UNIVERSIDADE DE SÃO PAULO
FFCLRP - DEPARTAMENTO DE COMPUTAÇÃO E MATEMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

Busca por similaridade utilizando grafo de interações NK

José Carlos Bueno de Moraes

Dissertação apresentada à Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da USP, como parte das exigências para a obtenção do título de Mestre em Ciências, Área: Computação Aplicada.

RIBEIRÃO PRETO - SP
2020

Busca por similaridade utilizando grafo de interações NK

José Carlos Bueno de Moraes

Orientador: Prof. Dr. Renato Tinós

Versão Original

2020

Moraes, José Carlos Bueno de

Busca por similaridade utilizando grafo de interações NK. Ribeirão Preto, 2020.

71 p. : il. ; 30 cm

Dissertação apresentada à Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da USP. Área de concentração: Neurologia.

Orientador: Tinós, Renato.

1. busca por similaridade 2. grafo de interações NK 3. Inteligência artificial

AGRADECIMENTOS

Agradeço a Deus pelo dom da vida, que me permitiu viver este caminho e que nele colocou pessoas essenciais para que tudo se tornasse melhor.

Agradeço à meus pais Eufrosino Carlos Bueno de Moraes e Teresinha Pastre Bueno de Moraes, por todo amor, apoio e educação que me proporcionaram almejar amplos horizontes.

Ao meu orientador Dr. Renato Tinós, pela imensa oportunidade de ser seu aluno. Obrigado pela compreensão, encorajamento e por todo apoio na realização deste trabalho. Gratidão pela sua confiança!

À minha namorada Aline Zanatta. Obrigado principalmente pelo apoio e compreensão nessa reta final na conclusão deste trabalho.

Aos meus amigos de pós-graduação: Diogenes, Leandro e Emerson pelos dias que compartilhamos no início dessa caminhada. Obrigado por todos os momentos de descontração e amizade.

Aos funcionários do departamento de Computação e Matemática de Ribeirão Preto, pela notável paciência e disponibilidade na resolução de eventuais problemas que surgiram pelo caminho.

RESUMO

Com o crescimento do volume de dados ao longo dos anos, foram desenvolvidas técnicas de busca por similaridade para responder às necessidades dos usuários de diversos segmentos. A evolução das técnicas de busca por similaridade vem permitindo recuperar objetos presentes em grandes bases de dados similares a um objeto fornecido pelo usuário, auxiliando na tomada de decisão cada vez mais correta utilizada em diversos segmentos de estudo e aplicação. Um método de busca por similaridade baseado no grafo de interações NK é proposto. O grafo de interações NK foi originalmente empregado para agrupamento e é construído com base na distância e densidade espacial dos objetos em um conjunto de dados. Duas variações do método são investigadas. Nas duas variações, *k* objetos são retornados visitando-se vértices do grafo de interações NK a partir do vértice inicial relacionado ao exemplo do conjunto de dados que está mais próximo do objeto a ser consultado. No método NK A, os *k* objetos relacionados a vértices com arestas incidentes ao vértice inicial são retornados. No método NK B, *k* vértices são visitados a partir do vértice inicial. O próximo vértice visitado é aquele com aresta incidente ao vértice atual e que está mais próximo do novo objeto a ser consultado. Os *k* objetos relacionados aos vértices visitados são retornados. Os algoritmos propostos são comparados entre si e com a busca por similaridade baseada apenas na distância. Os resultados experimentais indicam que os métodos propostos apresentam bom desempenho quando existem agrupamentos com formas arbitrárias no conjunto de dados.

Palavras chaves: busca por similaridade; grafo de interações NK; inteligência artificial.

ABSTRACT

With the growth in the volume of data over the years, similarity search techniques were applied to respond to the needs of users in different segments. The evolution of similarity search techniques allows retrieving objects present in large databases, such as an object provided by the user, assisting in the decision making increasingly correct, used in several studies and applications. A similarity search method based on the NK interaction graph is proposed. The NK interaction graph was originally employed for clustering and is built based on distance and spatial density of the objects in a dataset. Two variations of the method are investigated. In the two variations, *k* objects are returned by visiting vertices of the NK interaction graph from the initial vertex related to the example of the dataset that is closer to the object to be consulted. In NK A, the *k* objects related to vertices with edges incident to the initial vertex are returned. In NK B, *k* vertices are visited starting from the initial vertex. The next visited vertex is that one with edge incident to the current vertex and that is closest to the new object to be consulted. The *k* objects related to the visited vertices are returned. The proposed algorithms are compared with each other and with the search for similarity based only on distance. The experimental results indicate that the proposed methods present good performance when there are clusters with arbitrary shapes in the dataset.

Key words: search similarity, NK interaction graph; artificial intelligence.

LISTA DE FIGURAS

LISTA DE TABELAS

Tabela 17: Tabela contendo as informações de cada imagem resultante da consulta por similaridade KNN adaptado para o terceiro exemplo.

Tabela 18: Tabela contendo as informações de cada imagem resultante da consulta por similaridade NK B para o quarto exemplo.

Tabela 19: Tabela contendo as informações de cada imagem resultante da consulta por similaridade KNN adaptado para o quarto exemplo.

LISTA DE ABREVIATURAS E SIGLAS

USP	Universidade de São Paulo
FFCLRP	Faculdade de Filosofia Ciências e Letras de Ribeirão Preto
KNN	*K*-vizinhos Mais Próximos
NKGA	Algoritmo Genético Híbrido NK
NKCV2	Critério de Validação de Cluster NK
SGBDS	Sistemas de Gerenciamento de Banco de Dados
Acc	Acurácia

Sumário

Capítulo 1

1.INTRODUÇÃO

Com o crescimento do volume de dados ao longo dos anos, foram desenvolvidas técnicas de busca por similaridade para responder às necessidades dos usuários em diversos segmentos do conhecimento (HSINCHUN et al., 2012). A evolução das técnicas de busca de similaridade vem permitindo recuperar objetos presentes em grandes bases de dados similares a um objeto fornecido pelo usuário de maneira eficiente, auxiliando na tomada de decisão em diversas aplicações. Por exemplo, na área da Medicina, busca por similaridade de exames (como imagens médicas, exames laboratoriais, entre outros) e laudos têm potencial para aumentar a eficiência das decisões médicas, reduzir custos e otimizar o tempo dos especialistas na análise de casos (CARPINETO & ROMANO, 2012).

Dentre as técnicas de aprendizado de máquina supervisionadas, a técnica mais comumente utilizada de busca por similaridade é aquela baseada em distância. O algoritmo dos *K*-vizinhos mais próximos (*K-nearest neighbours* - KNN) (AHA et al., 1991) pode ser adaptado para retornar os *K* objetos de uma base de treinamento mais similares ao objeto que está sendo consultado. No entanto, as informações sobre densidade espacial de objetos não são consideradas quando apenas KNN é empregado.

Informações adicionais sobre densidade espacial podem ser úteis especialmente em bases de dados com agrupamentos com formas arbitrárias. A densidade espacial de objetos é explorada por algumas técnicas de clusterização para, entre outros, produzir agrupamentos que não são necessariamente hiper-esféricos (ESTER et al., 1996; RODRIGUEZ & LAIO, 2014; Tinós et al., 2018). Técnicas de agrupamento tem sido aplicada nas mais diversas áreas do conhecimento (HRUSCHKA et al., 2009). Conceitos utilizados em agrupamento podem ser especialmente úteis na recuperação de informação e na visualização de dados (XU & TIAN, 2015).

Em TINÓS et al. (2018), o NKGA (*NK Hybrid Genetic Algorithm*) foi proposto para o problema de clustering. O NKGA utiliza tanto a distância entre objetos como a densidade espacial para o agrupamento de objetos. Para avaliar a soluções (particionamentos dos objetos da base de dados), o NKGA usa uma função de validação interna chamada NKCV2. Esta função utiliza informações sobre a disposição de *N* objetos, sendo *N* o número de objetos na base de dados. Cada grupo é composto de *K*+1 objetos, sendo *K* um parâmetro definido pelo usuário. As informações sobre os grupos de objetos são capturadas no grafo de interações NK. Tanto informações sobre densidade como de distância entre objetos são utilizadas para construir o

grafo de interações NK. Resultados experimentais mostram que agrupamentos de dados com formas arbitrárias podem ser identificados usando NKGA com *K* pequeno.

Neste trabalho, propomos um método de busca por similaridade baseado no grafo de interações NK. Duas variações do método são investigadas. Nas duas variações, *k* objetos são retornados percorrendo-se o grafo de interações NK a partir do vértice inicial v_x relacionado ao objeto da base de dados mais similar ao objeto ***x*** à ser consultado. Na primeira variação (método NK A), *k*=*K*+1, sendo que os *k* objetos cujos vértices têm arestas incidentes no vértice v_x são retornados. Na segunda variação (método NK B), caminha-se a partir de $v_x$sempre alcançando o vértice, com aresta incidente, cujo objeto é mais próximo ao objeto consultado. Após *k* passos, *k* objetos são então retornados: aqueles relacionados aos *k* vértices visitados. Os algoritmos propostos são comparados entre si e com a busca por similaridade baseada em distância (chamada aqui, por simplicidade, de KNN adaptado).

1.1 Motivação

A área de consultas por similaridade tem evoluído consideravelmente nos últimos anos. Vários algoritmos têm sido propostos (HSINCHUN et al., 2012), contribuindo para inúmeras áreas de aplicação. Entretanto, alguns problemas relacionados ao tema ainda são frequentemente abordados, tais como:

- Detecção de agrupamentos
- Busca por similaridade em dados n-dimensionais

Existem algumas técnicas que buscam resolver estes problemas, contribuindo para inúmeras áreas de aplicação. Entretanto, nenhuma leva em consideração distância e densidade espacial. Portanto, o escopo desta pesquisa foi definido para resolução de problemas reais encontrado em diversas áreas do conhecimento, ou seja, buscar exemplos similares considerando distância e densidade espacial a partir de um conjunto de dados.

1.2 Objetivo

O objetivo deste trabalho de pesquisa consiste no desenvolvimento e análise de um novo algoritmo de busca por similaridade baseado no grafo de interações NK. Esse novo método apresenta uma abordagem para encontrar similaridade em dados n-dimensionais a partir do grafo de interações NK. Para isso foram utilizadas três abordagens em que são divididas:

- Método de busca de similaridade baseado nos vértices de arestas incidentes no grafo de interações NK (NK A).
- Método de busca por similaridade baseado na busca dos vértices, com aresta incidente, cujo objeto é mais próximo ao objeto consultado (NK B).
- Aplicar os algoritmos implementados em problemas simples e em problemas do mundo real. Comparar o desempenho do método proposto com o do método baseado no KNN adaptado.

1.3 Organização

O restante deste documento está organizado da seguinte maneira: no Capítulo 2 são descritos os referenciais teóricos sobre medidas de similaridade, consulta por similaridade, princípios básicos da teoria dos grafos e grafo de interações NK. No Capítulo 3 é descrita a metodologia utilizada no desenvolvimento deste trabalho, bem como, dois novos algoritmos de busca por similaridade via grafo de interações NK. No Capítulo 4, são descritos os experimentos, e resultados obtidos na análise da busca por similaridade, proposto por esse trabalho, comparando-o com a busca dos k-vizinhos mais próximo, existente na literatura. Finalmente, no Capítulo 5 são apresentadas as conclusões, principais contribuições e trabalhos futuros.

Capítulo 2

2.REFERENCIAL TEÓRICO

2.1 Medidas de similaridade

O conhecimento sobre similaridade e dissimilaridade é fundamental em muitas áreas da ciência da computação, como por exemplo, para mineração de dados, reconhecimento de padrões, aprendizado de máquinas e inteligência artificial (BERGADANO & De RAEDT; 1994).

A Figura 1 ilustra um pequeno exemplo do que é similaridade, pode-se notar que a estrela (A) é similar a estrela (C). As estrelas (A), (B) e (C) têm o mesmo tamanho, enquanto (A), (C) e (D) apresentam a mesma cor. As características de cor e tamanho são exemplos de variáveis que podem ser mensuradas. Similaridade pode ser difícil de ser formalmente definida. Intuitivamente, ela pode ser entendida como um valor que reflete a distância entre dois objetos ou duas características, sendo esse valor entre 0 a 1 (TEKNOMO & KARDI, 2015).

(A) (B) (C) (D)

Figura 1: Representação de quatro estrelas (A), (B), (C) e (D), na qual as estrelas estão diferenciadas em cor e tamanho.

A abordagem mais comum de entender similaridade é representar os objetos como pontos em um espaço geométrico. De acordo com essa abordagem, a semelhança entre dois objetos é definido através de uma função de distância, ou função de dissimilaridade $d(\boldsymbol{x},\boldsymbol{y})$, que retorna zero se ambos os objetos $\boldsymbol{x}$ e $\boldsymbol{y}$ forem idênticos, e um valor positivo quanto maior a distância ou dissimilaridade entre os objetos. Sendo a distância medida por alguma função métrica **(**BERDADANO & De RAEDT; 1994**)**.

Muitas medidas de distância existem na literatura **(**CHA & SUNG-HYUK; 2007**)**. Este trabalho concentra-se na medida de similaridade definida de acordo com a distância euclidiana. A compreensão de similaridade presente em um conjunto de dados é interessante para muitos

campos de pesquisa porque pode revelar informações importantes sobre o problema em questão, como por exemplo:

a. Distinguir um objeto de outro;
b. Agrupar de acordo com sua similaridade e dissimilaridade;
c. Investigar e compreender as características de cada grupo;
d. Explicar a formação de agrupamentos;
e. Organizar e recuperar informações de maneira mais eficientes;
f. Agrupar um novo objeto em grupos existentes;
g. Prever o comportamento de um novo objeto;
h. Descobrir a estrutura dentro do conjunto de dados;

2.1.2 Métrica

Métrica é um conceito que generaliza a idéia geométrica de distância. Um conjunto em que há uma métrica definida recebe o nome de espaço métrico. Para ser uma métrica, uma medida *d* deve satisfazer as seguintes operações **(**HUANG & ANNA; 2008**)**.

a. A distância entre quaisquer dois pontos não deve ser negativa, isto é, $d(\mathbf{x},\mathbf{y}) \geq 0$.

b. A distância entre dois objetos deve ser zero se, e somente se, os dois objetos são idênticos, isto é, $d(\mathbf{x},\mathbf{y}) = 0$, se e somente se, $\mathbf{x} = \mathbf{y}$.

c. A distância deve ser simétrica, isto é, a distância de **x** à **y** é a mesma de **y** à **x**, ou seja, $d(\mathbf{x},\mathbf{y}) = d(\mathbf{y},\mathbf{x})$.

d. A medida deve satisfazer a desigualdade triangular, que é $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$.

2.1.3 Distância Euclidiana

A distância Euclidiana é uma métrica padrão para problemas geométricos. Ela é amplamente utilizada em problemas de agrupamento. Ela satisfaz todas as quatro condições acima e, portanto, é considerada uma métrica. É também a medida de distância padrão utilizada com o algoritmo de agrupamento *k-means* e no algoritmo de classificação *K*-vizinhos próximos (KNN) **(**HUANG & ANNA; 2008**)**.

