UNIVERSIDADE ESTADUAL DE CAMPINAS

Instituto de Matemática, Estatística e Computação Científica

WANDERSON LUIZ DA SILVA

MÉTODOS DE AGRUPAMENTOS COM RESTRIÇÕES E COM BUSCA EM VIZINHANÇA VARIÁVEL COM APLICAÇÕES EM SÉRIES TEMPORAIS DE IMAGENS NDVI

CAMPINAS
2017

Wanderson Luiz da Silva

MÉTODOS DE AGRUPAMENTOS COM RESTRIÇÕES E COM BUSCA EM VIZINHANÇA VARIÁVEL COM APLICAÇÕES EM SÉRIES TEMPORAIS DE IMAGENS NDVI

Tese apresentada ao Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Doutor em Matemática Aplicada.

Orientador: Francisco de Assis Magalhães Gomes Neto

ESTE EXEMPLAR CORRESPONDE À VERSÃO FINAL DA TESE DEFENDIDA PELO ALUNO WANDERSON LUIZ DA SILVA E ORIENTADA PELO PROF. DR. FRANCISCO DE ASSIS MAGALHÃES GOMES NETO.

Campinas, SP
2017

Agência(s) de fomento e nº(s) de processo(s): CAPES

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

Si38m

Silva, Wanderson Luiz da, 1978-
Métodos de agrupamentos com restrições e com busca em vizinhança variável com aplicações em séries temporais de imagens NDVI / Wanderson Luiz da Silva. – Campinas, SP : [s.n.], 2017.

Orientador: Francisco de Assis Magalhães Gomes Neto.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

1. Análise por agrupamento - Processamento de dados. 2. Meta-heurística. 3. Algoritmo k-means. 4. Imagens de sensoriamento remoto. I. Gomes Neto, Francisco de Assis Magalhães, 1964-. II. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Restricted clustering and variable neighborhood search with time-series applications of NDVI images
Palavras-chave em inglês:
Cluster analysis - Data processing
Metaheuristic
k-means algorithm
Remote-sensing images
Área de concentração: Matemática Aplicada
Titulação: Doutor em Matemática Aplicada
Banca examinadora:
Francisco de Assis Magalhães Gomes Neto [Orientador]
João Francisco Gonçalves Antunes
Aurélio Ribeiro Leite de Oliveira
Jurandir Zullo Júnior
Helenice de Oliveira Florentino Silva
Data de defesa: 11-12-2017
Programa de Pós-Graduação: Matemática Aplicada

Tese de Doutorado defendida em 11 de dezembro de 2017 e aprovada pela banca examinadora composta pelos Profs. Drs.

Prof(a). Dr(a). FRANCISCO DE ASSIS MAGALHÃES GOMES NETO

Prof(a). Dr(a). AURELIO RIBEIRO LEITE DE OLIVEIRA

Prof(a). Dr(a). HELENICE DE OLIVEIRA FLORENTINO SILVA

Prof(a). Dr(a). JOÃO FRANCISCO GONÇALVES ANTUNES

Prof(a). Dr(a). JURANDIR ZULLO JUNIOR

As respectivas assinaturas dos membros encontram-se na Ata de defesa

Aos meus pais José Luiz, Vanda Baldez e Maria Reine, à minha esposa Sheila e aos meus filhos Lucca, Alice e Rafael.

Agradecimentos

A linguagem escrita aproxima pobremente o sentimento daqueles que escrevem, como um espantalho que tenta imitar o homem. Posso dizer o nome daqueles que foram importantes para este trabalho, mas sem a pretenção de que o que vou dizer seja capaz de descrever a importância que tiveram. Agradeço ao meu orientador Francisco, que com sua paciência e dedicação me inspirou a concluir esta tese e a fazê-la da melhor maneira possível. Agradeço ao professor Stanley, que além de me ajudar a trabalhar os aspectos técnicos de *data mining*, auxiliou-me em tantos sentidos diferentes, que é importante que se diga que, sem ele, este trabalho simplesmente não seria possível. Agradeço também ao professor Jurandir, que com sua gentileza e competência me mostrou o tipo de pesquisador e professor que eu quero ser.

Eu sou feito do que recebo dos outros e recebi dos meus pais, José Luiz e Vanda Baldez, a vida e os valores que me definem. Meus pais se tornaram meus melhores amigos e a simples lembrança de um sorriso deles me recobra o ânimo mesmo em dias nebulosos. Recebi de meus irmãos Wagner, Washington e Júnior o companheirismo típico dos irmãos que se amam e por meus sobrinhos Diogo e Leonardo, eu recebi a vontade de ser um bom exemplo. Da minha esposa e alma gêmea, eu recebi o apoio sem o qual nada do que fiz seria possível. E por fim, de meus três filhos, Lucca, Alice e Rafael (todos nascidos durante o doutorado), eu recebi um motivo para viver e para me tornar a cada dia alguém melhor. Quero ser um pai que eles tenham orgulho.

Acredito que meus amigos desconhecem a importância que eles têm em minha vida, são irmãos e irmãs de espírito que sempre me apoiaram quando as coisas não estavam bem e comemoraram comigo quando as vitórias vinham. 19, Abel, Charles, Daniel e Raniere me acompanharam em grande parte deste trabalho e sabem o que ele significa para mim. Leandro e Ana me deram dois dos meus maiores tesouros e eu aprendi a ser pai, antes de ser pai, pelo amor que dediquei aos meus afilhados Iori e Larissa. Hidelbrando é simplesmente uma das melhores pessoas que conheci em minha vida. Flávio e Ana são pessoas que, em um primeiro momento, olhos menos atentos podem deixar passar o quão excepcional este casal é, tê-los como amigos é uma honra. Glauco e Aline são parceiros que fazem a vida ser melhor.

Parece incomum que eu lembre de professores de meu ensino médio, mas foram eles os meus primeiros parceiros dos sonhos. Foram eles que permitiram que um *office boy* de BH, com deficiência em aspectos elementares de formação, se tornasse doutor pela melhor universidade da América Latina. Rodrigo, Rommel, Paulo, Bernadelli e Marcelo, cada conquista que eu alcançar tem o dedo de vocês.

