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

Projeto de Pesquisa
PIBIC – CNPq

Área Prioritária do MCTIC: *Tecnologia Habilitadora I - Inteligência Artificial*

Classificação Baseada em K-Vizinhos Mais Próximos e no Grafo de Interações Nk

Aluno: Gustavo Fernandes Carneiro de Castro

Orientador: Renato Tinós

Resumo: K-Vizinhos mais Próximos (*K-Nearest Neighbors* – KNN) é um algoritmo de classificação não-paramétrica utilizado em problemas de diversas áreas. O KNN é bastante simples e intuitivo: um novo objeto é classificado com o rótulo da classe majoritária entre os *K* vizinhos mais próximos ao objeto. No KNN original, os *K* vizinhos mais próximos são determinados de acordo com a distância até o novo objeto. Geralmente a distância Euclidiana é empregada, o que facilita a formação de aglomerados hiper-elipsóides. Neste projeto, propõe-se o uso do grafo de interações Nk para retornar os *K* vizinhos mais próximos no algoritmo KNN. O grafo de interações Nk, originalmente empregado em clusterização, é construído com base na distância e densidade espacial de objetos em pequenos grupos formados por *k* objetos. Ao usar a distância aliada à densidade espacial, possibilita-se a formação de aglomerados com formatos arbitrários. Duas variações do método serão investigadas; os algoritmos diferem na forma em que os vértices associados aos *N* objetos da base de treinamento são visitados. Os *K*=*k* objetos relacionados aos vértices visitados são retornados como vizinhos mais próximos. As variações propostas serão comparadas entre si e com o KNN original em experimentos com bases de dados com propriedades diversas.

Ribeirão Preto, 17 de maio de 2021

1. Introdução e Justificativa

Classificação é um importante problema estatístico e de aprendizado de máquina [KOTSIANTIS *et al*, 2007]. Atualmente, classificação é requisito para diversas tecnologias aplicadas no nosso dia-a-dia, como reconhecimento de fala, identificação biométrica, visão computacional, sistemas de recomendação, entre outros.

O algoritmo dos *K-vizinhos mais próximos* (*K-nearest neighbors algorithm* - KNN) é um dos primeiros métodos eficientes de classificação não-paramétrica [FIX, 1985]. O KNN é bastante simples e intuitivo, servindo de base para o desenvolvimento de diversos outros algoritmos. Aqui, iremos estudá-lo no contexto de classificação, mas ele também pode ser utilizado em regressão e, com algumas modificações, para outros problemas de aprendizado de máquina.

Quando KNN é usado para classificação, um novo objeto é classificado com o rótulo da classe majoritária entre os *K* vizinhos mais próximos ao objeto [AHA *et al*., 1991]. No KNN original, os *K* vizinhos mais próximos são determinados de acordo com a distância até o novo objeto. Geralmente a distância Euclidiana é empregada, o que facilita a formação de aglomerados hiper-elipsóides. Tal característica é interessante para diversas bases de dados. Entretanto, em alguns problemas, os objetos de uma mesma classe estão dispostos em aglomerados arbitrários definidos tanto pela distância como pela densidade espacial [RODRIGUEZ & LAIO, 2014], [ESTER *et al*., 1996]. O KNN ignora a densidade espacial, levando em conta apenas a distância dos objetos para determinar os *K* vizinhos mais próximos ao novo objeto a ser classificado.

Em [TINÓS *et al*., 2018], o *NK hybrid genetic algorithm* (NKGA) foi proposto para *clustering*. O NKGA utiliza tanto a distância como a densidade espacial para o agrupamento de objetos. Para avaliar os particionamentos das bases 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 por *k* objetos. As informações sobre os grupos de objetos são capturadas no grafo de interações Nk[1]. 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.