Para medir a distância entre dois objetos, representados pelos seus vetores de características **X** e **Y** respectivamente, a distância euclidiana desses dois objetos é definida como:

$$d(\boldsymbol{X},\boldsymbol{Y}) = (\sum_{i=1}^{n}(\boldsymbol{x_i} - \boldsymbol{y_i})^2)^{1/2} \quad (1)$$

onde os vetores de características são $\mathbf{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ e $\mathbf{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n\}$, respectivamente e *n* é o número de dimensões dos objetos.

2.2 Consultas por similaridade

Em contraste com as consultas tradicionais, tais como feitas nos sistemas de gerenciamento de banco de dados (SGBD), que oferecem recursos eficazes para realizar buscas sobre dados usando relações de igualdade e operadores lógicos nos dados armazenados, as consultas por similaridade buscam por dados complexos (tais como, imagens, vídeos, sequência de DNA, textos longos e impressões digitais) usando as relações de similaridade entre os objetos (vetores de características). Nestes casos, operadores lógicos muitas vezes não se aplicam, ou simplesmente, são de pouca serventia.

Para esses tipos de dados é mais significativo fazer uso de consultas por similaridade, onde a busca por elementos em um conjunto é feita de acordo com algum critério de similaridade. Em outras palavras, a busca por similaridade consiste em comparar cada elemento de um conjunto com um elemento consultado, e selecionar somente aqueles que satisfazem ao critério de similaridade. O tipo mais básico de consulta por similaridade é a utilização do algoritmo de distância, nesse caso o principal algoritmo é a consulta baseada nos *K*-vizinhos mais próximos (BOHM et al., 2001; CHÁVEZ et al., 2001).

2.2.1 Consulta aos *K*-vizinhos mais próximos

O algoritmo *K*-vizinhos mais próximos (*K-Nearest Neighbor - KNN*) é um algoritmo de classificação no qual um novo objeto $\boldsymbol{Z}$ é classificado examinando a classe majoritária entre os $\boldsymbol{k}$ objetos da base de dados mais próximos a $\boldsymbol{Z}$ (AHA et al., 1991). Um método similar a KNN pode ser utilizado em busca por similaridade. Neste algoritmo, busca por similaridade chamaremos de KNN ou KNN adaptado, os $\boldsymbol{k}$ objetos da base de dados mais próximos a **z** são retornados como resultado da busca por similaridade. É importante salientar que esse tipo de algoritmo pode retornar menos que *k* elementos se o conjunto de dados for menor do que *k*. Como também, pode existir dois ou mais exemplos com a mesma distância do novo elemento

de consulta. Neste caso, existem algumas abordagens utilizadas na literatura, como decidir entre os objetos arbitrariamente, ou selecionar todos os exemplos com mesma distância, retornando mais do que *k* elementos (FERREIRA et al., 2011).

Um exemplo de consulta aos *K*-vizinhos mais próximos em uma base de dados é representado pela Figura 2, que ilustra uma consulta aos 5-vizinhos mais próximos em um domínio bidimensional empregando a função de distância Euclidiana. O ponto **z** representa um novo objeto no conjunto de dados e os pontos $\{\boldsymbol{z_1}, \boldsymbol{z_2}, \boldsymbol{z_3}, \boldsymbol{z_4}, \boldsymbol{z_5}\}$ são os pontos mais próximos do ponto principal **z**. Portanto, os pontos são retornados como objetos mais similares ao objeto de entrada **z**.

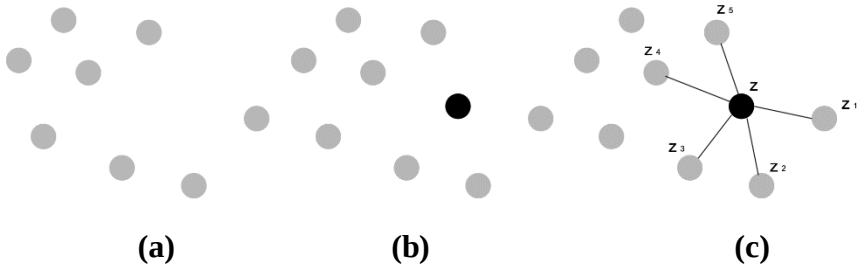

Figura 2: Representação da consulta aos *K*-vizinhos mais próximos. Na figura (a) ilustra o conjunto de dados existente, na figura (b) a entrada de um novo objeto a base de dados, e na figura (c) o cálculo da distância do novo objeto **z** aos k objetos mais próximos.

2.3 Conceitos básicos da teoria de grafos

2.3.1 Tipos de grafos

A teoria dos grafos estuda a relações entre os objetos de um determinado conjunto. Existem diversas formas de modelarmos um problema utilizando grafos. Cada forma ou tipo possui um conjunto de características para representar adequadamente um certo problema. Um grafo G (V,A) é definido pelo par de conjunto V e A, onde:

- G representa a função que associa cada aresta **a** a um par não ordenado (y,w) de vértices chamados extremos de **a**.
- V representa o conjunto não vazio de vértices.
- A representa o conjunto de pares ordenados a = (y,w), onde y e w são pertencentes a V.

Os exemplos mais simples conhecidos podem ser agrupados em quatro definições principais (FIGUEIREDO, 2011).

- Grafo simples refere-se ao grafo em que não tem laços nem duas ligações distintas com o mesmo par de extremos. Sua definição de grafo *G* é representada por dois conjuntos, *G* = (*V, A*), onde *V* e *A* representam os vértices e arestas de *G*, respectivamente.
- Grafo direcionado é um grafo simples, onde os pares de arestas *e* = (*y, w*) e *t* = (*w, y*) são considerados distintos, ou seja, o par (*y*,*w*) indica que aresta está direcionada de *y* para w, enquanto, o par (*w*,*y*) indica o oposto.
- Grafo ponderado refere-se a u m grafo simples, onde suas arestas possuem pesos. Esses pesos são valores definidos por uma função de mapeamento *A* : *P* -> $\mathbb{R}$, sendo $\mathbb{R}$ o conjunto dos números reais.
- Grafo rotulado refere-se a um grafo simples, quando cada vértices ou arestas possuem rótulos. Os rótulos são definidos por uma função de mapeamento, onde *V* = *v* -> *L*, sendo *L* o conjunto de rótulos, enquanto *A* = (*y, w, p*), em que *y* a *w* indica que aresta está direcionada de *y* para *w*, sendo *p* o valor da aresta .

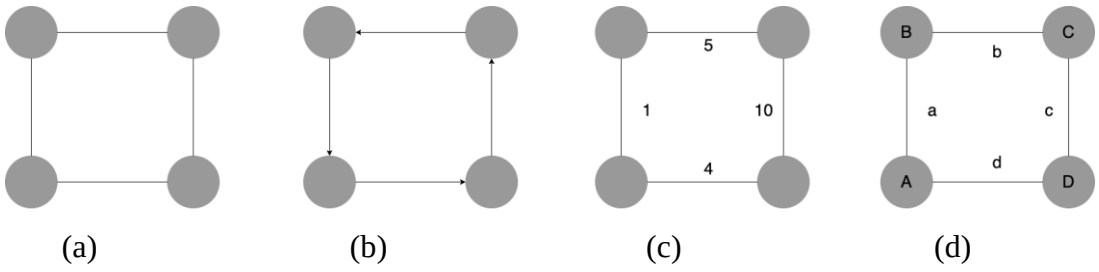

Figura 3: Tipos de grafos. (a) Grafo simples (b) Grafo direcionado (c) Grafo ponderado (d) Grafo rotulado

Na Figura 3 é ilustrado a representação das definições para cada tipos de grafo. É importante ressaltar que outros modelos de grafos podem existir, combinando as características das definições anteriormente apresentadas (GROSS & YELLEN, 1999).

2.3.2 Subgrafos, passeio, caminhos e ciclo

Um grafo *G* pode ser considerado como a concatenação de várias estruturas ou grafos menores. Essas estruturas são conhecidas por subgrafos. Um grafo *H* é um subgrafo de um grafo

G se todo vértice de *H* é vértice de *G* e toda aresta de *H* é aresta de *G*. Um passeio em um grafo é uma sequência de vértices dotada da seguinte propriedade: se *y* e *w* são vértices consecutivos, então *y*-*w* é uma aresta do grafo. Um caminho em um grafo é um passeio sem arestas repetidas, ou seja, um passeio em que as arestas são todas diferentes entre si. Um ciclo, por sua vez, apresenta uma estrutura parecida a um caminho simples, contudo com o primeiro e o último vértice da cadeia iguais. Todos as arestas de um ciclo apontam no mesmo sentido, de um vértice do ciclo para o seu sucessor. Na Figura 4 é ilustrado a representação para cada tipos de grafo apresentado.

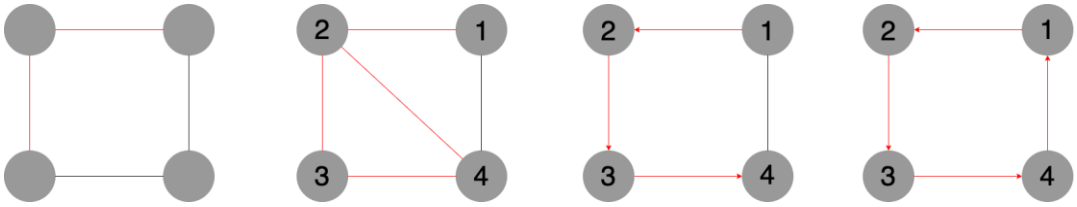

Figura 4: Modelo de grafos (a) Subgrafo (b) Passeio (c) Caminho (d) Ciclo

2.3.3 Grafo de Interações NK

Em **(**TINÓS *et al*., 2018**)**, o NKGA (NK Hybrid Genetic Algorithm) foi proposto para o problema de agrupamento. O NKGA utiliza tanto a distância entre objetos como a densidade espacial para o agrupamento de objetos. Para avaliar a soluções (particionamentos dos objetos da base de dados), o NKGA usa uma função de validação interna chamada NKCV2. Esta função utiliza informações sobre a disposição de *N* pequenos grupos de objetos, sendo *N* o número de objetos na base de dados. Cada grupo é composto de *K*+1 objetos, sendo *K* um parâmetro definido pelo usuário. As informações sobre os grupos de objetos são capturadas no grafo de interações NK. Tanto informações sobre densidade como de distância entre objetos são utilizadas para construir o grafo de interações NK. Resultados experimentais mostram que agrupamentos de dados com formas arbitrárias podem ser identificados usando NKGA com *K* pequeno.

NKCV2 é dado por:

$$f(y) = \sum_{i=1}^{n} f_i(\boldsymbol{Y}, \boldsymbol{M}_i) \qquad (3)$$

onde $\{\boldsymbol{Y} \in Z^n\} : 1 \leq y_j \leq N_c\,(\boldsymbol{Y})$ é a solução candidata, i.e., **Y** é um vetor de inteiros, sendo que cada elemento $\boldsymbol{y}_j$ define o cluster assinalado ($N_c\,(\boldsymbol{Y})$ agrupamentos) para o j-ésimo objeto $\mathbf{x}_j$ da base de dados. A máscara $\boldsymbol{M_i} \in B^n$ indica os elementos de **y** que influenciam a subfunção f_i. Cada subfunção é influenciada por *K*+1 objetos. Assim, *K* indica o grau de interações entre os objetos. Pode-se observar que a função NKCV2 (Eq. 3) é semelhante à função usada no modelo NK landscapes (Kauffman, 1993). No modelo NK landscapes, grafos de interações aleatórios ou adjacentes são geralmente empregados e os valores para as subfunções f_i são gerados aleatoriamente. No NKCV2, o grafo de interações é dado pela disposição dos objetos no espaço n-dimensional.

As interações entre os elementos de **y** nas sub-funções f_i podem ser representadas por um vetor de listas de adjacências $M = [m_1 m_2 \ldots m_N]$. O grafo direcionado correspondente ao vetor **M** é chamado de grafo de interações, G_{ep}. No grafo de interações, o vértice v_i, *i* = 1,...,*N*, representa o i-ésimo elemento do vetor solução **y**. Cada aresta (v_j, v_i) indica que o elemento $y_j$influencia a subfunção f_i. Resta, portanto, definir como o grafo de interações é ligado.

No agrupamento, a interação entre os objetos em um mesmo grupo deve-se dar principalmente entre os objetos mais próximos, i.e., mais similares. Dessa maneira, arestas entre objetos próximos são criadas de modo a gerar o grafo de interações, G_{ep}. Entretanto, antes, para cada vértice uma aresta é criada até o vértice correspondente ao objeto mais próximo com maior densidade **(**TINÓS et al., 2018**).** Este passo assegura que agrupamentos diferentes sejam ligados por arestas no grafo de interações.

A Figura 5 apresenta esse processo de criação do grafo de interações, G_{ep}.

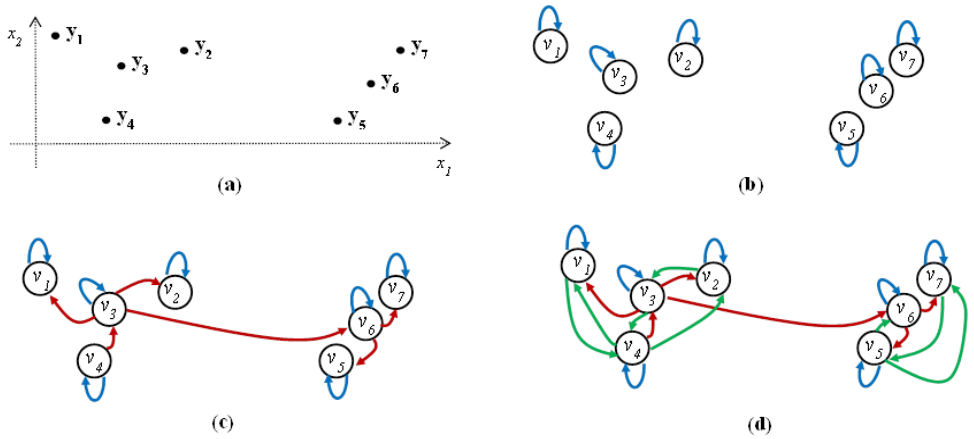

Figura 5 - Exemplo de construção do grafo de interações NK com K = 2 para um conjunto com 7 objetos bidimensionais (N = 7, n = 2). Cada objeto da base de dados (a) é associado com um vértice com auto-loop (b). A densidade dos objetos é calculada e cada vértice é ligado ao vértice associado com o objeto mais próximo com maior densidade (c). O próximo passo é adicionar arestas para os vértices representando objetos mais próximos até que o grau de entrada de cada vértice seja igual a K+1. O gráfico de interações (d) tem N = 7 vértices e $N(K+1)$ arestas.

