Nesta etapa, quando o desvio padrão de uma coluna possui valor igual a zero, seleciona-se o atributo com o menor valor de desvio padrão aplicando um peso pré-definido *weight* para ressaltar o atributo naquela dimensão, onde $weight = 1/(0.5 * min(std))$. Por fim, é realizada a normalização do vetor de pesos (c) para manter $\sum wi = 1$, reduzindo a disparidade de escala e a carga computacional.

[1]Rahman, T. (2022). Covid-19 radiography database.

[2]Al-Dhabyani, W., Gomaa, M., Khaled, H., and Fahmy, A. (2020). Dataset of breast ultrasound images.

[3]Wang, J. Z., Li, J., and Wiederhold, G. (2001). Simplicity: Semantics-sensitive integrated matching for picture libraries. IEEE Transactions on pattern analysis and machine intelligence, 23(9):947–963.

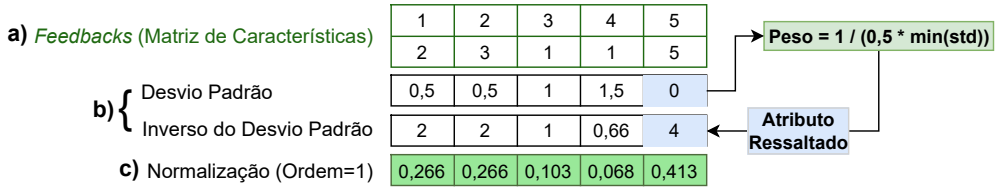

Figura 4. Processo de geração do vetor de pesos sobre a matriz de características de imagens relevantes, indicadas pelo usuário durante a RR.

4.3. Abordagem *Tuning Metrics Relevance Feedback* e Inclusão de Funções de Distância Ponderada na *Slim-Tree*

Como apresentado na Subseção 2.3, a *Slim-Tree* é um MAM eficiente para consultas por similaridade sobre dados complexos. Porém, a indexação com funções de distâncias fixas pode limitar a resposta semântica das consultas, a qual é a imprecisão dos resultados de uma consulta por similaridade. Esse trabalho mostra que a eficácia semântica da estrutura pode ser aprimorada com a inclusão de funções de distâncias ponderadas. Observa-se que a ponderação de uma função de distância onera o cálculo de similaridade, devido ao recálculo de distâncias, porém o ganho semântico obtido geralmente compensa o pequeno tempo adicional demandado.

Para incluir funções de distância ponderada na *Slim-Tree*, é necessário modificar o cálculo da distância entre dois elementos utilizando as funções Lp ponderadas (Subseção 2.1). A *Slim-Tree* também deve ser modificada para permitir que os usuários forneçam um vetor de pesos para cada atributo de forma dinâmica. Para isto, é realizada a construção dos índices da *Slim-Tree* com a definição de uma métrica de similaridade ponderada com todos os valores do vetor de pesos inicialmente iguais a um. A inclusão do vetor de pesos é feita adicionando um novo parâmetro ao método de consulta da *Slim-Tree*, que permite que o usuário especifique os pesos para cada atributo, os quais são aprendidos pela **RR Desvio Padrão**. A inclusão do vetor de pesos deve multiplicar cada dimensão do espaço pela sua respectiva ponderação antes de calcular a distância, realizando o cálculo de distâncias no MAM para entregar um novo resultado obtido pela ponderação, tal procedimento é aplicado nos ciclos de RR até ocorrer uma reindexação da estrutura. O Algoritmo 1 apresenta o pseudocódigo principal para calcular a distância entre dois elementos com ponderação de pesos considerando a distância Euclidiana. Isto permite que após o aprendizado de pesos, o vetor de ponderação obtido seja enviado como parâmetro para que os cálculos de distâncias sejam refeitos para obter os resultados ponderados. No caso de não haver um vetor de pesos submetido como parâmetro, todos os pesos são iniciados com valor igual a um, atuando como uma função de distância padrão.