Agradeço também à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela bolsa de estudos de Doutorado. E por fim, em um nível emocional, muitas pessoas resistem à ideia de que as influências aleatórias são importantes. Apesar de compreendermos isso intelectualmente, subestimamos o papel do acaso na vida de todos nós. Pensando nisso achei adequado agradecer a Deus pelas pessoas e oportunidades que sempre recebi.

Resumo

O problema tratado foi o de formular k grupos a partir de um subconjunto discreto e finito em um espaço d-dimensional. Os grupos constituídos devem obedecer a uma divisão que privilegie a alta similaridade entre elementos do mesmo grupo e a alta dissimilaridade entre elementos de grupos distintos. Trata-se de um problema de clusterização partitiva, que possui aplicações em um grande número de áreas, desde a segmentação de mercado até a análise de sequências macromoleculares.

Nesta tese, propomos dois métodos para realizar este agrupamento: um algoritmo semissupervisionado, baseado em uma variante do k-médias (pk-means), que usa restrições espaciais para os centroides. Ele permitiu realizar agrupamentos com o uso de amostras classificadas que seriam descartadas por uma abordagem não supervisionada e que, por outro lado, não seriam suficientes para induzir um classificador supervisionado. A segunda proposta é um esquema de vizinhança para uso da VNS (meta-heurística de busca em vizinhança variável) como apoio a algoritmos de clusterização que são incorporados como métodos de busca local (variantes VNS).

Fizemos uma grande variedade de experimentos computacionais que confirmaram que tanto o pk-means quanto as variantes VNS geram grupos mais homogêneos e coesos. Acrescido a isso, temos que os métodos propostos podem ser aplicados inclusive a dados dinâmicos, como séries temporais de imagens de satélite, pois apresentamos uma transformação que troca as séries temporais por parte dos seus harmônicos. Isso permitiu que realizássemos um estudo de caso, usando os métodos propostos para classificar imagens NDVI extraídas do sensor AVHRR/NOAA 17 e TERRA/MODIS.

Palavras-chave: análise de clusters-análise de dados; meta-heurística; algoritmo k-means; imagens de sensoriamento remoto

Abstract

The problem treated is one of formulating k groups from a discrete and finite subset in a d-dimensional space. The groups must have high similarity between elements of the same group and high dissimilarity between elements of different groups. It is a partitive clustering problem that has applications in multiple areas like market segmentation or macromolecular sequence analysis.

In our work, we propose two methods to perform this grouping: a semi-supervised algorithm, based on a k-means variant (pk-means), which uses spatial constraints for centroids. It allowed grouping by the use of classified sample that would be discarded by an unsupervised approach and, on the other hand, would not be sufficient to induce a supervised classifier. The second method is based on a neighborhood for VNS metaheuristic (variable neighborhood search) as support for clustering algorithms, that is used like local search (VNS variants).

We confirmed by a great variety of computational experiments that both, pk-means and VNS variants, generate more homogeneous and cohesive groups. In addition to this, we applied the proposed methods to dynamic data, such time series of satellite images, because we also present a transformation that exchange the time series by their harmonics. This allowed us to classify NDVI images extracted from the AVHRR / NOAA 17 and TERRA / MODIS sensors.

Keywords: cluster analysis - data processing; meta-heuristic; k-means algorithm; remote sensing images

Lista de Figuras

Lista de Tabelas

Notação

$\mathscr{C}$: Agrupamento definido por uma partição. *Exemplo*:

$$\left\{\{\mathbf{x}_1,\cdots,\mathbf{x}_{n_1}\},\{\mathbf{x}_{n_1+1},\cdots,\mathbf{x}_{n_2}\},\cdots,\{\mathbf{x}_{n_k},\cdots,\mathbf{x}_n\}\right\}$$

C_i : Conjunto de centroides, i.e. $C_i = \{\mathbf{c}_1,\mathbf{c}_2,\cdots,\mathbf{c}_k\}$. Também pode ser entendido com uma matriz $k \times d$ onde cada uma das linhas é um centroide.

$\mathbf{c}_i$: Centroide do grupo i.

$\mathbf{e}_j$: Vetor canônico definido por $(e_j)_i = 0 se i \neq j$ e $(e_j)_j = 1$.

k : Número de grupos.

k' : Número de restrições.

k_d : Quando um algoritmo de busca converge para $k - k_d < k$ grupos, o valor $k_d > 0$ é o número de degeneração.

$\mathbf{X}$: Conjunto de elementos a serem agrupados, i.e. $\mathbf{X} = \{\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n\}$.

$\mathbf{x}$: Elemento a ser agrupado.

$\mathbf{f}(\mathbf{a})\cdot\mathbf{e}_j$: Produto escalar entre o campo vetorial $\mathbf{f}$ e $\mathbf{e}_j$, que equivale a j-ésima coordenada da imagem $\mathbf{f}(\mathbf{a})$.

x_j : j-ésima coordenada de $\mathbf{x}$.

Siglas e termos

AST	:	Agrupamento de séries temporais.
AVHRR	:	*Advanced Very High Resolution Radiometer*
Cluster	:	Anglicismo da palavra grupo, coleção.
Clusterização	:	Aportuguesamento do anglicismo *clustering*, significando agrupamento.
Clustering	:	Anglicismo de agrupamento, equivalente a clusterização.
DM	:	Sigla referente a mineração de dados. Suas iniciais decorrem do termo inglês: *Data Mining*.
KDD	:	Sigla referente a descoberta de conhecimento de base de dados, do inglês *knowledge-discovery in databases*.
NDVI	:	*Normalized Difference Vegetation Index*
VNS	:	*Variable Neighborhood Search*
HANTS	:	*Harmonic ANalysis of Time Series*
MODIS	:	*Moderate Resolution Imaging Spectroradiometer*
NOAA	:	*National Oceanic and Atmospheric Administration*
mvc	:	*Maximum Value Composites*
REM	:	Radiação eletromagnética
Padrão ouro	-	(ou verdade terrestre) é a referência superivisionada de classificação de um grupo.
SR	-	sensoriamento remoto