O grafo de interações NK é um grafo direcionado com N vértices, cada um com grau de entrada K+1. Dada uma base de dados (treinamento) com N objetos n-dimensionais, o primeiro passo para a construção do grafo é adicionar vértices v_i, i = 1,...,N, para cada objeto $\mathbf{y}_i$da base de dados. Cada vértice possui auto-loop, que aqui poderia ser ignorado. Se ignorarmos os auto-loops, o grau de entrada para cada vértice é então igual a K; entretanto, para fins de uniformidade com (TINÓS *et al*., 2018), o auto-loop será preservado na descrição do grafo, apesar de não ser utilizado pela busca por similaridade. Cada aresta (v_j, v_i) indica que o j-ésimo objeto é relacionado com o i-ésimo objeto.

É importante ressaltar que a construção das arestas leva em consideração a distância Euclidiana para objetos próximos e a densidade dos objetos. Após a criação dos vértices com auto-loop, a densidade dos objetos é calculada. Para o i-ésimo objeto, a densidade é dada por:

$$\rho_i = \sum_{j=1}^{N} K(\boldsymbol{y}_i - \boldsymbol{y}_j) \tag{4}$$

sendo **K**(.) a função Kernel dada por:

$$K(y_i - y_j) = e^{\frac{-\|y_i - y_j\|^2}{2\epsilon^2}} \tag{5}$$

sendo ϵ o parâmetro que define a distância de corte. Aqui, ϵ é escolhido de modo que o número médio de vizinhos de um objeto seja 2% do total de objetos da base de treinamento (RODRIGUEZ & LAIO, 2014; TINÓS *et al.,* 2018).

Para cada vértice v_i, o vértice v_{ai}representando o objeto mais próximo que possui densidade maior que o objeto relacionado a v_ié identificado. Então, uma aresta (v_{ai}, v_i) é criada. O último passo é adicionar arestas para os vértices representando objetos mais próximos até que o grau de entrada de cada vértice seja igual a *K*+1. A Figura 5 apresenta um exemplo do processo de criação do grafo de interações NK.

2.4 Linguagem de programação Python

A linguagem de programação Python está se estabelecendo como uma das linguagens mais populares para computação científica **(**Estas são as 10 linguagens de programação mais populares atualmente; 2020**)**. Por se tratar de uma linguagem livre sob a licença *Python Software Foundation License*, multiplataforma e de ser de fácil aprendizado, a linguagem Python também fornece um grande ecossistema de bibliotecas, sendo uma escolha atraente para desenvolvimento algoritmo e análise de dados (Guia qual a melhor linguagem para ciência de dados; 2020**)**. A biblioteca *NetworkX* é um pacote Python para criação de grafos.

Com a *NetworkX* é possível fazer manipulação, estudo de estrutura, dinâmica e funções de redes complexas. Portanto, foi utilizado Python como linguagem de programação para o desenvolvimento deste trabalho.

Capítulo 3

3. METODOLOGIA

Os métodos propostos são comparados nos experimentos ao KNN adaptado, no qual, ao invés de utilizar KNN para classificar novos objetos a partir da distância para objetos de uma base de dados (também chamada de base de treinamento), utiliza-se a distância para cálculo da dissimilaridade e são retornados os *k* objetos mais próximos ao objeto consultado. A seguir, baseados no grafo de interações NK, são apresentados os métodos propostos, bem como os conjuntos de dados. Em seguida, o método de avaliação é apresentado.

3.1 Base de dados

Para analisar o comportamento dos métodos propostos e compará-los ao KNN adaptado, foram realizados experimentos utilizando os conjuntos de dados: Covid, Íris, *E Coli* ,*Pathbased*, *Spiral, Aggregation, A.K Jain's Toy Problem, R15, D31, Flame, Zahn's Compound.* As propriedades destes conjuntos são mostradas na Tabela 1. Os 8 últimos conjuntos de dados pertencem ao benchmark Shape Sets (Franti & Sieranoja, 2018), enquanto que os conjuntos Íris e E Coli são do repositório de Aprendizado de Máquina UCI (DUA & GRAFF, 2019).

Todos os conjuntos, com exceção do conjunto Covid são previamente conhecidos quais objetos pertencem à mesma classe e quais não pertencem. Apesar de busca por similaridade ser prioritariamente uma tarefa não-supervisionada, o uso de bases com exemplos rotulados permite verificar se os exemplos retornados pela busca possuem similaridade em relação à classe do exemplo original.

Informações sobre os conjuntos do benchmark Shape Sets podem ser encontradas em *http://cs.joensuu.fi/sipu/datasets/*. Todos os conjuntos têm dimensão n=2, o que facilita a visualização da disposição dos objetos. Alguns dos conjuntos de Shape Sets possuem agrupamentos que não são hiper esféricos, sendo difíceis de serem detectados por algoritmos de agrupamento que utilizam apenas a distância entre os objetos para o particionamento, como o *k-means*. Já os conjuntos *Aggregation e R15* possuem apenas agrupamentos hiper esféricos. Portanto, as bases deste conjunto são interessantes para verificar o desempenho dos algoritmos de busca por similaridade quando existem diferentes combinações de agrupamentos hiper esféricos e de forma arbitrária. São também interessantes por serem formadas por um número variado de agrupamentos.

Os conjuntos de dado Íris e E Coli são conjuntos bastante utilizados na literatura de Aprendizado de Máquina. Diferente dos conjuntos da base Shape Sets, ambos possuem mais que dois atributos no conjunto de dados. Todos os conjuntos citados foram utilizados na análise quantitativa utilizada na comparação dos algoritmos.

O conjunto de dados Covid contém exemplos da base COVID-19 Image Data Collection, tendo sido utilizado na análise qualitativa apresentada no Capítulo 4. Esse conjunto de dados público contém imagens de raio-X de pacientes com COVID-19, uma doença respiratória aguda causada pelo coronavírus (SARS-CoV-2). Novos dados são adicionados constantemente ao repositório COVID-19 Image Data Collection, vindas de fontes públicas, tais como médicos e hospitais. O repositório pode ser encontrado no endereço *https://github.com/ieee8023/covid-chestxray-dataset*. No momento da aquisição das imagens para este estudo, o repositório contava com 183 imagens de raio-X de pacientes diagnosticados com COVID-19. A base Covid utilizada neste trabalho foi preparada por Rafael Del Lama, que a utilizou, com algumas modificações, em seu trabalho de mestrado (DEL LAMA, 2020). A biblioteca *pyRadiomics* (VAN GRIETHUYSEN et al., 2017) foi utilizada para se extrair das imagens 99 atributos, como de forma e textura, que estão relacionados à distribuição de níveis de cinza das imagens de Raio X.

Tabela 1: Descrição dos conjuntos de dados com valor real utilizados nos experimentos.

Nome	**Nº Objetos**	**Dimensão**	**Descrição**
Aggregation	788	2	Dados sintéticos
Coumpond	399	2	Dados sintéticos
Pathbased	300	2	Dados sintéticos
Spiral	312	2	Dados sintéticos
D31	3100	2	Dados sintéticos
R15	600	2	Dados sintéticos
Jain	373	2	Dados sintéticos
Flame	240	2	Dados sintéticos
Iris	150	4	Contém dados sobre flores do gênero Iris dividida em três classes
E Coli	336	7	Dados reais
Covid	183	99	Dados reais

3.2 Busca por similaridade via grafo de interações NK

Dado um novo objeto $\boldsymbol{x}$, desejamos encontrar os k objetos similares a $\boldsymbol{x}$. Aqui, o grafo de interações NK é utilizado para encontrar a similaridade entre objetos. O grafo de interações NK é representado por listas de adjacências na qual, para cada vértice v_i, são apresentados os (K+1) vértices com arestas incidentes a v_i.

Após a criação do grafo de interações NK, a próxima etapa é calcular a distância Euclidiana do novo objeto $\boldsymbol{x}$ para cada um dos objetos do conjunto de treinamento. O vértice relacionado ao objeto mais próximo de $\boldsymbol{x}$ é definido como v_x . Aqui são apresentadas duas

variações para a busca de similaridade baseada no grafo de interações NK. O grafo de interações é utilizado para retornar quais são os objetos mais similares ao objeto $\boldsymbol{x}$. Em ambos os métodos, o grafo de interações NK é criado com $K=k$-1, sendo k o número de objetos a ser retornado pela busca. De fato, o grafo de interações NK para os dois métodos é igual, diferindo apenas a maneira como os vértices são percorridos no grafo. Vale ressaltar que, dada uma base de treinamento, o grafo de interações NK é criado uma única vez para cada valor de k.

No método *NK A*, após a identificação do vértice inicial v_x, os K nós com arestas incidentes a v_x são identificados, i.e., retorna-se a lista de adjacências para os nós incidentes a v_x. O objeto associado a v_x e os objetos associados ao $K=k$-1 vértices com arestas incidentes a v_x são então retornados pelo método como os mais similares ao objeto $\boldsymbol{x}$.

No método *NK B*, após a identificação de v_x, encontra-se o vértice com arestas incidentes a v_x cujo objeto é mais próximo (de acordo com a distância Euclidiana) à $\boldsymbol{x}$. Então, este novo vértice é visitado e repete-se o processo até que k vértices sejam visitados. Os objetos relacionados aos vértices visitados são então retornados pelo método como os mais similares ao objeto $\boldsymbol{x}$.

3.3 Avaliação

Experimentos foram executados com diferentes valores de k (número de objetos retornados pela busca por similaridade). Na próxima seção, são apresentados experimentos para valores de k entre 1 e 14, ou seja, são retornados de 1 a 14 objetos para cada objeto consultado. De modo a avaliar os métodos, cada base de dados é dividida em conjunto de treinamento e conjunto de testes. O conjunto de testes é composto pelos objetos novos que devem ser consultados em relação à similaridade para os objetos do conjunto de treinamento. Nos métodos propostos, o conjunto de treinamento é utilizado para a criação do grafo de interações NK. No Capítulo 4, são apresentados resultados de experimentos considerando-se a validação cruzada de 10-folds para avaliação dos métodos propostos para cada k. Na validação cruzada, 9 subconjuntos (folds) do conjunto de dados são utilizados para treinamento e um para teste, i.e., contém os objetos a serem consultados. Para cada valor de k, cada algoritmo é executado 10 vezes, mudando os subconjuntos para treinamento e teste. Após finalizar toda execução é realizado uma média dos resultados para cada retorno de k. A busca por similaridade não requer que os objetos da base de dados seja rotulado. Entretanto, para fins de validação e comparação, consideramos que os métodos devem retornar exemplos da mesma classe que o exemplo a ser consultado. Por exemplo, em Medicina queremos que, quando uma imagem é consultada, os

métodos retornem imagens da mesma classe (por exemplo, mesma doença) ou do mesmo agrupamento da imagem consultada.

Assim, para cada objeto do conjunto de teste, é calculada a acurácia em relação à classe dos objetos retornados. O valor médio de acurácia para todos os objetos do conjunto de teste é então apresentado nas tabelas e figuras. A acurácia para o conjunto de teste é dada por:

$$Acc = \frac{1}{M\,k} \sum_{j=1}^{M} \sum_{i=1}^{k} hit_{(j,i)} \qquad (6)$$

sendo M o número de objetos no conjunto de teste, k o número de objetos retornados pelo método e $hit_{(j,i)}$ é igual a 1 se os rótulos do j-ésimo objeto do conjunto de teste e do i-ésimo objeto retornado pelo método são iguais e 0 caso contrário.

Por fim, para identificar se houve alguma diferença significativa entre o comportamento das amostras, foram aplicados testes estatísticos não paramétricos conhecido na literatura como teste de Wilcoxon (CONOVER, 1971), dentro de um intervalo de confiança de 95%.

Capítulo 4

4. RESULTADOS

Para testar a qualidade do nosso algoritmo de busca por similaridade realizamos dois tipos de experimentos, um quantitativo e outro qualitativo. Os experimentos foram executados em um computador Intel Core i5 2,7GHz, com 8GB de memória RAM.

4.1 Análise quantitativa

No experimento quantitativo foram comparados os dois modelos propostos e o KNN adaptado para evidenciar as diferentes abordagens para cada conjunto de banco de dados. São utilizados 8 conjuntos de dados do benchmark *Shape Sets*: *Pathbased, Spiral, Aggregation, A.K Jain's Toy Problem, R15, D31, Flame, Zahn's Compound.* além dos conjuntos *Íris e E Coli* do Repositório de Aprendizado de Máquina UCI. A seguir, os resultados são apresentados para k=1 até 14 objetos retornados. Lembrando que os resultados correspondem à validação cruzada com 10 folds, ou seja, para cada valor de k, os experimentos são repetidos 10 vezes, variando o particionamento entre os conjuntos de dados e testes.

4.1.1 Conjunto *Aggregation*

A Figura 6 apresenta a matriz de dispersão para o conjunto de dados *Aggregation*. Como pode ser observado na Tabela 2, para esse conjunto de dados, o método proposto NK B apresentou os melhores valores de acurácia para todos os experimentos com valores $k > 2$. É interessante notar que nesse conjunto de dados, a matriz de dispersão indica 7 agrupamentos bem distintos, sendo que 6 deles têm formatos circulares. Para o teste estatístico de Wilcoxon, quando comparada a distribuição NK A com NK B a maioria dos casos foram significativos, demonstrado que rejeita a hipótese H_0, ou seja, as distribuições das amostras não são iguais. Já quando comparado KNN com NK B, apesar dos melhores valores de NK B para k>2, as diferenças não foram estatisticamente significantes. Assim, apesar de o resultado de NK B quando comparados com KNN terem sido melhores, não se pode afirmar que diferenças são significativas.

.

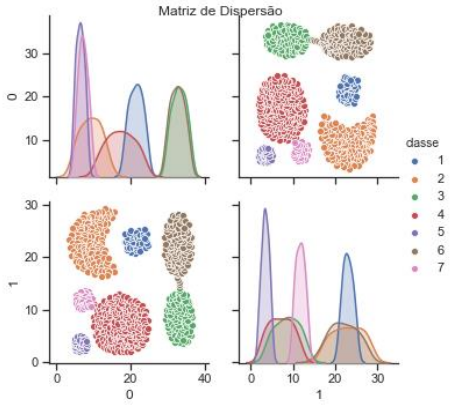

Figura 6: Matriz de dispersão do conjunto de dados *Aggregation.*

Tabela 2. Resultados expressos em média da acurácia e desvio padrão para conjuntos da base de dados *Aggregation*.