Posteriormente em [MORAES & TINÓS, 2020], o grafo de interações Nk foi utilizado para o problema de busca por similaridade. Basicamente, o método proposto em [MORAES & TINÓS, 2020] retorna *K* objetos similares ao objeto consultado visitando *K* vértices do grafo de interações Nk que são vizinhos. O método proposto se mostrou interessante para a consulta de objetos em bases de dados com aglomerados com formato

[1] Originalmente, o grafo foi chamado de "*grafo de interações NK*". Aqui, de modo a não confundir um dos parâmetros do grafo com o parâmetro *K* do KNN, utilizaremos para o grafo a letra "*k"* e chamaremos o grafo de "*grafo de interações Nk*". O parâmetro *k* é definido arbitrariamente, mas aqui será utilizado *k*=*K* sendo que *K* neste caso é o parâmetro do KNN. Além disso, aqui o grafo é definido como o transposto do grafo original.

arbitrário. Isso ocorre porque o método leva em consideração tanto a distância entre os objetos como a densidade espacial nos *N* pequenos grupos.

2. Objetivos

O método proposto em [MORAES & TINÓS, 2020] para o problema de busca por similaridade baseado no grafo de interações Nk não é supervisionado e não foi proposto para o problema de classificação. Entretanto, ele pode ser adaptado para ser utilizado no algoritmo KNN em problemas de classificação.

A hipótese que será investigada neste trabalho é que: "*Utilizar o grafo de interações Nk para retornar os K objetos da base de treinamento que serão utilizados para a rotulação do novo objeto melhorará o desempenho do KNN em bases de dados que contenham aglomerados de formato determinado tanto por relações de distância como por relações de densidade espacial entre objetos*".

Portanto, o objetivo principal deste projeto é: "*Investigar estratégias eficientes para a utilização do grafo de interações Nk para retornar os K objetos da base de treinamento que serão utilizados para a rotulação do novo objeto em KNN*".

Aproveitando da simplicidade e praticidade do método de classificação KNN, além de sua sólida base conceitual, a proposta principal deste trabalho é a criação de um algoritmo de classificação baseado no KNN, aplicado ao grafo de interações Nk, visando reduzir as limitações do algoritmo original na aplicação em determinados problemas.

Entre os objetivos secundários deste projeto, mas não menos importantes, estão:

- Explorar as possibilidades oferecidas pela fusão de KNN e das propriedades do grafo de interações Nk;
- Modificar a maneira em que o grafo de interações Nk é construído. Atualmente, cada um dos *N* vértices é ligado a ele mesmo e a outros *k*-1 vértices. Destas *k*-1 conexões, uma aresta é definida por relação envolvendo a densidade dos objetos e as restantes por relação envolvendo a distância entre objetos. A flexibilização desta regra será explorada neste projeto, permitindo alterar a razão entre arestas definidas de acordo com densidade espacial e distância;
- Comparar variantes do KNN modificado com o KNN original em bases de dados, com propriedades distintas, de repositórios públicos para diferentes valores de *K*;
- Comparar variantes do KNN modificado com o KNN original em um problema prático da área de Medicina para diferentes valores de *K*;
- A capacitação do aluno nas áreas de:
    - Inteligência Artificial;
    - Aprendizado de Máquina;
    - Teoria dos Grafos;
    - Classificação.

3. Metodologia

Este é um trabalho derivado da pesquisa feita durante o mestrado de José Carlos Bueno de Moraes no Programa de Computação Aplicada [MORAES, 2020]. Na pesquisa realizada no mestrado, o grafo de interações Nk [TINÓS *et al.*, 2018], desenvolvido inicialmente para o problema de *clustering*, foi utilizado para o problema de busca por similaridade. Aqui, o grafo de interações Nk será utilizado no KNN. O KNN e o grafo de interações Nk serão discutidos a seguir.