Algorithm 1 Calibração da distância Euclidiana entre dois objetos na *Slim-Tree*

Input: $Ob1, Obj2$
Output: $\sqrt{(D)}$
$W \leftarrow GetWeights()$ ▷ Captura o último vetor de pesos aprendido
function GETDISTANCE($Obj1, Obj2$)
  if W is empty **then** ▷ Se vetor de pesos for vazio, é uma distância padrão
    for each integer i in $Obj1.size()$ **do**
      $W_i \leftarrow 1$
    end for
  end if
  for each integer i in $Obj1.size()$ **do** ▷ Conforme a equação (2)
    $Tmp \leftarrow (Obj1.GetFeatures()[i] - Obj2.GetFeatures()[i])$
    $D \leftarrow D + Tmp^2 * W_i$ ▷ Calibração das distâncias
  end for
end function

4.4. Abordagem *Tuning Metrics Relevance Feedback* e Reindexação da *Slim-Tree*

As funções de distâncias ponderadas têm melhor eficácia em transformar o espaço de características em um novo espaço que melhor atenda ao usuário. Entretanto, a atualização constante dos pesos durante as consultas implica em um aumento no número de cálculos de distâncias e acessos a disco. Para manter a eficiência das estruturas dos MAMs a longo prazo, é necessário a recriação da árvore métrica, ou seja, reindexando o MAM com o espaço métrico previamente ponderado para preservar a eficiência de recuperação da estrutura. Este tipo de implementação nos leva a um problema a ser respondido: *Quando e de que forma é o melhor momento para reindexar um MAM depois do usuário aprimorar a medida de similaridade ao longo de ciclos de realimentação de relevância?*

A metodologia desenvolvida para definir a necessidade e o momento de reindexação do MAM utiliza o histórico de consultas por similaridade que foram ativadas por RR, durante o armazenamento dos vetores de pesos. Utiliza-se uma janela deslizante sobre os vetores de pesos para realizar cálculos de vetores de pesos médios. Este vetor de pesos médios representa uma média geral de ponderação das características, gerando espaços métricos transformados em conjunto com a redução da dimensionalidade por meio do *t-Distributed Stochastic Neighbor Embedding* (t-SNE) [Van der Maaten and Hinton 2008]. Sobre o espaço métrico transformado t-SNE, é realizado o agrupamento com *k-means* e analisada a qualidade do agrupamento com o coeficiente de *Silhoutte*. Ao receber como entrada os resultados do *k-means*, o coeficiente de *Silhoutte* calcula a média dos valores de *Silhoutte* para todos os pontos em cada agrupamento. Essa métrica varia de -1 a 1, onde valores mais próximos de 1 indicam uma separação melhor, e valores mais próximos de -1 indicam que os pontos foram atribuídos erroneamente. Com base nesta métrica, a primeira reindexação no MAM ocorre com o espaço métrico ponderado que possui o maior coeficiente de *Silhoutte*. Conforme o histórico de consultas no sistema aumenta, uma nova reindexação será realizada quando o coeficiente de *Silhoutte* for maior que a reindexação do espaço métrico mais recente, a Figura 5 ilustra o diagrama dessa metodologia.

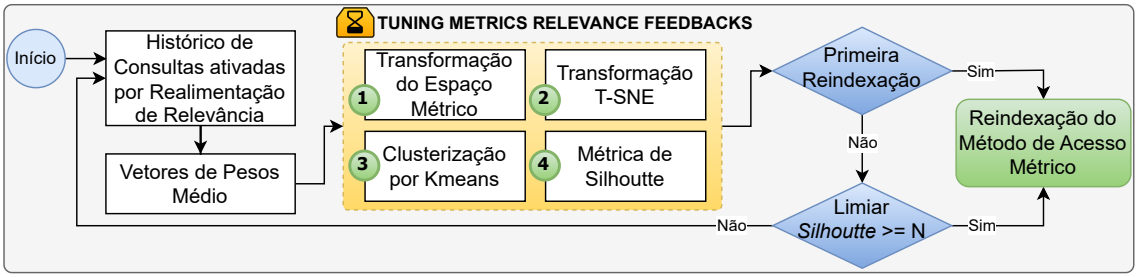

Figura 5. Processo para realização da reindexação do MAM.