		ACURÁCIA		
		KNN	**NK A**	**NK B**
k	**1**	1.000 +- 0.000	0.990 +- 0.011	0.998 +- 0.004 #
	2	0.997 +- 0.005	0.989 +- 0.009	0.998 +- 0.004 #
	3	0.998 +- 0.004	0.992 +- 0.012	0.998 +- 0.004
	4	0.996 +- 0.007	0.992 +- 0.009	0.998 +- 0.004 #
	5	0.996 +- 0.007	0.993 +- 0.008	0.998 +- 0.004 #
	6	0.997 +- 0.005	0.993 +- 0.008	0.998 +- 0.004 #
	7	0.997 +- 0.005	0.993 +- 0.008	0.998 +- 0.004 #
	8	0.996 +- 0.007	0.993 +- 0.008	0.997 +- 0.006 #
	9	0.994 +- 0.007	0.993 +- 0.008	0.997 +- 0.006 #
	10	0.994 +- 0.007	0.993 +- 0.008	0.997 +- 0.006 #
	11	0.994 +- 0.007	0.993 +- 0.008	0.997 +- 0.006 #
	12	0.994 +- 0.007	0.993 +- 0.008	0.997 +- 0.006 #
	13	0.994 +- 0.007	0.994 +- 0.007	0.997 +- 0.006
	14	0.994 +- 0.007	0.993 +- 0.008	0.997 +- 0.006 #

* p<0,05 vs KNN ; # p<0,05 vs NK A

4.1.2 Conjunto *Compound*

Para esse conjunto de dados (Figura 7), o algoritmo KNN adaptado obteve melhores resultados para todos os valores de *k* (Tabela 3). Já em relação aos métodos NK A e B, mostra que o método B apresentou uma acurácia melhor, como pode ser verificado na Tabela 3. Para o teste estatístico de Wilcoxon, quando comparada a distribuição KNN com NK B, a maioria dos casos foram significativos, demonstrando que é rejeitada a hipótese H_0, ou seja, as distribuições das amostras não são iguais. Já quando comparado NK A com NK B, apenas alguns casos foram significativos.

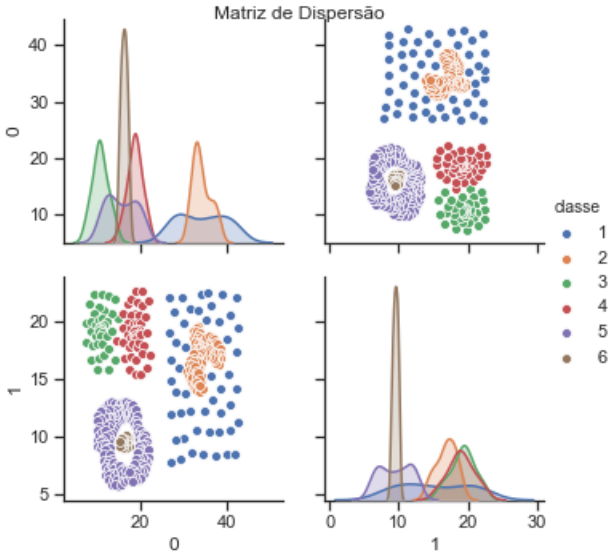

Figura 7: Matriz de dispersão do conjunto de dados *Compound.*

Tabela 3. Resultados expressos em média da acurácia e desvio padrão para conjuntos da base de dados *Compound*.

		ACURÁCIA		
		KNN	**NK A**	**NK B**
k	**1**	0.974 +- 0.024	0.927 +- 0.034	0.959 +- 0.033 *#
	2	0.973 +- 0.024	0.939 +- 0.034	0.955 +- 0.036 *#
	3	0.972 +- 0.021	0.947 +- 0.030	0.957 +- 0.031 *#
	4	0.974 +- 0.022	0.945 +- 0.029	0.952 +- 0.030 *#
	5	0.969 +- 0.024	0.943 +- 0.035	0.948 +- 0.029 *
	6	0.966 +- 0.026	0.938 +- 0.032	0.942 +- 0.032 *
	7	0.960 +- 0.024	0.935 +- 0.031	0.939 +- 0.033 *
	8	0.954 +- 0.023	0.931 +- 0.032	0.935 +- 0.031 *#
	9	0.953 +- 0.022	0.928 +- 0.034	0.932 +- 0.033 *
	10	0.946 +- 0.025	0.924 +- 0.033	0.928 +- 0.032 *
	11	0.942 +- 0.026	0.921 +- 0.032	0.925 +- 0.032 *
	12	0.936 +- 0.024	0.918 +- 0.029	0.923 +- 0.034
	13	0.931 +- 0.025	0.914 +- 0.031	0.918 +- 0.033 *
	14	0.929 +- 0.025	0.910 +- 0.031	0.914 +- 0.033 *

*p<0,05 vs KNN ; # p<0,05 vs NK A

4.1.3 Conjunto *D31*

A Figura 8 apresenta a matriz de dispersão para o conjunto de dados *D31*. Como pode ser observado na Tabela 4, no conjunto *D31* o algoritmo NK B obteve os melhores resultados. É interessante notar que nesse conjunto de dados, a frequência de registros entre as classes é próxima e sua dispersão é em forma de grupos circulares, o que facilita a detecção por algoritmos de clustering que exploram a distância entre os objetos, como pode ser verificado na Figura 8. Para o teste estatístico de Wilcoxon, quando comparada a distribuição NK A com NK B a maioria dos casos foram significativos. Já quando comparado KNN com NK B, apenas para k = 12 foi notada diferença estatisticamente significativa. Assim, apesar de os resultados de NK B quando comparados com KNN terem sido melhores, não se pode afirmar que diferenças são significativas.

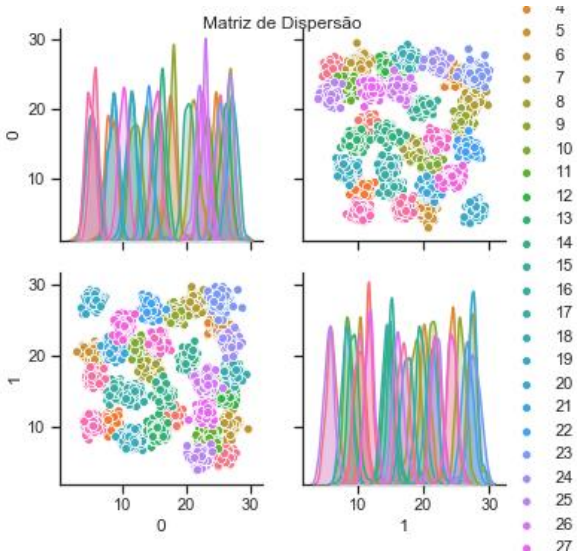

Figura 8 : Matriz de dispersão do conjunto de dados *D31*.

Tabela 4. Resultados expressos em média da acurácia e desvio padrão para conjuntos da base de dados *D31*.

		ACURÁCIA		
		KNN	**NK A**	**NK B**
k	**1**	0.960 +- 0.008	0.947 +- 0.013	0.963 +- 0.012 #
	2	0.960 +- 0.009	0.950 +- 0.012	0.965 +- 0.010 #
	3	0.961 +- 0.007	0.952 +- 0.011	0.965 +- 0.010 #
	4	0.959 +- 0.008	0.952 +- 0.011	0.964 +- 0.011 #
	5	0.960 +- 0.006	0.953 +- 0.009	0.964 +- 0.011 #
	6	0.960 +- 0.006	0.952 +- 0.011	0.963 +- 0.009 #
	7	0.958 +- 0.007	0.953 +- 0.009	0.963 +- 0.009 #
	8	0.958 +- 0.007	0.953 +- 0.009	0.963 +- 0.009 #
	9	0.958 +- 0.007	0.953 +- 0.009	0.963 +- 0.009 #
	10	0.958 +- 0.007	0.954 +- 0.009	0.963 +- 0.009 #
	11	0.958 +- 0.007	0.954 +- 0.009	0.963 +- 0.009 #
	12	0.957 +- 0.009	0.954 +- 0.009	0.963 +- 0.009 *#
	13	0.957 +- 0.011	0.953 +- 0.010	0.963 +- 0.009 #
	14	0.957 +- 0.011	0.953 +- 0.009	0.963 +- 0.009 #

*p<0,05 vs KNN ; # p<0,05 vs NK A

4.1.4 Conjunto *Flame*

Neste conjunto de dados, KNN se destaca apenas quando k é igual a 1. Para k subsequentes, NK B apresenta acurácias melhores. O método NK A foi pior em acurácia como pode ser observado na Tabela 5. Em relação ao número de registros e dispersão dos dados, apresentados na Figura 9, observa-se um volume de dados maior para classe 1 do que a 2, e dados mais centralizados. Já a classe 2 apresenta um volume menor e seus dados mais dispersos. Para o teste estatístico de Wilcoxon, apenas a distribuição de NK A com NK B para $k = 2$ foi significativo, sendo que para todos os demais, não foram observadas diferenças significativas. Assim, apesar de o resultado de NK B quando comparado com KNN ter sido melhor, não se pode afirmar que as diferenças são significativas.

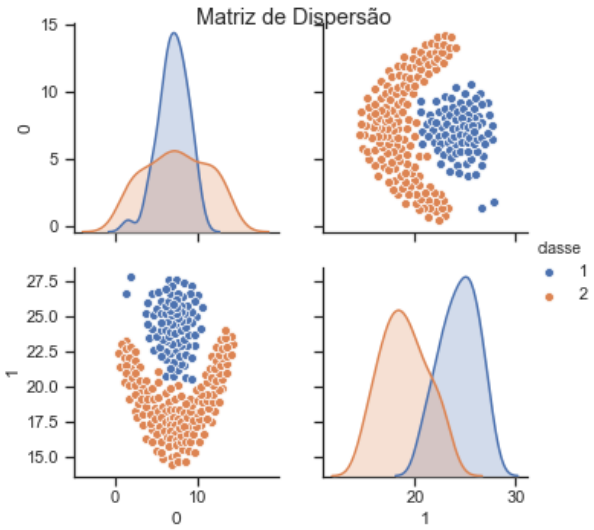

Figura 9: Matriz de dispersão do conjunto de dados Flame

Tabela 5. Resultados expressos em média da acurácia e desvio padrão para o conjuntos da base de dados *Flame*.

		ACURÁCIA		
		KNN	**NK A**	**NK B**
k	**1**	1.000 +- 0.000	0.984 +- 0.027	0.992 +- 0.016
	2	0.992 +- 0.010	0.984 +- 0.020	0.992 +- 0.016 #
	3	0.988 +- 0.010	0.985 +- 0.019	0.989 +- 0.015
	4	0.987 +- 0.009	0.987 +- 0.012	0.987 +- 0.014
	5	0.981 +- 0.011	0.984 +- 0.013	0.987 +- 0.014
	6	0.986 +- 0.010	0.986 +- 0.012	0.986 +- 0.014
	7	0.981 +- 0.011	0.984 +- 0.011	0.986 +- 0.014
	8	0.981 +- 0.011	0.982 +- 0.012	0.985 +- 0.014
	9	0.982 +- 0.013	0.984 +- 0.014	0.986 +- 0.014
	10	0.979 +- 0.014	0.980 +- 0.014	0.985 +- 0.014
	11	0.977 +- 0.013	0.980 +- 0.013	0.984 +- 0.014
	12	0.978 +- 0.012	0.978 +- 0.009	0.985 +- 0.014
	13	0.976 +- 0.012	0.977 +- 0.009	0.984 +- 0.014
	14	0.976 +- 0.012	0.977 +- 0.009	0.984 +- 0.014

*p<0,05 vs KNN ; # p<0,05 vs NK A

4.1.5 Conjunto *Jain*

Neste conjunto de dados houve um equilíbrio na acurácia dos três classificadores até *k* igual à 11 (Tabela 6). Após esse valor de *K*, apenas o método NK B apresentou acurácia de 100%, como pode ser observado na Tabela 6. Em relação ao número de registros e dispersão dos dados, apresentados na Figura 10, observa-se um volume de dados maior para classe 1 do que a 2, e ambos os dados dispersos em forma de *U*. Para o teste estatístico de Wilcoxon, não se pode afirmar que as diferenças são significativas.

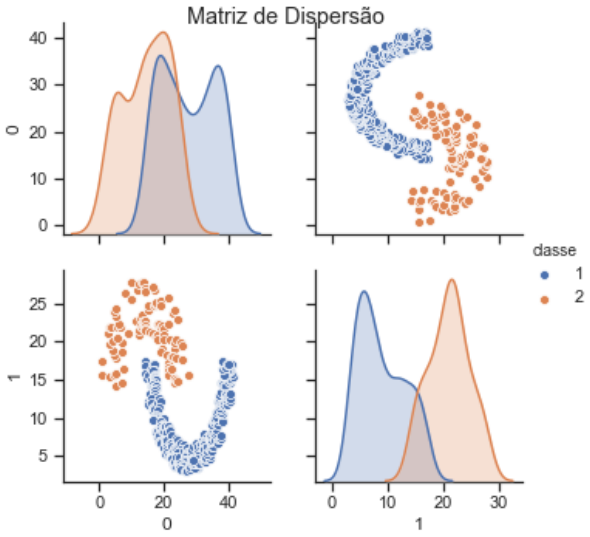

Figura 10: Matriz de dispersão do conjunto de dados Jain

Tabela 6. Resultados expressos em média da acurácia e desvio padrão para conjuntos da base de dados *Jain*.

		ACURÁCIA		
		KNN	**NK A**	**NK B**
k	**1**	1.000 +- 0.000	1.000 +- 0.000	1.000 +- 0.000
	2	1.000 +- 0.000	1.000 +- 0.000	1.000 +- 0.000
	3	1.000 +- 0.000	1.000 +- 0.000	1.000 +- 0.000
	4	1.000 +- 0.000	1.000 +- 0.000	1.000 +- 0.000
	5	1.000 +- 0.000	0.999 +- 0.003	1.000 +- 0.000
	6	1.000 +- 0.000	1.000 +- 0.000	1.000 +- 0.000
	7	1.000 +- 0.000	1.000 +- 0.000	1.000 +- 0.000
	8	1.000 +- 0.000	1.000 +- 0.000	1.000 +- 0.000
	9	1.000 +- 0.000	1.000 +- 0.000	1.000 +- 0.000
	10	1.000 +- 0.000	1.000 +- 0.000	1.000 +- 0.000
	11	1.000 +- 0.000	1.000 +- 0.000	1.000 +- 0.000
	12	1.000 +- 0.000	0.999 +- 0.003	1.000 +- 0.000
	13	1.000 +- 0.000	0.999 +- 0.003	1.000 +- 0.000
	14	0.999 +- 0.003	0.998 +- 0.004	1.000 +- 0.000

*p<0,05 vs KNN ; # p<0,05 vs NK A

4.1.6 Conjunto *Path-based*

A Figura 11 mostra a disposição de dados da base *Path-based*. Observa-se que os objetos das classes 2 e 3 formam agrupamentos agrupados dentro de círculos, enquanto que os objetos da classe 1 apresentam dispersão ao longo de uma linha formando um semicírculo. Este último cluster é difícil de ser detectado por algoritmos que utilizam apenas a distância entre os objetos. A Tabela 7 mostra o resultado de acurácia para os três modelos. Observa-se que os métodos baseados no grafo de interações NK obtiveram melhor acurácia que o método KNN adaptado. Quando os dois métodos propostos são comparados, os melhores resultados são alcançados pelo método NK B. Para o teste estatístico de Wilcoxon, quando comparada a distribuição NK A com NK B, a maioria dos casos foram significativos, demonstrando que é rejeitada a hipótese H_0, ou seja, as distribuições das amostras não são iguais. Já quando comparado KNN com NK B, observa-se diferença significativa a partir de k igual a 9.