Sumário

Capítulo 1

Introdução

"The task is not to see what has never been seen before, but to think what has never been thought before about what you see everyday."
- Erwin Schrodinger

Nesse trabalho investigamos técnicas de resolução do problema de agrupamento k através de uma abordagem partitiva, supondo conhecido *a priori* um pequeno conjunto de informações sobre a classe de alguns dos elementos. O problema de agrupamento k é formalmente apresentado na Seção 3.1, mas pode ser entendido como o problema de gerar k grupos (sendo k conhecido) a partir de uma coleção discreta $\mathbf{X} \subset \mathbb{R}^d$ de objetos, onde d dimensiona as características do objeto. Essa divisão deve privilegiar a alta similaridade entre elementos associados ao mesmo grupo e a alta dissimilaridade entre elementos postos em grupos distintos. Nesse formato, esse é um problema de clusterização semissupervionada e possui aplicações em uma infinidade de áreas, desde a segmentação de mercado até a análise de sequências macromoleculares.

Para os fins aqui propostos, os objetos a serem agrupados são séries temporais de imagens de satélite. Uma série temporal de imagens de satélite pode ser entendida, de forma simplificada, como uma coleção de pixels em que a cada pixel está associada a uma série temporal das refletâncias de uma região de interesse. Essa abordagem possui o inconveniente da perda do contexto espacial, mas é uma forma simples de interpretar os dados e tem alcançado bons resultados para o propósito de classificação.

O problema em domínio é complexo demais para ser resolvido por uma disciplina específica, requerendo uma resolução interdisciplinar que leve em conta o uso de grandes acervos de informações relacionadas ao problema e formas eficientes de gerar um particionamento. Daí, naturalmente, emerge a proposta baseada no uso conciliado de técnicas de mineração de dados (DM, do inglês *Data Mining*) e técnicas de otimização.

1.1 Interface entre DM e otimização

"In God we trust, all others bring data."
- W. Edwards Deming

Em 2012, foram criados cerca de 2,5 quintilhões de bytes por dia. Isso preencheria dez trilhões de livros com mil páginas cada, o que corresponde a um acervo aproximadamente 66 mil vezes

maior que o da biblioteca do congresso nos EUA, considerada a maior biblioteca do mundo. Estima-se que 90% de todos os dados da atualidade foram produzidos nos últimos dois anos[1]. Estes dados são gerados pelas mais variadas fontes: redes sociais, sites, vídeos, sensores orbitais e vários outros. E nesse arcabouço de dados existem padrões ocultos, improváveis ou impossíveis de serem descobertos mesmo por especialistas.

Analisar grandes volumes de dados se tornou uma tarefa essencial para as atividades sociais. Suas aplicações se estendem a todos os ramos do conhecimento, desde a identificação de anomalias e tendências de epidemias à melhoria em inteligência de negócios. Sendo assim, existe uma clara demanda por técnicas que possam transformar esse amontoado de informação em conhecimento estratégico. Este é o foco de uma área de pesquisa chamada mineração de dados, que se formou em meados da década de 90, a partir da contribuição de diversas áreas como estatística, aprendizado de máquina e banco de dados.

Pela importância dessa área, existe uma tendência mundial a desenvolver múltiplas abordagens por meio de trabalhos combinados de pesquisadores de áreas diferentes e, em especial, da área de matemática. Como exemplo, cabe citar que um dos melhores classificadores da atualidade, a máquina de vetores de suporte (*SVM*, do inglês *support vector machine*), incorpora princípios de otimização para gerar hiperplanos separadores em um espaço de Hilbert. Trabalhos na área de agrupamento (*clustering*) estão sendo feitos pelos proponentes da meta-heurística de busca em vizinhança variável (*VNS*, do inglês *variable neighborhood search*) e existem trabalhos de fuzzificação de algoritmos *crisp* consagrados na literatura, como o *c-means*. Entretanto, no Brasil, mineração de dados é quase que exclusivamente tratada por pesquisadores da área de ciência da computação.

Os matemáticos brasileiros, mesmo os que pesquisam na área de aplicações, têm demonstrado pouco interesse nesta área. Entretanto, essa ausência de interesse pode ser encarada como uma oportunidade para trabalhar em um segmento teórico pouco explorado e de grande utilidade. Uma ponte entre métodos de pesquisa operacional, técnicas de otimização, análise funcional e equações diferenciais com técnicas de DM se justifica tanto quanto outras relações já consagradas como biomatemática, física matemática e geofísica matemática.

Em uma época marcada pela pletora de dados, os conceitos de DM podem ser flexibilizados para diversas áreas, tais como a epidemologia, com previsões eficientes de atividades de gripe pela frequência de pesquisa de certos termos (*google flu trends*), ou a educação, com um crescente conjuntos de técnicas denominadas mineração de dados educacionais (*EDM*, do inglês *educational data mining*).

Estabelecer um diálogo entre grupos de pesquisa na área de matemática aplicada e mineração de dados é estender o atual domínio de contribuição da matemática a uma área cujo o escopo de utilidades é uma espécie de "***novo petróleo***".

[1] http://www-01.ibm.com/software/data/bigdata em 10/06/2013.

1.2 Aplicação na agricultura

"(...) we paid no attention to disciplinary boundaries; we blithely followed problems wherever they led. For better or for worse, I've never been able to shake this approach."
- Alan Dowty

O Brasil é um país de vocação agrícola. Portanto, pensar sobre aplicações na área agrícola é quase mandatório. Acrescido a isso, os dados de sensoriamento remoto têm aumentado continuamente em volume e velocidade[2]. Estes dados podem ser usados para auxiliar o monitoramento e a previsão de safra do Brasil, que na atualidade são feitos de forma censitária ou por levantamento de amostras no campo [61]. Em especial, eles podem gerar estimativas de produção para culturas como a cana-de-açúcar, que é usada para a produção de biocombustível e tem importância estratégica para a economia do país[3]. Entretanto, a análise humana dos dados, em tempo hábil, se tornou impossível. Daí a necessidade de se usar técnicas que sejam capazes de extrair, de forma automática, padrões, tendências e relações das séries temporais de imagens de satélite. As técnicas de mineração de dados (DM) parecem ser adequadas para transformar esse labirinto de imagens em ferramentas para auxiliar na tomada de decisão em demandas do agronegócio [75].