5. Resultados

A implementação de uma função de distância ponderada em um MAM visa ampliar a capacidade semântica da estrutura na recuperação de dados. Verificamos o tempo gasto pelas consultas por similaridade até retornar 100% da base de dados COVID-19 a qual possui a maior quantidade de elementos, comparando o tempo de consulta do MAM com a de busca sequencial, que é a abordagem passível de comparação com outros métodos. É possível observar que mesmo com o aumento de cálculos de distâncias e acessos

a disco pelos métodos de buscas utilizando funções de distâncias ponderadas como a ***Range_Slim_Weighted*** e ***kNN_Slim_Weighted***, o tempo gasto com o uso do MAM ainda é muito menor em relação ao sequencial, sendo 70% mais rápido recuperando 10% de toda a base, como apresentado na Figura 6. Vale ressaltar que quando o MAM é reindexado, o tempo gasto nas consultas e o número de cálculos de distância e acessos a disco se equiparam aos da ***Range_Slim*** e ***kNN_Slim*** por não realizarem o reconstrução da árvore métrica.

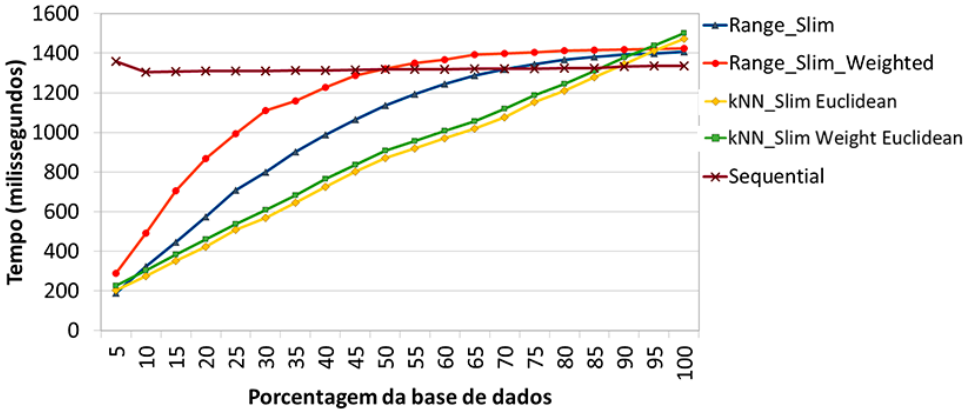

Figura 6. Tempo de consulta em milissegundos sobre a base de COVID-19.

A RR visa aprimorar o potencial semântico das consultas por similaridade. Na Tabela 3 comparamos o efeito de ciclos de RR de até 5 iterações sobre consultas nas bases de dados, cujo objetivo é demonstrar como a RR que aprende um vetor de pesos consegue reorganizar os elementos melhorando significamente o *rank* de uma consulta. Para isto são calculados o *Average Precision* do ciclo de realimentação e ***Mean Average Precision*** sobre as *top* 10 e 20 imagens nas bases de dados. É possível observar que a ponderação da função de distância melhora significativamente a precisão das consultas por similaridade nos diferentes descritores usados nos experimentos, obtendo um ganho de até 42% nas *top* 20 imagens do extrator GLCM.

Tabela 3. *Mean Average Precision* sobre os top 10 e 20 imagens, de cada extrator sem RR e com RR Desvio Padrão em diferentes bases de imagens.