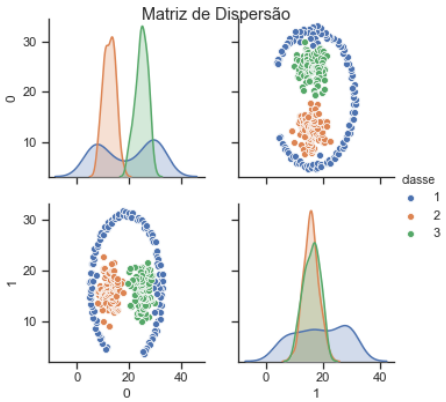

Figura 11 : Matriz de dispersão do conjunto de dados *Path-based.*

Tabela 7. Resultados expressos em média da acurácia e desvio padrão para conjuntos da base de dados *Path-based*.

		ACURÁCIA		
		KNN	**NK A**	**NK B**
k	**1**	1.000 +- 0.000	0.991 +- 0.014	1.000 +- 0.000
	2	0.994 +- 0.009	0.985 +- 0.015	0.996 +- 0.008 #
	3	0.992 +- 0.009	0.986 +- 0.012	0.992 +- 0.012 #
	4	0.989 +- 0.012	0.979 +- 0.009	0.989 +- 0.012 #
	5	0.987 +- 0.011	0.978 +- 0.010	0.989 +- 0.014 #
	6	0.982 +- 0.011	0.973 +- 0.013	0.987 +- 0.011 #
	7	0.979 +- 0.016	0.969 +- 0.014	0.988 +- 0.012 #
	8	0.970 +- 0.017	0.962 +- 0.015	0.983 +- 0.012 #
	9	0.963 +- 0.014	0.955 +- 0.019	0.984 +- 0.010 *#
	10	0.954 +- 0.016	0.951 +- 0.017	0.980 +- 0.010 *#
	11	0.945 +- 0.017	0.942 +- 0.018	0.975 +- 0.012 *#
	12	0.937 +- 0.016	0.933 +- 0.020	0.971 +- 0.009 *#
	13	0.930 +- 0.020	0.922 +- 0.021	0.967 +- 0.011 *#
	14	0.924 +- 0.017	0.916 +- 0.021	0.961 +- 0.014 *#

*p<0,05 vs KNN ; # p<0,05 vs NK A

4.1.7 Conjunto *R15*

Na Figura 12 é apresentada a matriz de dispersão do conjunto de dados *R15*. Observe que este conjunto é formado por objetos dispostos em 15 agrupamentos na forma de círculos. Em seu centro, os objetos estão mais próximos uns aos outros, enquanto ao redor apresenta-se grupos melhor separados, ou seja, a densidade é diferente ao longo do espaço de exemplos. Pode-se observar na Tabela 8 é que o método baseado no grafo de interações NK obtiveram melhor acurácia que o método KNN adaptado. Quando os dois métodos propostos são comparados, os melhores resultados são também alcançados pelo método NK B. Para o teste estatístico de Wilcoxon, quando comparada ambas as distribuições com NK B, a maioria dos casos foram estatisticamente significativos, mostrando que é rejeitada a hipótese H_0, ou seja, as distribuições das amostras não são iguais.

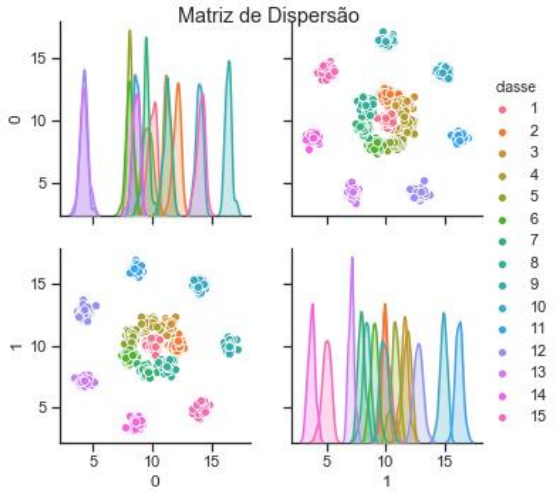

Figura 12 : Matriz de dispersão do conjunto de dados *R15*

Tabela 8: Resultados expressos em média da acurácia e desvio padrão para conjuntos da base de dados *R15*.

		ACURÁCIA		
		KNN	**NK A**	**NK B**
k	**1**	0.993 +- 0.011	0.965 +- 0.034	0.996 +- 0.008 #
	2	0.989 +- 0.009	0.972 +- 0.017	0.996 +- 0.008 *#
	3	0.990 +- 0.009	0.980 +- 0.013	0.996 +- 0.008 *#
	4	0.989 +- 0.009	0.982 +- 0.012	0.996 +- 0.008 *#
	5	0.987 +- 0.012	0.985 +- 0.011	0.996 +- 0.008 *#
	6	0.986 +- 0.013	0.984 +- 0.013	0.996 +- 0.008 *#
	7	0.988 +- 0.012	0.984 +- 0.013	0.996 +- 0.008 *#
	8	0.988 +- 0.015	0.985 +- 0.011	0.996 +- 0.008 *#
	9	0.988 +- 0.012	0.985 +- 0.011	0.996 +- 0.008 *#
	10	0.987 +- 0.012	0.984 +- 0.010	0.996 +- 0.008 *#
	11	0.987 +- 0.012	0.986 +- 0.009	0.996 +- 0.008 *#
	12	0.986 +- 0.013	0.986 +- 0.009	0.996 +- 0.008 *#
	13	0.987 +- 0.012	0.986 +- 0.009	0.996 +- 0.008 *#
	14	0.987 +- 0.012	0.987 +- 0.010	0.996 +- 0.008 *#

*p<0,05 vs KNN ; # p<0,05 vs NK A

4.1.8 Conjunto *Spiral*

Na Figura 13 é apresentada a matriz de dispersão do conjunto de dados *Spiral*. Observe que este conjunto é formado por objetos dispostos em 3 agrupamentos na forma de espirais. Observa-se na Tabela 9 que os métodos baseados no grafo de interações NK obtiveram melhor acurácia que o método KNN adaptado. Quando os dois métodos propostos são comparados, os melhores resultados são também alcançados pelo método NK B. Para o teste estatístico de Wilcoxon, quando comparada a distribuição NK A com NK B, a maioria dos casos foram estatisticamente significativos. Já quando comparado KNN com NK B, observa-se diferença estatisticamente significativa a partir de *k* igual à 7.

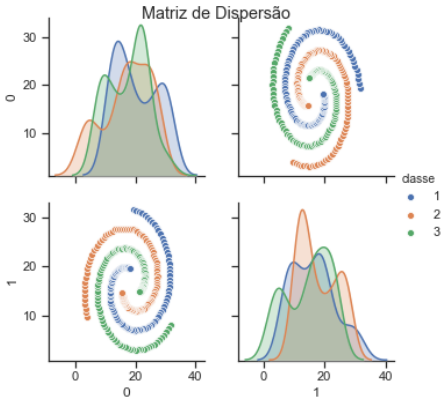

Figura 13: Matriz de dispersão do conjunto de dados *Spiral.*

Tabela 9: Resultados expressos em média da acurácia e desvio padrão para conjuntos da base de dados *Spiral*.

		ACURÁCIA		
		KNN	**NK A**	**NK B**
k	**1**	1.000 +- 0.000	0.985 +- 0.015	1.000 +- 0.000 #
	2	1.000 +- 0.000	0.990 +- 0.010	1.000 +- 0.000 #
	3	1.000 +- 0.000	0.994 +- 0.005	1.000 +- 0.000 #
	4	0.998 +- 0.004	0.992 +- 0.007	0.998 +- 0.004
	5	0.996 +- 0.005	0.987 +- 0.011	0.998 +- 0.004 #
	6	0.992 +- 0.006	0.974 +- 0.017	0.996 +- 0.007 #
	7	0.982 +- 0.010	0.963 +- 0.023	0.994 +- 0.012 *#
	8	0.971 +- 0.012	0.942 +- 0.027	0.992 +- 0.013 *#
	9	0.948 +- 0.017	0.918 +- 0.030	0.987 +- 0.016 *#
	10	0.928 +- 0.022	0.889 +- 0.030	0.987 +- 0.016 *#
	11	0.899 +- 0.025	0.862 +- 0.034	0.984 +- 0.018 *#
	12	0.873 +- 0.028	0.838 +- 0.032	0.976 +- 0.020 *#
	13	0.847 +- 0.028	0.812 +- 0.034	0.973 +- 0.021 *#
	14	0.823 +- 0.028	0.783 +- 0.029	0.965 +- 0.025 *#

*p<0,05 vs KNN ; # p<0,05 vs NK A

4.1.9 Conjunto Íris

Na Figura 10 é apresentada a matriz de dispersão do conjunto de dados *Íris*. Observe que este conjunto é formado por objetos dispostos em 3 agrupamentos. Enquanto pode-se observar que um dos agrupamentos é claramente identificado, os dois restantes apresentam sobreposição. Observa-se na Tabela 10 que o método NK B obteve melhor acurácia que o método KNN adaptado e NK A. Para o teste estatístico de Wilcoxon, quando comparada a distribuição NK A com NK B, a maioria dos casos foram significativos, demonstrando que rejeita a hipótese H_0, ou seja, as distribuições das amostras não são iguais. Já quando comparado KNN com NK B, não se pode afirmar que as diferenças são estatisticamente significantes.

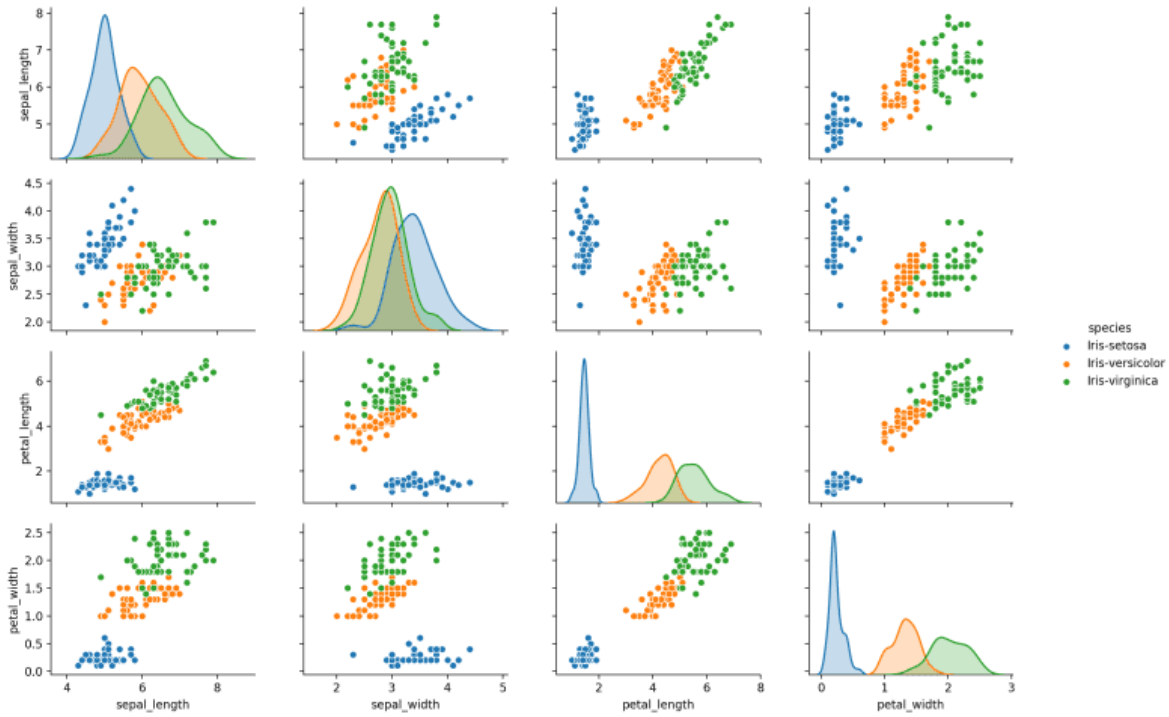

Figura 14: Matriz de dispersão do conjunto de dados Iris.

Tabela 10: Resultados expressos em média da acurácia e desvio padrão para conjuntos da base de dados *Iris*.

		ACURÁCIA		
		KNN	**NK A**	**NK B**
k	**1**	0.953 +- 0.051	0.945 +- 0.040	0.946 +- 0.049
	2	0.949 +- 0.049	0.947 +- 0.059	0.946 +- 0.049
	3	0.954 +- 0.045	0.947 +- 0.056	0.949 +- 0.049
	4	0.951 +- 0.045	0.939 +- 0.054	0.952 +- 0.049
	5	0.946 +- 0.047	0.935 +- 0.060	0.953 +- 0.049
	6	0.945 +- 0.044	0.930 +- 0.060	0.954 +- 0.049 #
	7	0.942 +- 0.046	0.928 +- 0.054	0.952 +- 0.050 #
	8	0.945 +- 0.045	0.933 +- 0.051	0.953 +- 0.049 #
	9	0.943 +- 0.048	0.925 +- 0.052	0.949 +- 0.054 #
	10	0.939 +- 0.047	0.922 +- 0.048	0.947 +- 0.056 #
	11	0.934 +- 0.051	0.915 +- 0.050	0.944 +- 0.061 #
	12	0.932 +- 0.048	0.915 +- 0.049	0.941 +- 0.063 #
	13	0.930 +- 0.046	0.913 +- 0.049	0.939 +- 0.064 #
	14	0.929 +- 0.052	0.910 +- 0.051	0.938 +- 0.066 #

*p<0,05 vs KNN ; # p<0,05 vs NK A

4.1.10 Conjunto *E Coli*

A Figura 15 apresenta a matriz de dispersão para o conjunto de dados *E Coli*. Como pode ser observado na Tabela 11, todos os métodos tiveram uma acurácia abaixo de 0.84, o que pode ser explicado pela sobreposição que existe entre os agrupamentos das diferentes classes. Entretanto, o método NK B apresentou uma acurácia maior quando comparado ao KNN. Para o teste estatístico de Wilcoxon, quando comparada ambas as distribuições com NK B, a maioria dos casos foram estatisticamente significativos.

Figura 15: Matriz de dispersão para o conjunto de dado *E Coli*.

Tabela 11: Resultados expressos em média da acurácia e desvio padrão para conjuntos da base de dados *E coli*.