3.1. K-Vizinhos Mais Próximos (KNN)

O KNN é um algoritmo de classificação no qual um novo objeto **x** é rotulado examinando-se a classe majoritária entre os *K* objetos da base de treinamento mais próximos a **x** [AHA et al., 1991]. O KNN pode ser modificado para ser utilizado em problemas de regressão, analisando-se os valores dos *K* objetos mais próximos a **x** e, por exemplo, calculando-se a média dos valores associados a este objeto. O KNN pode utilizar pesos baseados na distância [ALTMAN, 1992] ou a área para definir os *K* vizinhos mais próximos. Na abordagem baseada em área, o parâmetro *K* define o raio da hiper-esfera em torno do objeto **x**. Os objetos da hiper-esfera são utilizados então para rotular **x**.

Geralmente a distância Euclidiana é empregada para a determinação dos vizinhos mais próximos. De qualquer modo, sempre se utiliza a distância para determinar os *K* objetos da base de treinamento mais próximos a **x**. O parâmetro *K* definido pelo usuário tem impacto significativo na definição das regiões de decisão e, em consequência, no desempenho do classificador. A Figura 1 mostra um exemplo. Observa-se a mudança provocada pelo aumento de *K* na definição das regiões de decisão. Nesta pesquisa, os algoritmos propostos serão comparados ao KNN original para diferentes valores de *K*.

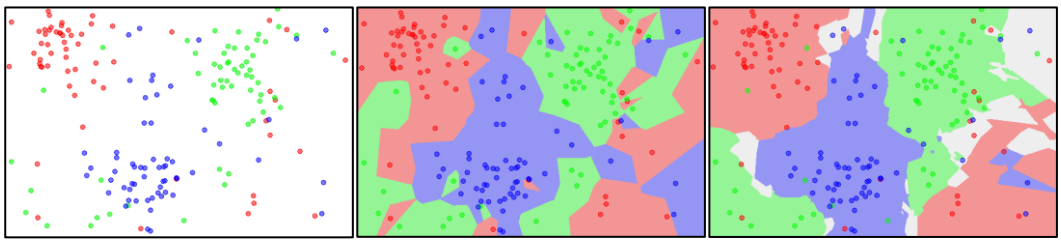

Figura 1. Regiões de decisão geradas por KNN para a base de dados bidimensional mostrada à esquerda para *K*=1 (centro) e *K*=5 (direita). Os valores da coordenadas não são apresentados por simplicidade.
Fonte: https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

3.2. Grafo de Interações Nk

Diversos algoritmos de *clustering*, como *k-means* [MACQUEEN, 1967], utilizam apenas a distância entre objetos para a definição dos agrupamentos. Outros, como *density-based spatial clustering of applications with noise* (DBSCAN) [ESTER *et al*., 1996], utilizam a densidade espacial dos objetos para a definição dos agrupamentos. O NKGA **[**TINÓS *et al*., 2018**]** utiliza tanto a distância entre objetos como a densidade espacial para criar *N* pequenos grupos, cada um com *k* objetos, com o objetivo de agrupar exemplos de uma base de dados de tamanho *N*. As informações sobre os grupos de objetos são capturadas no grafo de interações Nk.

O grafo de interações Nk é um grafo direcionado com *N* vértices, cada um com grau de saída *k*. Cada aresta (v_j , v_i) do grafo indica que o *j*-ésimo objeto é relacionado com o *i*-ésimo objeto. Dada uma base de dados (treinamento) com *N* objetos *n*-dimensionais, o primeiro passo para a construção do grafo é adicionar vértices v_i , $i = 1,...,N$, com auto-loop para cada objeto $\mathbf{y}_i$ da base de dados.

A construção das *k*-1 arestas de saída restantes 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:

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

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

$$K(\mathbf{x}) = e^{\frac{-\|\mathbf{x}\|^2}{2\epsilon^2}} \qquad (2)$$

e ϵ o parâmetro que define a distância de corte. No grafo de interações Nk, 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_j representando o objeto mais próximo a $\mathbf{y}_i$ e que possui densidade maior que $\mathbf{y}_i$ é identificado. Então, uma aresta (v_i, v_j) é criada. O último passo é adicionar arestas de v_i até os vértices representando os objetos mais próximos à $\mathbf{y}_i$. Isso é feito até que o grau de saída de cada vértice seja igual a *k*. A Figura 2 apresenta um exemplo do processo de criação do grafo de interações Nk.