Dataset	Extrator	Sem RR		RR Desvio Padrão		Ganho (%)	
		P@10	P@20	P@10	P@20	P@10	P@20
COREL-1000	**Gray Hist**	61,10	56,13	**79,65**	**71,80**	30	28
COVID-19	**Gray Hist**	81,57	78,26	**91,21**	**86,88**	12	11
	GLCM	63,36	58,18	**84,96**	**82,70**	34	42
	FOS	70,10	65,37	**81,95**	**76,50**	17	17
	SFM	71,96	68,72	**83,10**	**80,62**	15	17
	LTE	64,96	60,78	**75,07**	**70,39**	16	16
BUSI-BREAST	**Gray Hist**	71,68	65,57	**83,07**	**72,00**	16	10
	GLCM	61,82	56,88	**74,28**	**66,20**	20	16
	FOS	64,20	59,55	**73,12**	**64,8**	14	9
	SFM	67,74	63,23	**76,41**	**70,58**	13	12
	LTE	63	58,86	**76,17**	**70,23**	21	19

A Figura 7 apresenta a comparação da quantidade de acessos a disco (Fig.7(a)) e

cálculos de distâncias (Fig.7(b)) da função de distância ponderada em relação à função de distância padrão no MAM na base COVID-19. Apesar do aumento do custo computacional das consultas **Range_Slim Weight Euclidean** e **kNN_Slim Weight Euclidean**, a acurácia e precisão das consultas por similaridade é significativamente aumentada pela calibração das distâncias utilizando a técnica RR, na qual ao longo de comparação da métrica *Silhoutte* é realizada a reindexação do MAM, atingido a mesma quantidade de números de cálculos e acesso a disco do **Range_Slim Euclidean** e **kNN_Slim Euclidean**. Para a análise de acurácia na recuperação e tempo computacional foi utilizado um computador com o sistema operacional Ubuntu 24.04, memória de 8G RAM e processador (8 CPUs) 2.0GHz, as análises foram executadas em memória secundária SDD.

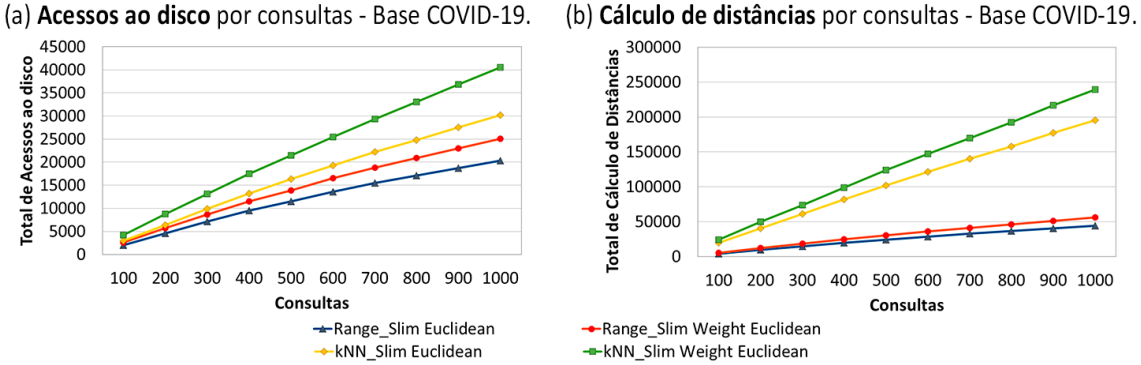

Figura 7. Quantidade de cálculos de distâncias e acesso a disco entre função de distância com e sem ponderação.

A Figura 8 apresenta a análise de Silhueta (*Silhouette*) sobre um vetor de pesos aprendido pelo nosso método *RR Desvio Padrão* para determinar a reindexação do MAM. A Silhueta fornece uma maneira de avaliar parâmetros como o número de *clusters* encontrados. O gráfico de Silhueta (Fig. 8(a)) mostra que o espaço métrico transformado é uma escolha boa para os dados fornecidos devido à presença de *clusters* com pontuações de Silhueta acima da média e também devido a pequenas flutuações no tamanho do gráfico de Silhueta, como também observado no gráfico de dispersão rotulado à direita (Fig.8(b)).