A classificação de imagens de satélite auxilia na estimativa de áreas cultivadas, que, por sua vez, favorece um reescalonamento do plantio das culturas, para que não haja a superprodução de um único produto e a escassez de outros. Também contribui para o abastecimento dos mercados interno e externo; para o fomento de alguns produtos tidos como essenciais para a economia nacional; para a estimativa dos prejuízos decorrentes de pragas, doenças e de fenômenos da natureza como seca, inundação, entre outros, que são comuns em países tropicais como o Brasil. A classificação de dados de sensoriamento remoto pode auxiliar no agronegócio, tendo um importante papel em estimativas de área cultivada, previsão de safra, monitoramento ambiental, identificação de uso e cobertura de terra e outros [1].

A classificação de imagens multitemporais de satélite tem a função de fornecer uma caracterização dinâmica de uma cultura e se justifica pela sua importância estratégica. Seja para estimativa de área ou auxílio da previsão de safra, é importante que se tenha, com alguma antecedência e de forma rápida e precisa, informações sobre a cultura da região de interesse. Atualmente, usa-se levantamentos de campo, de caráter censitário ou amostral, para suprir essas demandas [61]. Estimativas de áreas cultivadas com cana-de-açúcar, por exemplo, têm implicações na indústria, no sentido de ampliar seus lucros, e em políticas públicas, como as questões que envolvem o impacto do uso de biocombustível no clima e no valor dos alimentos.

Soluções para o problema de classificação de imagens multitemporais de satélite emergem naturalmente do uso de técnicas de classificação e clusterização. O desenvolvimento de algoritmos de classificação não supervisionada é provavelmente o problema mais estudado em mineração de dados [2] e pode permitir a obtenção automática ou semi-automática de informações a partir de extensos bancos de imagens e dados de sensoriamento remoto.

[2] O CEPAGRI/Unicamp (Centro de Pesquisas Meteorológicas e Climáticas Aplicadas à Agricultura) possui um acervo de imagens e dados relacionados a meteorologia e climatologia desde 1995 que já ultrapassou os 6 TB.

[3] O Brasil ocupa o papel de maior exportador de açúcar do mundo, gerando mais de dois bilhões de dólares por ano na balança comercial e onde o estado de São Paulo, que é o maior produtor nacional, tem aproximadamente 20% de sua área coberta por cana-de-açúcar

1.3 Proposta e objetivos

"I'm aiming for something higher"
- Richard Stallman

Uma abordagem possível para gerar agrupamentos é formular o problema de clusterização como um problema de otimização, em que a partição mais adequada maximiza ou minimiza uma ou múltiplas funções objetivos [35, 51, 54, 91]. A grande parte dos algoritmos baseados neste paradigma se apoia em métodos de busca local, tendo como desvantagem o aprisionamento em mínimos locais, além do inconveniente de serem sensíveis aos pontos iniciais [31]. O crescimento de investidas para sanar estes inconvenientes é facilmente atestada pelo aumento no número de trabalhos que usam meta-heurísticas para geração de agrupamentos.

Neste contexto propõe-se o desenvolvimento de um método eficiente para resolver o problema combinatorial de agrupamento k (com restrições), com posterior aplicação deste no domínio específico de agrupamentos de séries temporais de imagens de satélite. De forma mais específica, tem-se a seguinte hipótese:

> **Hipótese 1:** O uso de uma variante da heurística k-médias (*k-means*) que suporte restrições, quando agregada à meta-heurística de busca em vizinhança variável (VNS), permite a fuga de mínimos locais, redução da sensibilidade aos pontos iniciais e pode ser um método competitivo comparado às técnicas sugeridas na literatura.

Os objetivos por trás dessa hipótese são dois:

i) verificar o ganho que se estabelece em um agrupamento obtido pelo algoritmo k-médias ao se incorporar algumas restrições sobre os grupos, ou seja, ao se usar o paradigma semissupervisionado de clusterização;

ii) verificar a qualidade dos agrupamentos obtidos pela busca local simples comparada à qualidade dos agrupamentos obtidos com a incorporação da VNS.

A escolha do algoritmo k-médias se justifica pela facilidade de implementação, além do baixo custo computacional, enquanto a escolha do esquema básico de VNS se dá em parte pela forma simples como o método explora conceitos de vizinhanças para gerar variações na solução.

Neste ponto, os dados são séries temporais univariadas de valores entre -1 e 1. A maioria dos métodos de clusterização de séries temporais utiliza de forma implícita o conceito de descritor. O *descritor* é uma dupla $< E, d >$, onde E é uma função que extrai um vetor de características V de uma série temporal e d refere-se a uma métrica escolhida para ser usada sobre o espaço vetorial ao qual V pertence. O extrator E serve para transformar os dados brutos em um conjunto de características mais descritivas do evento investigado, enquanto a métrica d serve para detalhar a forma como se avalia a distância (ou similaridade) entre dois vetores característicos obtidos por E.

É comum definir métodos onde E é uma função identidade I, enquanto d é uma métrica adequada a quantizar as dissimilaridades entres séries temporais, como o caso de métodos baseados na *DTW* (*Dynamic Time Warping*) [102]. Também existem trabalhos [114] que usam dados brutos sem nenhuma modificação, o que seria equivalente a usar um descritor $< I, L_2 >$, onde L_2 é a distância usual de espaços reais, também conhecida como distância Euclidiana.