		ACURÁCIA		
		KNN	**NK A**	**NK B**
k	**1**	0.808 +- 0.058	0.762 +- 0.049	0.834 +- 0.065 *
	2	0.796 +- 0.047	0.778 +- 0.040	0.838 +- 0.066 *#
	3	0.794 +- 0.043	0.779 +- 0.049	0.838 +- 0.068 *#
	4	0.795 +- 0.047	0.777 +- 0.051	0.834 +- 0.068 #
	5	0.788 +- 0.049	0.771 +- 0.054	0.834 +- 0.069 #
	6	0.789 +- 0.046	0.767 +- 0.051	0.831 +- 0.071 #
	7	0.780 +- 0.046	0.759 +- 0.047	0.828 +- 0.072 *#
	8	0.777 +- 0.046	0.765 +- 0.049	0.827 +- 0.072 *#
	9	0.777 +- 0.042	0.764 +- 0.050	0.826 +- 0.073 *#
	10	0.775 +- 0.045	0.762 +- 0.050	0.824 +- 0.074 *#
	11	0.769 +- 0.047	0.759 +- 0.049	0.823 +- 0.073 *#
	12	0.768 +- 0.045	0.755 +- 0.049	0.823 +- 0.073 *#
	13	0.767 +- 0.044	0.753 +- 0.048	0.821 +- 0.072 *#
	14	0.763 +- 0.045	0.752 +- 0.045	0.817 +- 0.072 *#

*p<0,05 vs KNN ; # p<0,05 vs NK A

4.2 Análise Qualitativa

No experimento qualitativo foi executado uma consulta por similaridade utilizando o método baseado no grafo de interações NK que obteve os melhores resultados na análise quantitativa, o método NK B, com o método KNN adaptado na base Covid. Para executar a busca por similaridade, é necessário fornecer a imagem de referência a ser utilizada. Através da imagem de referência, executa-se uma busca por imagens similares.

A imagem de referência utilizada para as consultas foi no formato .jpeg. Foi utilizado o valor de K igual à 5 para o total de imagens mais semelhantes que se deseja recuperar. A consulta é rápida, levando menos de 1 segundo para ter o resultado exibido. Após o processamento, as imagens retornadas são mostradas em forma de *thumbnails*. Os resultados visualizados aqui, possuem como primeira imagem a imagem de referência, destacada com um

quadrado vermelho, e na sequência são listadas as imagens em ordem retornada pelo algoritmo. Aqui são também apresentadas, para cada busca, uma tabela com os descritivos de cada imagem retornada, nome do arquivo, gênero, idade, local de origem da imagem. Todas estas informações constam na COVID-19 Image Data Collection, que foi utilizada para a construção da base Covid. As Figuras 16-19 e as Tabelas 12-19 mostram os resultados obtidos pelos algoritmos para 4 imagens da base escolhidas aleatoriamente.

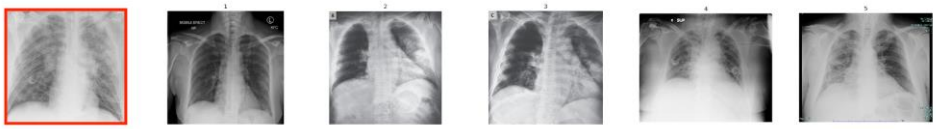

Figura 16: Imagens resultante da consulta por similaridade para o primeiro exemplo - consulta NK B.

Figura 17: Imagens resultante da consulta por similaridade para o primeiro exemplo - consulta KNN.

Tabela 12: Tabela contendo as informações de cada imagem resultante da consulta por similaridade NK B para o primeiro exemplo. A imagem consultada é apresentada na primeira linha (0).

	nome do arquivo	sexo	idade	localização
0	B2D20576-00B7-4519-A415-72DE29C90C34	M	60.0	Italy
1	41591_2020_819_Fig1_HTML.webp-day10	F	NaN	The Royal Melbourne Hospital, Melbourne, Austr..
2	auntminnie-b-2020_01_28_23_51_6665_2020_01_28_...	M	65.0	Cho Ray Hospital, Ho Chi Minh City, Vietnam
3	auntminnie-c-2020_01_28_23_51_6665_2020_01_28_...	M	65.0	Cho Ray Hospital, Ho Chi Minh City, Vietnam
4	85E52EB3-56E9-4D67-82DA-DEA247C82886	F	78.0	Italy
5	80446565-E090-4187-A031-9D3CEAA586C8	M	73.0	Italy

Tabela 13: Tabela contendo as informações de cada imagem resultante da consulta por similaridade KNN adaptado para o primeiro exemplo.

	nome do arquivo	sexo	idade	localização
0	B2D20576-00B7-4519-A415-72DE29C90C34	M	60.0	Italy
1	41591_2020_819_Fig1_HTML.webp-day10	F	NaN	The Royal Melbourne Hospital, Melbourne, Austr...
2	auntminnie-c-2020_01_28_23_51_6665_2020_01_28_...	M	65.0	Cho Ray Hospital, Ho Chi Minh City, Vietnam
3	85E52EB3-56E9-4D67-82DA-DEA247C82886	F	78.0	Italy
4	auntminnie-b-2020_01_28_23_51_6665_2020_01_28_...	M	65.0	Cho Ray Hospital, Ho Chi Minh City, Vietnam
5	F63AB6CE-1968-4154-A70F-913AF154F53D	F	58.0	Italy

Figura 18: Imagens resultante da consulta por similaridade para o segundo exemplo - consulta NK B.

Figura 19: Imagens resultante da consulta por similaridade para o segundo exemplo - consulta KNN.

Tabela 14: Tabela contendo as informações de cada imagem resultante da consulta por similaridade NK B para o segundo exemplo.

	nome do arquivo	sexo	idade	localização
0	03BF7561-A9BA-4C3C-B8A0-D3E585F73F3C	F	65.0	Italy
1	FE9F9A5D-2830-46F9-851B-1FF4534959BE	M	65.0	Italy
2	nejmc2001573_f1a	F	52.0	Changhua Christian Hospital, Changhua City, Ta...
3	1312A392-67A3-4EBF-9319-810CF6DA5EF6	M	50.0	Italy
4	ryct.2020200034.fig2	F	NaN	Hong Kong
5	1-s2.0-S1684118220300608-main.pdf-002	F	46.0	Taiwan

Tabela 15: Tabela contendo as informações de cada imagem resultante da consulta por similaridade KNN adaptado para o segundo exemplo.

	nome do arquivo	sexo	idade	localização
0	03BF7561-A9BA-4C3C-B8A0-D3E585F73F3C	F	65.0	Italy
1	FE9F9A5D-2830-46F9-851B-1FF4534959BE	M	65.0	Italy
2	nejmc2001573_f1a	F	52.0	Changhua Christian Hospital, Changhua City, Ta...
3	1312A392-67A3-4EBF-9319-810CF6DA5EF6	M	50.0	Italy
4	ryct.2020200034.fig2	F	NaN	Hong Kong
5	1-s2.0-S1684118220300608-main.pdf-002	F	46.0	Taiwan

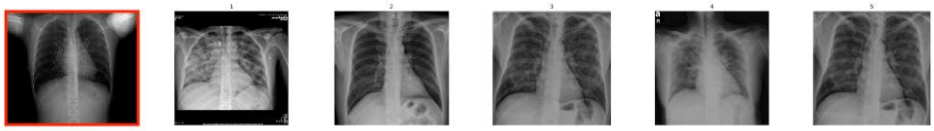

Figura 20: Imagens resultante da consulta por similaridade para o terceiro exemplo - consulta NK B.

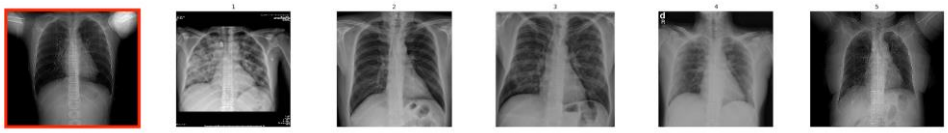

Figura 21: Imagens resultante da consulta por similaridade para o terceiro exemplo - consulta KNN.

Tabela 16: Tabela contendo as informações de cada imagem resultante da consulta por similaridade NK B para o terceiro exemplo.

	nome do arquivo	sexo	idade	localização
0	ciaa199.pdf-001-c	M	38.0	China
1	F2DE909F-E19C-4900-92F5-8F435B031AC6	M	58.0	Italy
2	7C69C012-7479-493F-8722-ABC29C60A2DD	M	62.0	Italy
3	F4341CE7-73C9-45C6-99C8-8567A5484B63	M	47.0	Italy
4	1-s2.0-S0929664620300449-gr2_lrg-a	F	55.0	Taoyuan General Hospital, Taoyuan, Taiwan
5	E63574A7-4188-4C8D-8D17-9D67A18A1AFA	M	47.0	Italy

Tabela 17: Tabela contendo as informações de cada imagem resultante da consulta por similaridade KNN adaptado para o terceiro exemplo.

	nome do arquivo	sexo	idade	localização
0	ciaa199.pdf-001-c	M	38.0	China
1	F2DE909F-E19C-4900-92F5-8F435B031AC6	M	58.0	Italy
2	7C69C012-7479-493F-8722-ABC29C60A2DD	M	62.0	Italy
3	F4341CE7-73C9-45C6-99C8-8567A5484B63	M	47.0	Italy
4	1-s2.0-S0929664620300449-gr2_lrg-d	F	55.0	Taoyuan General Hospital, Taoyuan, Taiwan
5	ciaa199.pdf-001-a	M	38.0	China

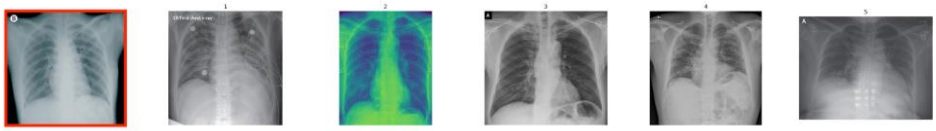

Figura 22: Imagens resultante da consulta por similaridade para o quarto exemplo- consulta NK B.

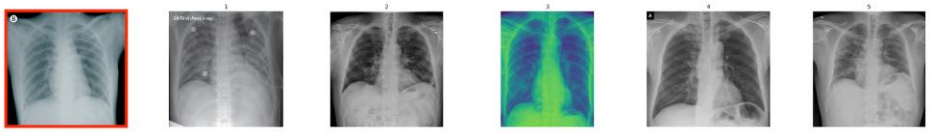

Figura 23: Imagens resultante da consulta por similaridade para o quarto exemplo- consulta KNN.

Tabela 18: Tabela contendo as informações de cada imagem resultante da consulta por similaridade NK B para o quarto exemplo.

	nome do arquivo	sexo	idade	localização
0	gr1_lrg-b	F	25.0	Thanh Hóa, Vietnam
1	lancet-case2a	NaN	NaN	Wuhan Jinyintan Hospital, Wuhan, Hubei Provinc...
2	ryct.2020200028.fig1a	F	59.0	Sichuan Provincial People's Hospital, Chengdu,...
3	ryct.2020200034.fig2	F	NaN	Hong Kong
4	1312A392-67A3-4EBF-9319-810CF6DA5EF6	M	50.0	Italy
5	nejmc2001573_f1a	F	52.0	Changhua Christian Hospital, Changhua City, Ta...

Tabela 19: Tabela contendo as informações de cada imagem resultante da consulta por similaridade KNN adaptado para o quarto exemplo.

	nome do arquivo	sexo	idade	localização
0	gr1_lrg-b	F	25.0	Thanh Hóa, Vietnam
1	lancet-case2a	NaN	NaN	Wuhan Jinyintan Hospital, Wuhan, Hubei Provinc...
2	9C34AF49-E589-44D5-92D3-168B3B04E4A6	M	67.0	Italy
3	ryct.2020200028.fig1a	F	59.0	Sichuan Provincial People's Hospital, Chengdu,...
4	ryct.2020200034.fig2	F	NaN	Hong Kong
5	1312A392-67A3-4EBF-9319-810CF6DA5EF6	M	50.0	Italy

Capítulo 5

5. ANÁLISE E CONCLUSÕES

Os resultados mostram que, ao utilizar a densidade, os métodos propostos permitiram retornar objetos dispostos nos agrupamentos que não são necessariamente hiper-esféricos. Ao usar apenas a distância, o KNN adaptado não foi capaz de explorar a disposição de tais agrupamentos. Os melhores resultados do KNN foram para o conjunto *Aggregation*, que possui agrupamentos compactos bem definidos. Além disso, a base de dados é maior que as outras, o que impacta a amostragem dos dados das bases de treinamento e teste.

Observa-se que, para as duas primeiras bases, conforme o número de objetos retornados (*k*) cresce, mais objetos de classes diferentes são retornados, ou seja, a acurácia diminui. Entretanto, o impacto de *k* foi mais significativo para o KNN adaptado que nos métodos propostos. Em geral, diminuir o tamanho do conjunto de treinamento também implicou em diminuir a acurácia dos métodos. Como os mesmos valores de *k* foram testados, o número de erros aumentou para um conjunto menor de dados de treinamento. Os melhores resultados foram alcançados pelo método NK B, que percorre o grafo de interações sempre alcançando o vértice, com aresta incidente, relacionado ao objeto mais próximo do novo objeto a ser consultado. O método NK A não leva em consideração a distância para o novo objeto consultado depois que o vértice inicial é visitado. O uso da distância para o novo objeto mostrou-se útil na busca por similaridade utilizando-se o grafo de interações.

Os métodos propostos se mostraram interessante para as bases de dados de formato arbitrário, ou seja, não necessariamente hiper esféricos. Isso ocorre porque os métodos levam em consideração tanto a distância entre os objetos como também a densidade local deles.

Em relação às análises qualitativas com banco de dados do Covid, foram observados alguns pontos interessantes. Tivemos retorno de pessoas com idade e sexo semelhantes, e as imagens retornadas de cada algoritmo também foram diferentes, evidenciando a diferença de busca por similaridade de cada método. Entretanto, na análise qualitativa feita, é difícil destacar qual método teve um melhor desempenho. Temos que levar em consideração que para uma melhor análise dessas imagens é necessário um profissional capacitado em radiologia (médicos radiologistas). Estes resultados abrem portas para trabalhos futuros, onde poderemos utilizar outras bases de dados da área médica com colaboração de profissionais da área em questão.

Além disso um trabalho futuro é investigar se o uso de um algoritmo de agrupamento antes da busca por similaridade pode melhorar o desempenho da acurácia. Contendo as informações de agrupamentos, a busca pode ser restrita aos objetos do cluster mais próximo ao objeto consultado, a ponto de ter uma diferença significativa. Por fim, o método de busca por

similaridade NK B vem se demonstrando promissor visto que demonstrou bom desempenho nos conjuntos de dados aplicados.

6. REFERÊNCIAS

AHA, D. W.; KIBLER, D. & ALBERT, M.K. (1991). “Instance-based learning algorithms”, *Machine Learning*, 6(1): 37-66.

BERGADANO, FRANCCESCO, AND LUC DE RAEDT. "*Machine Learning*: ECML-94: *European Conference on Machine Learning, Catania*", *Italy, April 6-8, (1994). Proceedings.* Vol. 784. Springer Science & Business Media, 1994. p. 49-63.

CARPINETO, C. & ROMANO, G. (2012). "A survey of automatic query expansion in information retrieval", *ACM Computing Surveys (CSUR)*, 44(1).

CHA, SUNG-HYUK. (2007) "Comprehensive survey on distance/similarity measures between probability density functions.: City, v. 1, n. 2, p. 1.

CONOVER, W. J. (1971). The Wilcoxon signed rank test. Practical nonparametric statistics. Wiley, New York, 206-216.