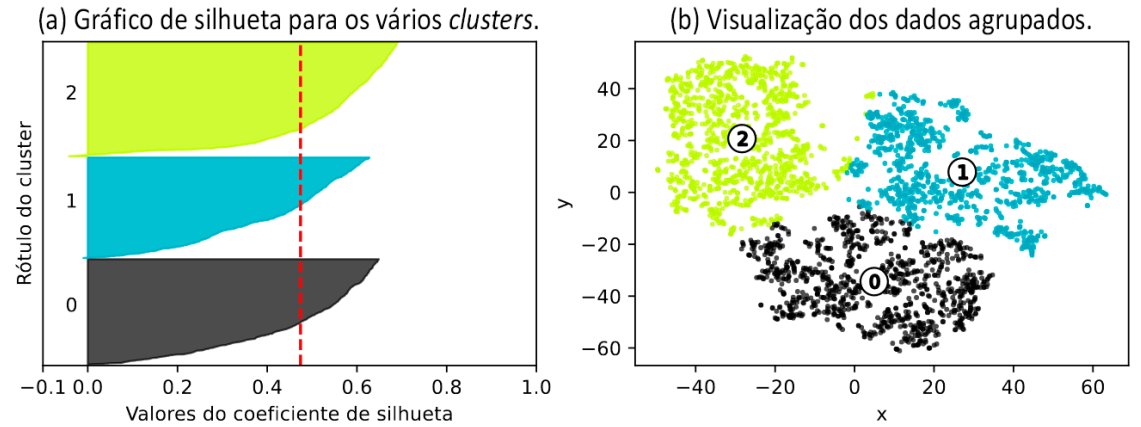

Figura 8. Exemplo de análise para validação da reindexação do MAM, onde a linha pontilhada em (a) é o valor médio das *Silhouettes* calculadas para cada *cluster* (b), indicando a medida geral da qualidade da clusterização.

A Figura 9 demonstra visualmente, por t-SNE, a melhoria semântica e de acurácia do espaço métrico original (Fig.9(a)), utilizando a técnica de *RR com desvio padrão* do extrator de textura GLCM sobre a base de imagens de COVID-19. Devido à limitação de espaço, a visualização dos resultados das demais bases de dados do estudo estão disponíveis no Github[4]. É possível observar a melhoria visual em regiões com elementos de mesma categoria (Fig.9(b)), resultado do ajuste do espaço métrico com a técnica proposta.

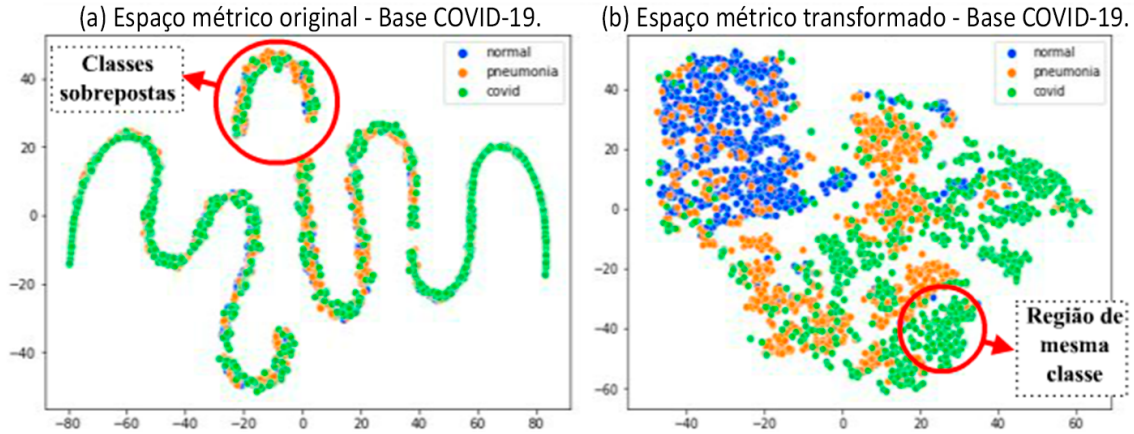

Figura 9. Visualização TSNE do espaço métrico modificado pela técnica RR Desvio Padrão após 5 ciclos de realimentação, indicando que a técnica permite diferenciar e separar as imagens da base em análise, segundo o tipo da categoria.