Para os fins deste trabalho usou-se a métrica usual (L_2) sobre um conjunto de características baseadas nos coeficientes do polinômio trigonométrico de ajuste da série temporal. Em outras palavras, tem-se a hipótese:

> **Hipótese 2:** Seja $\mathscr{F}$ uma extração de características baseada no ajuste pela base de Fourier (detalhes na Seção 4.5) de séries temporais do índice *NDVI* e L_2 a distância Euclidiana. As coleções dos vetores V obtidos por um descritor $< \mathscr{F}, L_2 >$ são melhor clusterizáveis que os dados originais, em medidas de qualidade extrínsecas. Isto é, o uso de técnicas de clusterização aplicadas sobre os dados transformados de uma série temporal pode gerar agrupamentos melhores do que o uso destas técnicas nas séries em si.

Essa é uma estratégia comum entre as propostas de agrupamento de séries temporais, que consiste em converter os dados dinâmicos em uma forma de dados estáticos para que se possa utilizar metodologias convencionais. O uso de um sistema ortogonal para projetar dados cronologicamente estruturados, usando a métrica Euclidiana, pode degradar informações que sejam intrínsecas ao desenho da curva formada pela série temporal. Por exemplo, uma translação vertical ou horizontal, em uma série temporal artificialmente gerada por um função $\text{sen}(t)$ pode torná-la mais semelhante a uma reta do que a sua própria forma não transladada.

A abordagem proposta consiste em analisar os coeficientes da curva de ajuste dos pontos da amostra em lugar dos dados originais. Assim, a partir da representação contínua dos dados por regressões por polinômios trigonométricos, onde é dado mais peso às características de oscilação da série do que a seus valores propriamente ditos.

1.4 Contribuições

A principal contribuição deste trabalho é a proposta de um novo algoritmo de agrupamento, o qual se baseia no uso do esquema básico da *VNS* acoplado a variantes que usam restrições (semissuperviosionadas) do algoritmo k-médias. Existem trabalhos que fazem uso da VNS com buscas locais baseadas em heurísticas simples como k-médias ou *k-harmonic means* [3, 17, 52] mas nenhum deles faz uso da incorporação de restrições e nenhum ajusta seu uso para dados dinâmicos.

Além disso, o descritor $< \mathscr{F}, L_2 >$ oferece uma forma de tratar séries temporais sob uma perspectiva funcional, o que permite uma correspondência de utilidade imediata entre dados dinâmicos e métodos de clusterização estática. A semente desta ideia veio da proposta de identificação de áreas plantadas (classificação) com o uso análise harmônica [7, 89] das séries temporais (*Harmonic ANalysis of Time Series*) de *NDVI* [9, 64, 65, 95]. Mas essa proposta não contempla toda a diversidade de possibilidades, como o uso de operadores diferenciais [22], ao analisar a curva de ajuste dos dados ao invés dos dados brutos.

Uma das possíveis aplicações do algoritmo proposto é o monitoramento da cultura da cana-de-açúcar no estado de São Paulo a partir da classificação de imagens de satélite sensor AVHRR/NOAA[4] com ganho de qualidade sobre estratégias tradicionais como o k-médias com

[4] O Radiômetro Avançado de Alta Resolução (*AVHRR*, do inglês *Advanced Very High Resolution Radiometer*) é um sensor orbital a bordo das plataformas orbitais da família *NOAA* (*National Oceanic and Atmospheric Administration*).

múltiplos recomeços (*k-means multistart*). O método proposto corrige o excesso de influência dos pontos iniciais e o inconveniente de convergências à mínimos locais.

Além da contribuição óbvia decorrente do cumprimento dos objetivos, a análise e compilação presentes nesse texto descreve um tema altamente multidisciplinar e uma linha incomum de pesquisa para a área matemática. Disso, têm-se a pretensão de que se possa desenvolver o interesse dos matemáticos brasileiros, que tiverem contato com este texto, a realizarem trabalhos que atuem na interface das áreas de otimização, clusterização e sensoriamento remoto. Ou de forma mais geral, na interface entre pesquisa operacional e mineração de dados com possíveis aplicações à agricultura. Mesmo que as pesquisas em matemática não precisem tradicionalmente serem justificadas por uma aplicação, o país possui prioridades estratégicas que podem levar órgãos de fomento a favorecerem áreas que tenham sua utilidade pública mais claramente justificada.

1.5 Organização da tese

Alguns termos específicos já estão consagrados na literatura em suas formas originais, ou não tem em nosso idioma uma palavra que corresponda exatamente ao vocábulo usado no idioma original. São os casos de *cluster*, *fuzzy*, *k-means*, *VNS* e *outlier*. Na tentativa de preservar ao máximo toda a riqueza de seu sentido, optou-se por manter esses termos em inglês, alternando-os, quando possível, com sua expressão em português. Os termos mantidos em seu formato original foram, na medida do possível, traduzidos ou explicados e sempre tipografados em itálico, atentando para o fato de que a forma aportuguesada, clusterização, foi usada como sinônimo de agrupamento.

Este texto está dividido em seis capítulos e três apêndices. Os apêndices incluíram textos que tornam a tese a mais auto contida possível. A ideia é de que, pelo caráter multidisciplinar do tema, conceitos relativamente introdutórios para alguns segmentos, como a conceituação da *KDD* (introduzido no Apêndice A.1), fossem apresentados de forma breve a especialistas de áreas alheias à ciência da computação, sem a necessidade de consulta externa. Os capítulos centrais da tese são descritos a seguir.

Capítulo 2 - Revisão Bibliográfica. Nesse capítulo, foram descritos os requisitos necessários à compreensão da proposta. Apresentou-se os conceitos fundamentais de sensoriamento remoto (SR) e análise de grupos (*clusters analysis*). Em SR, descreveu-se conceitos básicos necessários para compreender a natureza dos dados de sensores orbitais, principalmente os do tipo AVHRR/NOAA, e a sumarização destes dados em índices vegetativos, em especial o NDVI. Em análise de grupos, apresentou-se, além dos conceitos introdutórios, conceitos mais restritos, como as abordagens mais usuais para agrupamentos de séries temporais e a apresentação de um segmento relativamente novo, conhecido como clusterização semissupervisionada. Enfatizou-se os principais métodos de validação de *cluster*, em especial os de natureza extrínseca.