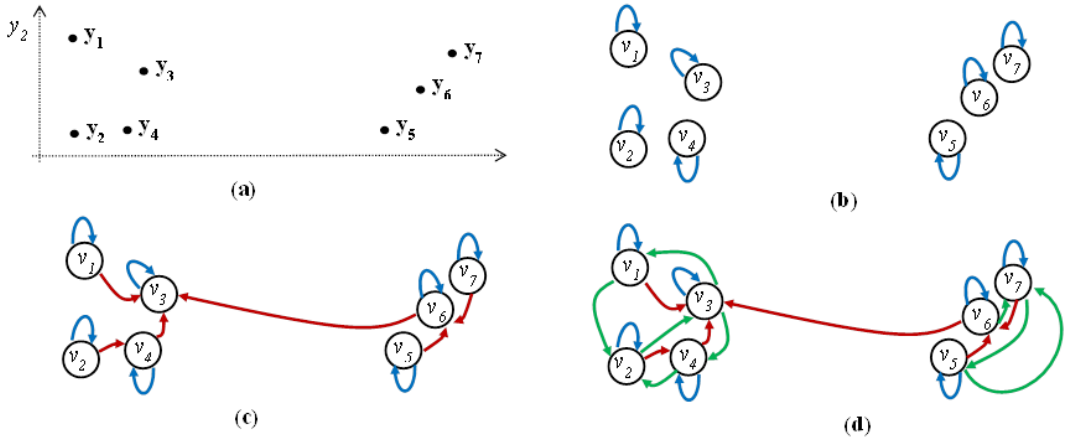

Figura 2: Exemplo de construção do grafo de interações Nk com *k* = 3 para um conjunto com 7 objetos bidimensionais (*N* = 7, *n* = 2). Cada objeto da base de dados (a) é associado com um vértice com auto-loop (b). A densidade dos objetos é calculada e cada vértice é ligado ao vértice associado com o 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 saída de cada vértice seja igual a *k*. O gráfico de interações (d) tem *N* = 7 vértices e *Nk* arestas.

3.3. KNN baseado no Grafo de Interações Nk

No KNN proposto, o grafo de interações Nk será utilizado para retornar os *K* vizinhos mais próximos a **x** no algoritmo KNN. Ao usar a distância aliada à densidade espacial, possibilita-se a formação de aglomerados com formatos arbitrários. A Figura 3 mostra um exemplo em que as distribuições dos objetos pertencentes a cada classe não são hiper-esféricas. Os agrupamentos nesta base de dados são mais bem descritos utilizando-se densidade espacial e relações de distância entre objetos vizinhos. Se KNN for aplicado neste problema com *K* grande, objetos de classes distantes devem ser escolhidos para a definição de rótulo, já que apenas a distância Euclidiana será levada em conta na definição dos *K* vizinhos mais próximos. Neste problema, a hiper-esfera não define os melhores agrupamentos [TINÓS *et al*., 2018]. O grafo de interações Nk (Figura 3) possibilita neste caso capturar informações para a definição de agrupamentos mais interessantes para este problema.