DEL LAMA, R. S. (2020). "Algoritmos Genéticos e Redes Neurais Convolucionais para Auxílio ao Diagnóstico de Fraturas Vertebrais por Compressão", Dissertação de Mestrado Apresentada ao Programa de Mestrado em Computação Aplicada, FFCLRP, USP.

DUA, D. & GRAFF, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

ESTER, M.; KRIEGEL, H.-P.; SANDER, J. & XU, X. (1996), “A density-based algorithm for discovering clusters in large spatial databases with noise”, *In the Proc. of the 2nd ACM Int. Conf. Knowl. Discovery Data Min. (KDD)*, 226–231.

ESTAS SÃO AS 10 LINGUAGENS DE PROGRAMAÇÃO MAIS POPULARES ATUALMENTE. COMPUTER WORLD. (2020). Disponível em: https://computerworld.com.br/carreira/estas-sao-as-10-linguagens-de-programacao-mais-populares-atualmente/. Acesso em: 09 de outubro de 2020.

FERREIRA, M. R. P.; SANTOS, L. F. D.; TRAINA, A. J. M.; DIAS, I.; CHABEIR, R.; TRAINA JR., C. (2011). “Algebraic properties to optimize KNN queries”, *Journal of Information and Data Management*, 2(3): 385–400.

FIGUEIREDO, DANIEL R.(2011). "Introdução a redes complexas". Atualizações em Informática, p. 303-358.

FRÄNTI, P. & SIERANOJA, S. (2018). “K-means properties on six clustering benchmark datasets”, *Applied Intelligence*, 48 (12), 4743-4759.

GUIA QUAL A MELHOR LINGUAGEM PARA CIÊNCIA DE DADOS. Disponível em: https://blog.geekhunter.com.br/guia-qual-a-melhor-linguagem-para-ciencia-de-dados/. Acesso em: 09 de outubro de 2020.

HRUSCHKA, E. R.; CAMPELLO, R. J. G. B.; FREITAS, A. A. & CARVALHO, A. C. P. L. F. (2009). "A survey of evolutionary algorithms for clustering", *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 39(2): 133-155.

HSINCHUN, C.; CHIANG, R. H. L. & STOREY, V. C. (2012). "Business intelligence and analytics: from big data to big impact", *MIS Quarterly*, 36(4): 1165-1188.

HUANG, ANNA. (2008). "Similarity measures for text document clustering". *Proceedings of the sixth new zealand computer science research student conference, Christchurch, New Zealand*. p. 49-56

J. J. VAN GRIETHUYSEN, A. FEDOROV, C. PARMAR, A. HOSNY, N. AUCOIN, V. NARAYAN, R. G, (2017). "Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. Aerts. Computational radiomics system to decode the radiographic phenotype". *Cancer research*, 77(21):e104–e107.

RODRIGUEZ, A. & LAIO, A. (2014). "Clustering by fast search and find of density peaks," *Science*, 344(6191): 1492–1496.

S. A. KAUFFMAN. (1993). "*The Origins of Order: Self-Organization and Selection in Evolution"*. New York, NY, USA: Oxford Univ. Press.

TEKNOMO, KARDI. (2015). "Similarity Measurement", Disponível em: http://people.revoledu.com/kardi/tutorial/Similarity. Acessado em: 09 de outubro de 2020.

TINÓS, R.; ZHAO, L.; CHICANO, F. & WHITLEY, D. (2018). "NK hybrid genetic algorithm for clustering", *IEEE Transactions on Evolutionary Computation*, 22(5): 748-761.

TINÓS, R.; ZHAO, L.; CHICANO, F. & WHITLEY, D. "A new evaluation function for clustering: The NK internal validation criterion". *In Proceedings of the Genetic and Evolutionary Computation Conference 2016 (GECCO '16)*, Denver, USA, p. 509-516.

XU, D. & TIAN, YA. (2015). "Comprehensive Survey of Clustering Algorithms", *Annals of Data Science*, 2: 165-193.

Apêndice

Artigo aceito no XVI Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)

Similarity Search using the NK Interaction Graph

José Carlos Bueno de Moraes, Renato Tinós

Departamento de Computação e Matemática, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto (FFCLRP), Universidade de São Paulo (USP), Ribeirão Preto, SP, Brasil

{josecbmoraes@gmail.com ,rtinos@ffclrp.usp.br}

Abstract: *A similarity search method based on the NK interaction graph is proposed. The NK interaction graph was originally employed for clustering and is built based on distance and spatial density of the objects in a dataset. Two variations of the method are investigated. In the two variations, k objects are returned by visiting vertices of the NK interaction graph from the initial vertex related to the example of the dataset that is closer to the object to be consulted. In NK A, the k objects related to vertices with edges incident to the initial vertex are returned. In NK B, k vertices are visited starting from the initial vertex. The next visited vertex is that one with edge incident to the current vertex and that is closest to the new object to be consulted. The k objects related to the visited vertices are returned. The proposed algorithms are compared with each other and with the search for similarity based only on distance. The experimental results indicate that the proposed methods present good performance when there are clusters with arbitrary shapes in the dataset.*

Resumo: *Um método de busca por similaridade baseado no grafo de interações NK é proposto. O grafo de interações NK foi originalmente empregado para agrupamento e é construído com base na distância e densidade espacial dos objetos em um conjunto de dados. Duas variações do método são investigadas. Em NK A, os k objetos relacionados a vértices com arestas incidentes ao vértice inicial são retornados. Em NK B, k vértices são visitados a partir do vértice inicial. O próximo vértice visitado é aquele com aresta incidente ao vértice atual e que está mais próximo do novo objeto a ser consultado. Os k objetos relacionados aos vértices visitados são retornados. Os algoritmos propostos são comparados entre si e com a busca por similaridade baseada apenas na distância. Os resultados experimentais indicam que os métodos propostos apresentam bom desempenho quando existem clusters com formas arbitrárias no conjunto de dados.*

1. Introdução

Com o crescimento do volume de dados ao longo dos anos, foram desenvolvidas técnicas de busca de similaridade para responder às necessidades dos usuários em diversos segmentos do conhecimento **[**HSINCHUN *et al*., 2012**]**. A evolução das técnicas de busca de similaridade vem permitindo recuperar objetos presentes em grandes bases de dados similares a um objeto fornecido pelo usuário de maneira eficiente, auxiliando na tomada de decisão em diversas aplicações. Por exemplo, na área da Medicina, busca por similaridade de exames (como imagens médicas, exames laboratoriais, entre outros) e laudos tem potencial para aumentar a

eficiência das decisões médicas, reduzir custos e otimizar o tempo dos especialistas na análise de casos **[**CARPINETO & ROMANO, 2012**]**.

Dentre as técnicas de aprendizado de máquina supervisionada, a técnica mais comumente utilizada de busca por similaridade é aquela baseada em distância. O algoritmo dos *K*-vizinhos próximos (*k-nearest neighbours - KNN*) [AHA *et al*., 1991] pode ser adaptado para retornar os *k* objetos de uma base de treinamento mais similares ao objeto que está sendo consultado. No entanto, as informações sobre densidade espacial de objetos não são consideradas quando apenas *KNN* é empregado. Informações adicionais sobre densidade espacial podem ser úteis especialmente em bases de dados com agrupamentos com formas arbitrárias. A densidade espacial de objetos é explorada por algumas técnicas de clusterização para, entre outros, produzir agrupamentos que não são necessariamente hiper-esféricos [TINÓS *et al*., 2018; RODRIGUEZ & LAIO, 2014; ESTER *et al*., 1996] . Técnicas de clusterização têm sido aplicadas nas mais diversas áreas do conhecimento [HRUSCHKA *et al*., 2009]. Conceitos utilizados em clusterização podem ser especialmente úteis na recuperação de informação e na visualização de dados.

Em [TINÓS *et al*., 2018], o NKGA (*NK Hybrid Genetic Algorithm*) foi proposto para o problema de clustering. O NKGA utiliza tanto a distância entre objetos como a densidade espacial para o agrupamento de objetos. Para avaliar a soluções (particionamentos dos objetos da base de dados), o NKGA usa uma função de validação interna chamada NKCV2. Esta função utiliza informações sobre a disposição de *N* pequenos grupos de objetos, sendo *N* o número de objetos na base de dados. Cada grupo é composto de *K*+1 objetos, sendo *K* um parâmetro definido pelo usuário. As informações sobre os grupos de objetos são capturadas no *grafo de interações NK*. Tanto informações sobre densidade como de distância entre objetos são utilizadas para construir o grafo de interações NK. Resultados experimentais mostram que agrupamentos de dados com formas arbitrárias podem ser identificados usando NKGA com *K* pequeno.

Neste trabalho, propomos um método de busca por similaridade baseado no grafo de interações NK. Duas variações do método são investigadas. Nas duas variações, *k* objetos são retornados percorrendo-se o grafo de interações NK a partir do vértice inicial v_x relacionado ao objeto da base de dados mais similar ao objeto **x** a ser consultado. Na primeira variação (*método NK A*), *k*=*K*+1, sendo que os *k* objetos cujos vértices têm arestas incidentes no vértice v_x são retornados. Na segunda variação (*método NK B*), caminha-se a partir de v_x sempre alcançando o vértice, com aresta incidente, cujo objeto é mais próximo ao objeto consultado. Após *k* passos, *k* objetos são então retornados: aqueles relacionados aos *k* vértices visitados. Os algoritmos propostos são comparados entre si e com a busca por similaridade baseada em distância (chamada aqui, por simplicidade, de KNN adaptado).

2. Metodologia

Os métodos propostos são aqui comparados ao KNN adaptado, no qual, ao invés de utilizar KNN para classificar novos objetos a partir da distância para objetos de uma base de dados (também chamada de base de treinamento), utiliza-se a distância para cálculo da dissimilaridade e são retornados os *k* objetos mais próximos ao objeto consultado. Os métodos propostos aqui são baseados no grafo de interações NK, que é descrito a seguir.

2.1 Grafo de Interações NK

O grafo de interações NK é um grafo direcionado com N vértices, cada um com grau de entrada K+1. Dada uma base de dados (treinamento) com N objetos n-dimensionais, o primeiro passo para a construção do grafo é adicionar vértices v_i , i = 1,...,N, para cada objeto $\mathbf{y}_i$ da base de dados. Cada vértice possui auto-loop, que aqui poderia ser ignorado. Se ignorarmos os auto-loops, o grau de entrada para cada vértice é então igual a K; entretanto, para fins de uniformidade com [TINÓS *et al.*, 2018], o auto-loop será preservado na descrição do grafo, apesar de não ser utilizado pela busca por similaridade. Cada aresta (v_j , v_i) indica que o j-ésimo objeto é relacionado com o i-ésimo objeto.

É importante ressaltar que a construção das arestas leva em consideração a distância Euclidiana para objetos próximos e a densidade dos objetos. Após a criação dos vértices com auto-loop, a densidade dos objetos é calculada. Para o i-ésimo objeto, a densidade é dada por:

$$\rho_i = \sum_{j=1}^{N} \mathbf{K}(\mathbf{y}_i - \mathbf{y}_j) \tag{1}$$

sendo **K**(.) a função Kernel dada por:

$$\mathbf{K}(\mathbf{y}_i - \mathbf{y}_j) = e^{\frac{-\|\mathbf{y}_i - \mathbf{y}_j\|^2}{2\epsilon^2}} \tag{2}$$

sendo ϵ o parâmetro que define a distância de corte. Aqui, ϵ é escolhido de modo que o número médio de vizinhos de um objeto seja 2% do total de objetos da base de treinamento [RODRIGUEZ & LAIO, 2014; TINÓS *et al.*, 2018].

Para cada vértice v_i , o vértice v_{ai} representando o objeto mais próximo que possui densidade maior que o objeto relacionado a v_i é identificado. Então, uma aresta (v_{ai}, v_i) é criada. O último passo é adicionar arestas para os vértices representando objetos mais próximos até que o grau de entrada de cada vértice seja igual a K+1. A Figura 1 apresenta um exemplo do processo de criação do grafo de interações NK.

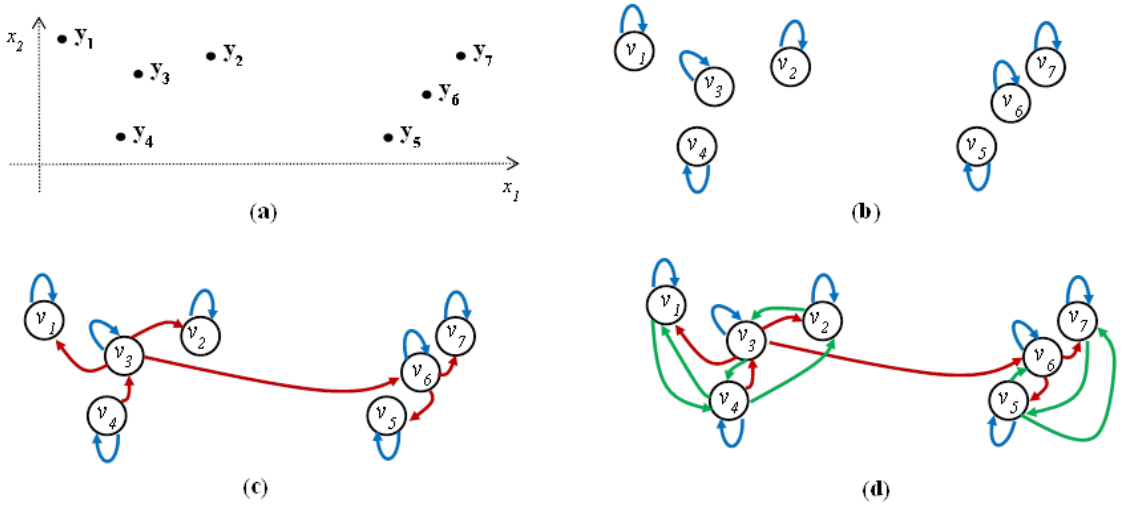

Figura 1: Exemplo de construção do grafo de interações NK com K = 2 para um conjunto com 7 objetos bidimensionais (N = 7, n = 2). Cada objeto da base de dados (a) é associado com um vértice com auto-

loop (b). A densidade dos objetos é calculada e cada vértice é ligado ao vértice associado com o objeto mais próximo com maior densidade (c). O próximo passo é adicionar arestas para os vértices representando objetos mais próximos até que o grau de entrada de cada vértice seja igual a K+1. O gráfico de interações (d) tem N = 7 vértices e $N(K+1)$ arestas.

2.2 Busca por Similaridade via Grafo de Interações NK

Dado um novo objeto **x**, desejamos encontrar os k objetos similares a **x**. Aqui, o grafo de interações NK é utilizado para encontrar a similaridade entre objetos. O grafo de interações NK é representado por uma matriz de adjacências na qual, para cada vértice v_i , são apresentados os (K+1) vértices com arestas incidentes a v_i .