Capítulo 3 - Clusterização por Meta-heurística. Neste capítulo, foram apresentados os requisitos e o estado da arte em relação aos tópicos necessários à compreensão matemática da proposta. O Capítulo 3 é o capítulo mais representativo deste trabalho, sob o ponto de

vista das teorias que foram efetivamente utilizadas nos algoritmos propostos. Descreveu-se os fundamentos da clusterização por meta-heurísticas, com um breve apanhado sobre meta-heurísticas e a formalização do problema de partição sobre o paradigma de resolução baseado em centroides. Além disso, introduziu-se os principais esquemas do algoritmo VNS e conceitos importantes como estruturas de vizinhança.

Capítulo 4 - VNS básico com busca local e restrições Este capítulo apresenta as propostas de algoritmos com uso do VNS e métodos de busca local com restrições. Além disso apresenta a formalização do problema de agrupamento k sobre o ponto de vista da programação matemática e o detalhamento do método de transformação usado pelo descritor $< \mathscr{F}, L_2 >$.

capítulo 5 - Experimentos Computacionais Neste capítulo, apresenta-se os resultados comparativos entre o algoritmo proposto e os algoritmos amplamente usados na literatura: k-médias com múltiplos reinícios e *COP k-means*. Aplicou-se o algoritmo proposto em bases sintéticas e na base Íris, frequentemente utilizada em testes comparativos. Os dados sintéticos permitiram a compreensão do funcionamento do algoritmo em um domínio de dados visualizável, enquanto a base Íris permitiu a comparação com os diversos outros trabalhos.

Para conclusão, aplicou-se o algoritmo nas imagens NDVI obtidas por composições de máximo valor (*mvc*) da cidade de Jaboticabal, São Paulo, obtidos pelo sensor *AVHRR* a bordo do satélite *NOAA*-17 e nas imagens *mvc*/NDVI do Mato Grosso, obtidas pelo sensor *MODIS* a bordo do satélite *TERRA*. Nestes últimos, os testes aplicados se diferenciam dos demais dados pela natureza tempo-dependente das séries temporais de imagens de satélite. Portanto é para estes dados que se fez uso do descritor $< \mathscr{F}, L_2 >$.

Capítulo 2

Conceitos Básicos

"As the pace of scientific discovery and innovation accelerates, there is an urgent cultural need to reflect thoughtfully about these epic changes and challenges. The challenges of the twenty-first century require new interdisciplinary collaborations, which place questions of meanings and values on the agenda."
- William Grassie

Neste capítulo, abordamos os fundamentos multidisciplinares necessários à compreensão do problema e da proposta apresentadas. Na Seção 2.1 apresentamos aquilo que é necessário para compreender a natureza dos dados de sensores orbitais (principalmente os do tipo *AVHRR/NOAA*) e a transformação de sumarização destes dados, conhecida como índices vegetativos (em especial o *NDVI*). Na seção 2.2 é apresentado, além dos conceitos gerais e de validação, conceitos mais restritos, como as abordagens usadas para agrupamentos de séries temporais e o estado da arte das técnicas de um segmento relativamente novo conhecido como clusterização semissupervisionada.

2.1 Sensoriamento remoto orbital aplicado ao monitoramento de áreas agrícolas

O agronegócio brasileiro movimentou, em 2010, o equivalente a 821 bilhões de reais, aproximadamente 22% do Produto interno Bruto (PIB) [29]. Além disso, seu papel em um contexto global tende a crescer, justificado pelo aumento crescente da população mundial e pelo fato de que os países super populosos terão dificuldades de atender demandas de grãos e fibras devido ao esgotamento de suas áreas agricultáveis. Neste cenário, é imprescindível o desenvolvimento de métodos de estimativa agrícola que sejam mais objetivos, precisos e rápidos. É neste contexto que o caráter multiespectral, sinóptico, repetitivo e global do sensoriamento remoto orbital (SR), em conjunto com tecnologias de geoprocessamento, tem grande potencial de uso em sistemas de estimativas agrícolas [63].

Definição 1 (Sensoriamento Remoto)**.** *O sensoriamento remoto é o processo de captação de informações dos fenômenos e feições terrestres por meio de sensores, sem contato direto com os mesmos, associado a metodologias e técnicas de armazenamento, tratamento e análise destas informações [**?**].*

O sensoriamento remoto orbital se ocupa em medir as propriedades da superfície da terra, sem possuir um vínculo físico com ela, através da análise e processamento das interações entre uma radiação incidente e a área de interesse. Embora sistemas que operam em aeronaves, radiômetros de campo e de laboratório, e sensores fotográficos façam parte do sensoriamento remoto, foram os satélites que se tornaram o instrumento de captação mais comum para análises e pesquisas na área. A Figura 2.1 exemplifica o funcionamento de um sistema de imageamento de SR passivo, ou seja, imageamento de áreas contando com uma fonte de iluminação/radiação natural.

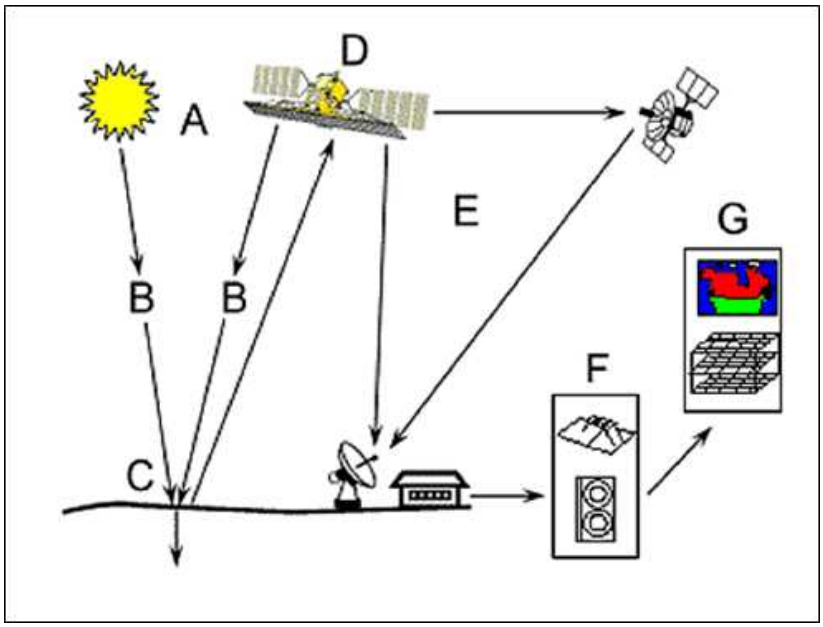


Fig. 2.1: Elementos de um sistema orbital passivo de sensoriamento remoto [28].