6. Conclusão

Este artigo propõe uma abordagem de inclusão de funções de distância ponderada no MAM *Slim-Tree* analisando o efeito da calibração dinâmica de distâncias. O algoritmo elaborado de RR foi aplicado em três bases de imagens públicas, alcançando aumentos significativos de precisão. A utilização da RR chega a ter ganhos de até 42% mantendo sempre aprimoramentos em relação aos extratores sem a RR em todas as bases de imagens. Também foram analisados o efeito de ponderar dinamicamente o MAM *Slim-Tree*, o qual mesmo tendo um pequeno aumento no número de cálculos de distâncias, ainda é significantemente mais rápido do que estratégias de recuperação sequencial. Além disso, foi proposta uma metodologia eficaz para reindexação do MAM visando ampliar a eficácia dos métodos que demandam consultas por similaridade. Como trabalhos futuros, prevê-se a inclusão de cálculos da distorção da projeção de espaços métricos quando houver reindexação do MAM. Esse ponto é importante para avaliar o grau de perda de informação da redução de dimensionalidade de espaços métricos e o quanto as distâncias dos espaços métricos foram modificadas.

Agradecimentos Os autores agradecem a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e a Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) Processos 2021/00366-1 e 2016/17078-0 pelo apoio à realização desta pesquisa.

[4]Resultados complementares, disponíveis em: https://github.com/renatomarcacini/Tuning-Metrics-Relevance-Feedback

Referências

Ahmed, A. (2020). Implementing relevance feedback for content-based medical image retrieval. *IEEE Access*, 8:79969–79976.

Bressan, R. S., Bugatti, P. H., and Saito, P. T. (2019). Breast cancer diagnosis through active learning in content-based image retrieval. *Neurocomputing*, 357:1–10.

Chang, Y.-J., Kamataki, K., and Chen, T. (2009). Mean shift feature space warping for relevance feedback. In *2009 16th IEEE International Conference on Image Processing (ICIP)*, pages 1849–1852. IEEE.

Giakoumoglou, N. (2021). Pyfeats: Open source software for image feature extraction. `https://github.com/giakou4/pyfeats`.

Guo, S., Ji, Y., Zhang, C., Xu, C., and Xu, J. (2020). vcbir: A verifiable search engine for content-based image retrieval. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1730–1733. IEEE.

Kim, D.-H. and Chung, C.-W. (2003). Qcluster: relevance feedback using adaptive clustering for content-based image retrieval. In *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, pages 599–610.

Kumaran, K., Papageorgiou, D. J., Takac, M., Lueg, L., and Sahinidis, N. V. (2021). Active metric learning for supervised classification. *Computers & Chemical Engineering*, 144:107132.

Li, Z., Zhang, X., Müller, H., and Zhang, S. (2018). Large-scale retrieval for medical image analytics: A comprehensive review. *Medical image analysis*, 43:66–84.

Mohanan, A. and Raju, S. (2017). A survey on different relevance feedback techniques in content based image retrieval. *Int. Res. J. Eng. Technol*, 4(2):582–585.

Rui, Y., Huang, T. S., Ortega, M., and Mehrotra, S. (1998). Relevance feedback: a power tool for interactive content-based image retrieval. *IEEE Transactions on circuits and systems for video technology*, 8(5):644–655.

Silva, M. P. d. (2009). *Processamento de consultas por similaridade em imagens médicas visando à recuperação perceptual guiada pelo usuário*. PhD thesis, Universidade de São Paulo.

Tian, D. (2018). A review on relevance feedback for content-based image retrieval. *J. Inf. Hiding Multim. Signal Process.*, 9(1):108–119.

Tyagi, V. (2018). *Understanding digital image processing*. CRC Press.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).