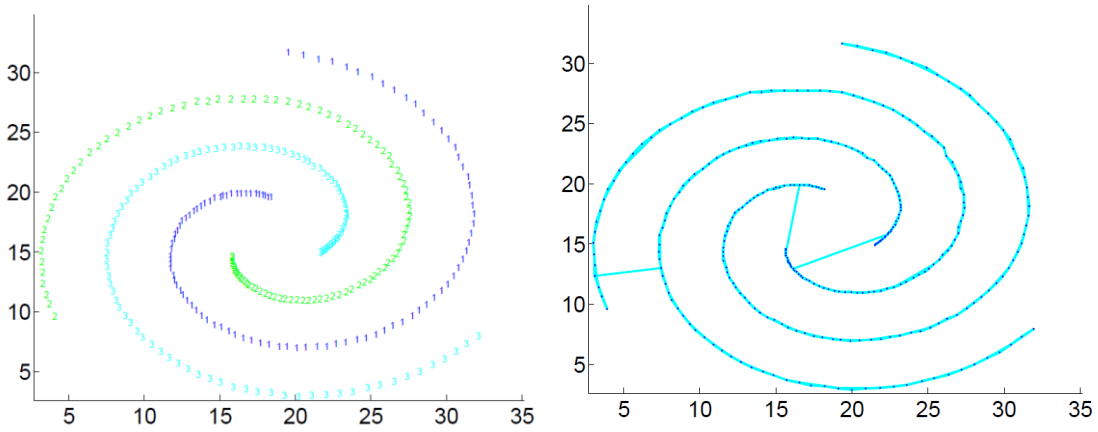

Figura 3. Distribuição dos objetos da instância *path-based 2* (*spiral*) do repositório *Shape Dataset* [FRÄNTI & SIERANOJA, 2018] e grafo de interações Nk (direita) para *k*=3 .
Fonte: Material Suplementar de [TINÓS *et al*., 2018].

Duas variações do método serão investigadas; os algoritmos diferem na forma em que os vértices associados aos N objetos da base de treinamento são visitados. Os $K=k$ objetos relacionados aos vértices visitados são retornados como vizinhos mais próximos. Os K objetos retornados são então utilizados para rotular o novo objeto **x**. 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. A definição do primeiro vértice visitado v_x também é igual nos dois métodos. Após a criação do grafo de interações Nk, calcula-se a distância Euclidiana do novo objeto **x** para cada um dos objetos do conjunto de treinamento, similarmente ao que é feito no KNN original. O vértice relacionado ao objeto mais próximo a **x** é definido como v_x .

No primeiro método, após a identificação do vértice inicial v_x, os k nós com arestas saindo de v_x são identificados, i.e., retorna-se a lista de adjacências para os nós saindo de v_x. Os objetos associados aos k vértices com arestas saindo de v_x (incluindo v_x) são então retornados pelo método como os mais similares ao objeto **x**. Lembrando que, como cada vértice possui auto-loop, o objeto associado a v_x estará entre os $k=K$ objetos retornados.

No segundo método, após a identificação de v_x, encontra-se o vértice não visitado v_j com arestas partindo de v_x cujo objeto $\mathbf{y}_j$ é mais próximo ao objeto **x** (de acordo com a distância Euclidiana). Então, este novo vértice v_j é visitado e repete-se o processo até que k vértices sejam visitados. Ou seja, ao visitar v_j, faz-se $v_x = v_j$ e encontra-se o próximo vértice não visitado mais próximo a **x** e assim por diante Os objetos associados ao k vértices visitados são então retornados pelo método como os K mais similares ao objeto **x**.

Ainda, variações nas quais a maneira em que o grafo de interações Nk é construído serão investigadas. Atualmente, cada um dos N vértices tem arestas para k-1 outros vértices, não levando em consideração o auto-loop. Como apresentado anteriormente, destas conexões uma aresta é definida por relação envolvendo a densidade espacial e as restantes por relação envolvendo a distância entre objetos. Uma variação em que $|\alpha\,(k\text{-}1)|$ arestas, sendo $0 \leq \alpha \leq 1$ um parâmetro real, são definidas por densidade e as restantes

((*k-1*) - |α (*k-1*) | arestas) são definidas por distância será investigada. Quando |α (*k-1*) |=1, o grafo será igual ao grafo de interações Nk original, ao passo que se |α (*k-1*) | = 0, apenas a distância será utilizada, como é atualmente feito no KNN.