A, fonte de energia: um requisito para o funcionamento de um sistema de SR é uma fonte de radiação eletromagnética sobre o alvo. O sol é a a principal fonte para sistemas passivos, pois emite uma iluminação abundante e composta por todas as diferentes regiões do espectro da luz.

B, radiação e atmosfera: a radiação viaja da fonte até o alvo e do alvo até o satélite. É nesse caminho que ocorrem distorções no sinal pela interação da onda eletromagnética com a atmosfera, uma vez que ocorre espalhamento e absorção da radiação por partículas e gases.

C, interação da radiação com o alvo: esse é o momento em que a onda eletromagnética interage com o alvo. É pela caracterização dessa interação que é possível realizar um imageamento do alvo, pois, a partir dos padrões de resposta das medidas de energia refletida ou emitida por estes alvos terrestres, em diferentes comprimentos de ondas, é possível distingui-los.

D, gravação da energia pelo sensor: o sensor embutido no satélite grava a intensidade de certas faixas espectrais da onda que foram emitidas ou refletidas pelo alvo.

E, transmissão, recepção e processamento: os dados brutos são transferidos a uma estação, onde são transformados em uma imagem. A transmissão dos dados pode ser feita imediatamente após a captação ou eles podem ser gravados a bordo e transmitidos posteriormente, ou mesmo transmitidos a outros satélites para que se descarregue os dados em estações específicas. O processamento dos dados brutos visa corrigir distorções atmosféricas e geométricas.

F, interpretação e análise: a imagem processada é interpretada com o fim de se extrair as informações do alvo. É neste estágio que usualmente se ajusta a imagem para atender os objetivos específicos, como integração de dados, filtragem e transformações.

G, aplicação: o propósito de um sistema de SR é que as informações obtidas a partir do acervo de imagens do alvo sejam usadas na compreensão ou resolução de um problema. As imagens podem ser usadas, por exemplo, para gerar mapas temáticos que auxiliem na estimativa de produção e, consequentemente, em políticas públicas.

2.1.1 Comportamento espectral

Os dados brutos de SR consistem, em grande parte, dos registros das intensidades de radiações eletromagnéticas (REM) captadas pelos sensores a bordo dos satélites. Mais especificamente, pode-se calcular as refletâncias das áreas visadas pelo sensor, sendo a refletância a razão entre a radiância refletida pela superfície de um alvo e a irradiância incidente sobre essa superfície [84]. Os diferentes materiais da natureza exibem distintas refletâncias, uma vez que cada material absorve e reflete maiores e menores quantidades de radiação eletromagnética em função de suas constituições físicas, biológicas e químicas. Essa diferença no comportamento de refletância dos objetos caracteriza uma espécie de assinatura espectral, que permite o reconhecimento da cobertura da terra ou o acompanhamento de mudanças de superfície. A curva de refletância espectral (Figura 2.2) mostra como a refletância varia em função comprimento de onda. É dessa forma que, a partir dos dados obtidos por sensores, pode-se inferir o tipo de cobertura presente na área investigada. Para realizar um acompanhamento agrícola, a curva mais relevante é a curva espectral da vegetação.

2.1.2 Satélites e sensores

O satélite é um veículo posto em órbita e é comumente composto por três grandes partes: plataforma, painel solar e carga útil. É na carga útil que se coloca os sensores, antenas e transmissores. Dentre esses, o sensor é responsável pela coleta contínua de propriedades primárias das áreas visadas, i.e. por registrar a radiação refletida e/ou emitida pela superfície. A radiação eletromagnética atravessa um sistema óptico e é focalizada sobre detectores. Estes transformam a radiação em sinais elétricos que são gravados em mídia digital. Neste ponto, a gravação do sinal pode ser influenciada pelas características do sensor, dependendo de suas resoluções espacial, radiométrica, espectral e temporal.

A resolução espacial se refere ao tamanho do pixel da imagem que se forma. Por exemplo, um sensor de resolução espacial de 1,6 km terá pixels de $1{,}6 \times 1{,}6$ km^2 de área. Em outras palavras, o registro da intensidade luminosa captada é associada a uma região definida pela área do pixel.

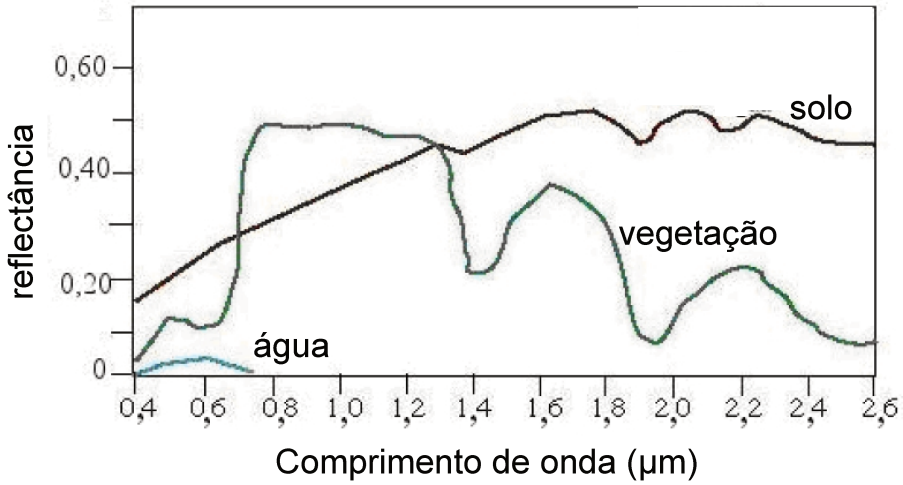


Fig. 2.2: Curva espectral do solo, água e vegetação
Fonte: Adaptado de Introduction to Remote Sensing, James B. Campbell, 2006.