Após a criação do grafo de interações NK, a próxima etapa é calcular a distância Euclidiana do novo objeto **x** para cada um dos objetos do conjunto de treinamento. O vértice relacionado ao objeto mais próximo de **x** é definido como v_x . Aqui são apresentados duas variações para a busca de similaridade baseada no grafo de interações NK. O grafo de interações é utilizado para retornar quais são os objetos mais similares ao objeto **x**. Em ambos os métodos, o grafo de interações NK é criado com $K=k$-1, sendo k o número de objetos a ser retornado pelo método. De fato, o grafo de interações NK para os dois métodos é igual, diferindo apenas a maneira como os vértices são percorridos no grafo. Vale ressaltar que, dada uma base de treinamento, o grafo de interações NK é criado uma única vez para cada valor de k.

No método *NK A*, após a identificação do vértice inicial v_x, os K nós com arestas incidentes a v_x são identificados, i.e., retorna-se a lista de adjacências para os nós incidentes a v_x. O objeto associado a v_x e os objetos associados ao $K=k$-1 vértices com arestas incidentes a v_x são então retornados pelo método como os mais similares ao objeto **x**.

No método *NK B*, após a identificação de v_x, encontra-se o vértice com arestas incidentes a v_x cujo objeto é mais próximo (de acordo com a distância Euclidiana) à **x**. Então, este novo vértice é visitado e repete-se o processo até que k vértices sejam visitados. Os objetos relacionados aos vértices visitados são então retornados pelo método como os mais similares ao objeto **x**.

2.3 Avaliação

Experimentos foram executados com diferentes valores de k. Na próxima seção, são apresentados valores de k entre 1 e 14, ou seja, são retornados de 1 a 14 objetos para cada objeto consultado. De modo a avaliar os métodos, cada base de dados é dividida em conjunto de treinamento e conjunto de testes. O conjunto de testes é composto pelos objetos novos que devem ser consultados em relação à similaridade para os objetos do conjunto de treinamento. Nos métodos propostos, o conjunto de treinamento é utilizado para a criação do grafo de interações NK. Na próxima seção, são realizados experimentos em que o conjunto de treinamento é composto por 90% ou 95% dos exemplos da base de dados. O restante dos dados compõe o conjunto de teste. A busca por similaridade não requer que os objetos da base de dados sejam rotulados. Entretanto, para fins de validação e comparação, consideramos que os métodos devem retornar exemplos da mesma classe que o exemplo a ser consultado. Por exemplo, em Medicina queremos que, quando uma imagem é consultada, os métodos retornem imagens da mesma classe (por exemplo, mesma doença) ou do mesmo agrupamento da imagem consultada.

Assim, para cada objeto do conjunto de teste, é calculada a acurácia em relação à classe dos objetos retornados. O valor total de acurácia para todos os objetos do conjunto de teste é então apresentado nas tabelas e figuras. A acurácia para o conjunto de teste é dada por:

$$Acc = \frac{1}{Mk} \sum_{j=1}^{M} \quad \sum_{i=1}^{k} \quad hit_{(j,i)} \tag{3}$$

sendo M o número de objetos no conjunto de teste, k o número de objetos retornados pelo método e $hit_{(j,i)}$ é igual a 1 se os rótulos do j-ésimo objeto do conjunto de teste e do i-ésimo objeto retornado pelo método são iguais e 0 caso contrário.

3. Resultados

Os experimentos descritos a seguir comparam os dois modelos propostos e o KNN adaptado para três conjunto de dados do benchmark *Shape Sets* [FRÄNTI & SIERANOJA, 2018]: *Pathbased*, *Spiral* e *Aggregation.* Todos os conjuntos têm dimensão n=2, o que facilita a visualização da disposição dos objetos. Os dois primeiros conjuntos possuem agrupamentos que são difíceis de serem detectados por algoritmos de clustering que utilizam apenas a distância entre os objetos para o particionamento, como o *k-means*. Já o conjunto *Aggregation* possui apenas agrupamentos que são mais fáceis de serem detectados por tais algoritmos. *Pathbased* possui 300 objetos e *Spiral* possui 312 objetos, ambos com 3 clusters cada (figuras 2 e 3). Já *Aggregation* possui 788 objetos com 7 clusters. Os experimentos foram executados em um computador Intel Core i5 2,7GHz, com 8GB de memória RAM.

3.1 Base de dados Pathbased

A figura 2 mostra a disposição de dados da base *Pathbased*. Observa-se que os objetos das classes 2 e 3 formam clusters agrupados dentro de círculos, enquanto que os objetos da classe 1 apresentam dispersão ao longo de uma linha formando um semicírculo. As tabelas 1 e 2 mostram os resultados de acurácia para, respectivamente, conjuntos de testes com 5% e 10% de objetos da base de dados. São mostrados os resultados para diversos valores de *k*. Observa-se em ambas as tabelas que os métodos baseados no grafo de interações NK obtiveram melhor acurácia que o método KNN adaptado. Quando os dois métodos propostos são comparados, os melhores resultados são alcançados pelo método NK B.

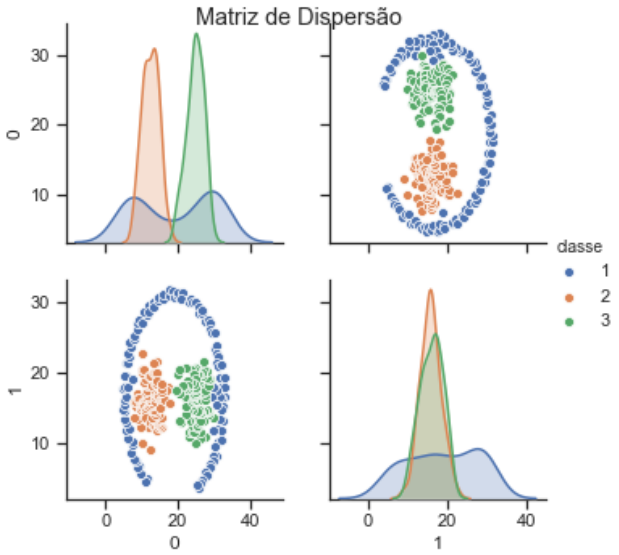

Figura 2: Matriz de dispersão do conjunto de dados *Pathbased.*

Tabela 1. Resultados para conjuntos de teste com 5% da base de dados *Pathbased*.

		k													
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Acc	**KNN**	1.00	1.00	0.98	0.97	0.93	0.93	0.92	0.93	0.93	0.91	0.90	0.90	0.88	0.87
	NK A	1.00	1.00	1.00	1.00	0.99	0.97	0.96	0.95	0.96	0.95	0.95	0.93	0.93	0.92
	NK B	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.98	0.98	0.97	0.97

Tabela 2. Resultados para conjuntos de teste com 10% da base de dados *Pathbased*.

		k													
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Acc	**KNN**	1.00	0.97	0.97	0.95	0.93	0.93	0.92	0.93	0.93	0.91	0.91	0.90	0.89	0.89
	NK A	1.00	1.00	1.00	0.99	0.95	0.98	0.95	0.97	0.96	0.96	0.95	0.94	0.94	0.93
	NK B	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99

3.2 Conjunto Spiral

Na Figura 3 é apresentada a matriz de dispersão do conjunto de dados *Spiral*. Observe que este conjunto é formado por objetos dispostos em 3 clusters na formas de espirais. As tabelas 3 e 4 mostram os resultados de acurácia para, respectivamente, conjuntos de testes com 5% e 10%

da base de dados. Observa-se novamente em ambas as tabelas que os métodos baseados no grafo de interações NK obtiveram melhor acurácia que o método KNN adaptado. Quando os dois métodos propostos são comparados, os melhores resultados são também alcançados pelo método NK B.

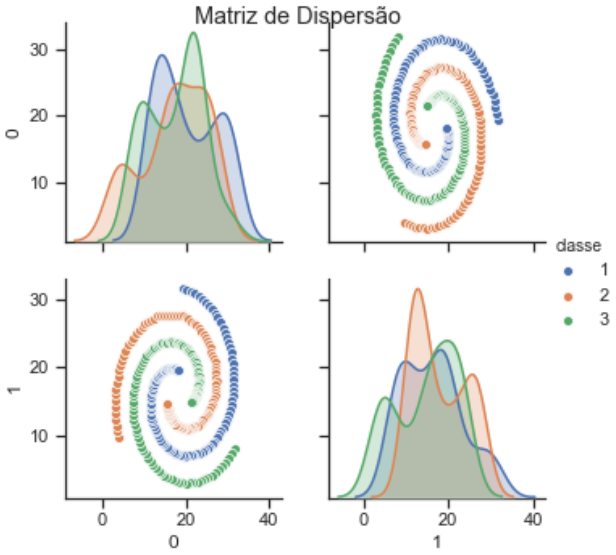

Figura 3: Matriz de dispersão do conjunto de dados *Spiral*.

Tabela 3. Resultados para conjuntos de teste com 5% da base de dados *Spiral*.

		k													
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Acc	**KNN**	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.96	0.93	0.92	0.92	0.89	0.86	0.85
	NK A	1.00	1.00	1.00	1.00	1.00	1.00	0.96	0.95	0.92	0.91	0.87	0.85	0.83	0.80
	NK B	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.98	0.98	0.98

Tabela 4. Resultados para conjuntos de teste com 10% da base de dados *Spiral*.

		k													
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Acc	**KNN**	1.00	1.00	1.00	1.00	0.97	0.95	0.93	0.90	0.87	0.85	0.84	0.78	0.75	0.74
	NK A	1.00	1.00	1.00	1.00	1.00	0.98	0.97	0.95	0.94	0.90	0.87	0.84	0.81	0.79
	NK B	1.00	1.00	1.00	1.00	1.00	0.98	1.00	0.98	0.98	0.97	0.97	0.96	0.95	0.94

3.3 Conjunto Aggregation

A Figura 4 apresenta a matriz de dispersão para o conjunto de dados *Aggregation*. Como pode ser observado nas tabelas 5 e 6, para esse conjunto de dados, os três métodos utilizados, tiveram uma acurácia de 100%, para o conjunto de teste com 5% dos dados. Para o conjunto de teste com 10% dos dados, apenas o método NK A apresentou acurácia abaixo de 100% (para *k* igual a 1 e 2). É interessante notar que nesse conjunto de dados, a matriz de dispersão indica 7

agrupamentos bem distintos, o que facilita a detecção por algoritmos de clustering que exploram a distância entre os objetos.

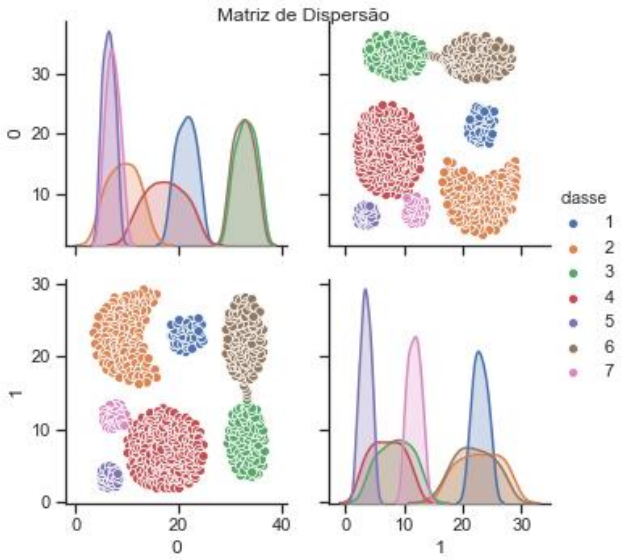

Figura 4: Matriz de dispersão do conjunto de dados *Aggregation.*

Tabela 5. Resultados para conjuntos de teste com 5% da base de dados *Aggregation*.

		k													
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Acc	**KNN**	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	NK A	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	NK B	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Tabela 6. Resultados para conjuntos de teste com 10% da base de dados *Aggregation*.

		k													
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Acc	**KNN**	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	NK A	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	NK B	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

4. Análise e Conclusões

Os resultados mostram que, ao utilizar a densidade, os métodos propostos permitiram retornar objetos dispostos nos clusters que não são necessariamente hiper-esféricos. Ao usar apenas a distância, o KNN adaptado não foi capaz de explorar a disposição de tais clusters. Os melhores resultados do KNN foram para o conjunto *Aggregation*, que possui agrupamentos compactos bem definidos. Além disso, a base de dados é maior que as outras, o que impacta a amostragem dos dados das bases de treinamento e teste.

Observa-se que, para as duas primeiras bases, conforme o número de objetos retornados (k) cresce, mais objetos de classes diferentes são retornados, ou seja, a acurácia piora. Entretanto, o impacto de k foi mais significativo para o KNN adaptado que nos métodos propostos. Em geral, diminuir o tamanho do conjunto de treinamento também implicou em diminuir a acurácia dos métodos. Como os mesmos valores de k foram testados, o número de erros aumentou para um conjunto menor de dados de treinamento. Os melhores resultados foram alcançados pelo método NK B, que percorre o grafo de interações sempre alcançando o vértice, com aresta incidente, relacionado ao objeto mais próximo do novo objeto a ser consultado. O método NK A não leva em consideração a distância para o novo objeto consultado depois que o vértice inicial é visitado. O uso da distância para o novo objeto mostrou-se útil na busca por similaridade.

Os métodos propostos se mostraram interessante para as bases de dados de formato arbitrário. Isso ocorre porque os métodos levam em consideração tanto a distância entre os objetos como também a densidade local deles. No futuro, testes com mais conjuntos de dados devem ser considerados. Além disso, testes com banco de imagens médicas devem ser realizados.

Referências

AHA, D. W.; KIBLER, D. & ALBERT, M.K. (1991). “Instance-based learning algorithms”, *Machine Learning*, 6(1): 37-66.

CARPINETO, C. & ROMANO, G. (2012). "A survey of automatic query expansion in information retrieval", *ACM Computing Surveys (CSUR)*, 44(1).

ESTER, M.; KRIEGEL, H.-P.; SANDER, J. & XU, X. (1996), “A density-based algorithm for discovering clusters in large spatial databases with noise”, *In the Proc. of the 2nd ACM Int. Conf. Knowl. Discovery Data Min. (KDD)*, 226–231.

FRÄNTI, P. & SIERANOJA, S. (2018). “K-means properties on six clustering benchmark datasets”, *Applied Intelligence*, 48 (12), 4743-4759.

HRUSCHKA, E. R.; CAMPELLO, R. J. G. B.; FREITAS, A. A. & CARVALHO, A. C. P. L. F. (2009). “A survey of evolutionary algorithms for clustering”, *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 39(2): 133-155.

HSINCHUN, C.; CHIANG, R. H. L. & STOREY, V. C. (2012). "Business intelligence and analytics: from big data to big impact", *MIS Quarterly,* 36(4): 1165-1188.

RODRIGUEZ, A. & LAIO, A. (2014). “Clustering by fast search and find of density peaks,” *Science*, 344(6191): 1492–1496.

TINÓS, R.; ZHAO, L.; CHICANO, F. & WHITLEY, D. (2018). “NK hybrid genetic algorithm for clustering”, *IEEE Transactions on Evolutionary Computation*, 22(5): 748-761.