As variações propostas serão comparadas entre si e com o KNN original em experimentos com bases de dados com propriedades diversas. Diferentes valores de *K* serão considerados. Em um primeiro momento, serão utilizadas instâncias das bases de repositórios públicos UCI [DUA & GRAFF, 2019] e *Shape* [FRÄNTI & SIERANOJA, 2018]. Posteriormente, o KNN modificado que apresentar o melhor desempenho para as bases públicas será comparado com o KNN original em um problema prático da área de Medicina.

4. Plano de Trabalho e Cronograma

A seguir, o cronograma para a pesquisa é apresentado.

	1º Bimestre		2º Bimestre		3º Bimestre		4º Bimestre		5º Bimestre		6º Bimestre	
Fase 1												
Fase 2												
Fase 3												
Fase 4												
Fase 5												
Fase 6												
Fase 7												
Fase 8												
Fase 9												

Fase 1) Levantamento Bibliográfico: estudo das referências relacionadas aos temas da pesquisa.
Fase 2) Implementação do grafo de interações Nk e do KNN original na linguagem Python. Esta será a linguagem utilizada no desenvolvimento do projeto.
Fase 3) Desenvolvimento da primeira variação do KNN baseado no grafo de interações Nk (ver Seção 3.3) .
Fase 4) Desenvolvimento da segunda variação do KNN baseado no grafo de interações Nk (ver Seção 3.3)
Fase 5) Testes com os métodos desenvolvidos em instâncias das bases *UCI* e *Shape*.
Fase 6) Desenvolvimento e testes da variação do KNN baseado no grafo de interações Nk modificado (ver Seção 3.3)
Fase 7) Testes com os métodos desenvolvidos em um problema de classificação da área de Medicina. O problema ainda será definido
Fase 8) Análise dos resultados obtidos.
Fase 9) Confecção de Relatórios e artigos científicos.

5. Considerações finais

O presente projeto será desenvolvido no Departamento de Computação e Matemática (DCM) da Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto (FFCLRP) da Universidade de São Paulo (USP). O DCM é responsável por dois cursos de graduação: Bacharelado em Ciência da Computação e Matemática Aplicada a Negócios. O DCM é responsável pelo programa de mestrado em Computação Aplicada (criado em Maio de 2015) e programa de mestrado em Matemática (criado em Setembro de 2020).

O aluno que está pleiteando a bolsa cursa atualmente o quinto semestre do curso de Bacharelado em Ciência da Computação. De acordo com informações obtidas no Sistema de Graduação JupiterWEB da USP em 07/05/2021, o aluno tem média ponderada 8,2, ao passo que a média ponderada dos alunos do curso é 7,6. O aluno não teve reprovações em disciplinas. A classificação do aluno é 5º entre os 20 alunos ingressantes da turma.

O presente trabalho será inicialmente executado remotamente devido às restrições impostas pela Pandemia de COVID19. Quando as aulas presenciais retornarem, o aluno utilizará o Laboratório de Sistemas Computacionais Complexos do DCM. Este laboratório, coordenado por 3 docentes e conta com, entre outros, 8 computadores pessoais e 2 estações de trabalho.

Referências Bibliográficas

ALTMAN, N. S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression", *The American Statistician*, 46(3): 175-185.

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

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.

FIX, E. (1985). "*Discriminatory analysis: nonparametric discrimination, consistency properties*", Technical Report, USAF School of Aviation Medicine.

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

KOTSIANTIS, S. B.; ZAHARAKIS, I. & PINTELAS, P. (2007). "Supervised machine learning: A review of classification techniques", *Emerging artificial intelligence applications in computer engineering*, 160(1): 3-24.

MACQUEEN, J. (1967). “Some methods for classification and analysis of multivariate observations”, *In Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1(14): 281-297.

MORAES, J. C. B. (2020). “*Busca por similaridade utilizando grafo de interações NK*”, Dissertação de Mestrado em Computação Aplicada, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo.

MORAES, J. C. B., & TINÓS, R. (2020). “Busca por Similaridade usando o Gráfico de Interação NK”, *Nos Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*, 222-233.

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.