A resolução radiométrica trata dos níveis de intensidade luminosa que se consegue distinguir. Isso se relaciona com quantas variações da radiância espectral recebida o sensor consegue medir. A radiância de cada pixel passa por uma codificação digital, obtendo um valor numérico, expresso em bits, denominado número digital (ND). Este valor é facilmente traduzido para uma intensidade visual ou ainda a um nível de cinza, localizado num intervalo finito $(0, K-1)$, onde K é o número de valores possíveis, denominados níveis de quantização [101].

A resolução espectral se refere as faixas do espectro de luz que o sensor do satélite pode coletar. Existem sensores que captam aspectos termais do alvo, relacionados ao infravermelho longo, enquanto outros tratam todo o espectro visível como uma única banda (pancromático).

A resolução temporal é tão somente o período de revisita. Satélites de alta resolução temporal, como o AVHRR/NOAA passam 2 vezes ao dia sobre o mesmo ponto. A resolução temporal é de grande interesse, especialmente em estudos relacionados a mudanças na superfície terrestre e no seu monitoramento.

2.1.3 Índices de vegetação

Observa-se que o sinal que chega a um sensor multiespectral é uma mistura das interações da luz solar com a vegetação e com elementos de degradação, como condições atmosféricas locais, geometria da iluminação, solos adjacentes e outros. A fim de destacar o brilho da vegetação e atenuar influências da atmosfera, geometria da cena e solo, criou-se os índices de vegetação.

Definição 2 (Índices de vegetação)**.** *Os índices de vegetação são funções das refletâncias de duas ou mais bandas espectrais com o objetivo de estabelecer relações entre os dados espectrais e as características da vegetação.*

Primeiramente propostos como uma simples razão baseada na refletância do infravermelho próximo (IVP) e vermelho (V) [68], os índices de vegetação cresceram em variantes, especialmente nas regiões do visível e do infravermelho próximo, em específico por se relacionarem a parâmetros agronômicos, como índice de área folear e biomassa.

Grande parte dos trabalhos que se propõem a analisar culturas agrícolas e coberturas por vegetação usam o índice de vegetação pela diferença normalizada (*NDVI*, do inglês *Normalized Difference Vegetation Index*) [98].

Definição 3 (*NDVI*)**.** *Sendo ρ_{nir} o valor percentual de refletância do espectro infravermelho próximo e ρ_r o valor percentual de refletância do espectro vermelho, temos:*

$$NDVI = \frac{\rho_{nir} - \rho_r}{\rho_{nir} + \rho_r}, \tag{2.1}$$

onde o valor percentual de refletância é a razão entre a radiância refletida pela superfície do alvo e a irradiância incidente sobre esta mesma superfície.

Gerado a partir da análise de imagens *ERTS*-1, antigo nome do programa *Landsat*, o NDVI foi considerado o índice mais adequado para avaliação das mudanças do vigor vegetal das plantas, mostrando ser exponencialmente relacionado ao índice de área foliar, biomassa vegetal e produtividade [58]. Seus valores variam entre -1 e 1, de forma que valores próximos de 1 indicam áreas altamente vegetadas ou vegetação sadia, enquanto valores próximos de zero, ou negativos, indicam ausência de vegetação ou vegetação com problemas de desenvolvimento.

2.1.4 Imagens digitais

As imagens SR são normalmente compostas por dois arquivos: um cabeçalho da imagem, que contém informações como identificação do satélite e do sensor, data, hora e tamanho do pixel, e o arquivo que comumente é chamado de imagem digital, que contém os valores numéricos correspondentes aos pixels da imagem, referentes à intensidade da REM registrada pelo sensor.

Definição 4 (Imagem digital)**.** *Uma imagem digital pode ser entendida como sendo um conjunto de pontos (pixels), em que cada qual corresponde a uma unidade de informação do terreno, formada através de uma função bidimensional $f(x,y)$, onde x e y são coordenadas espaciais e o valor de f no ponto (x,y) representa, em níveis digitais, o brilho ou radiância da área correspondente ao pixel do terreno. Tanto x e y quanto f assumem somente valores inteiros. Portanto, a imagem pode ser expressa numa forma matricial, onde a linha i e coluna j correspondem às coordenadas espaciais x e y, e f é o nível de cinza do pixel.*

Quanto maior o intervalo de possíveis valores do pixel, maior a sua resolução radiométrica. Quanto maior o número de elementos da matriz por unidade de área do terreno, maior a sua resolução espacial. Os níveis de cinza podem ser analisados através de um histograma, representando a frequência numérica ou porcentagem de ocorrência. A média dos níveis de cinza corresponde ao brilho da imagem, enquanto a variância refere-se ao contraste. Quanto maior a variância, maior será o contraste da imagem.

Além do benefício claro do imageamento de uma área de interesse, o uso de imagens multitemporais permite o estudo da dinâmica de uma vegetação. Para aplicações como a estimativa de produtividade agrícola, é necessário o acompanhamento frequente das culturas agrícolas, daí a demanda por satélites de alta resolução temporal. Plataformas com elevada resolução temporal permitem a avaliação de parâmetros de uma cultura agrícola que se alteram

no decorrer do tempo. O sensor *AVHRR*, a bordo dos satélites da família *NOAA*, por exemplo, possui abrangência espacial, longevidade e baixo custo para aquisição de imagens, o que permite coberturas diárias das culturas de interesse.

O caráter multitemporal permite a geração de perfis temporais de *NDVI*, a partir dos quais se obtêm informações sobre a biomassa ao longo dos seus estágios fisiológicos [55,65,67]. Acrescido a esses benefícios, o grande volume de imagens permite um tratamento mais adequado das distorções causadas por influência de nuvens, variações do ângulo de iluminação solar, efeitos de sombras e geometria de visada, o que é feito através da construção de imagens por composição dos valores máximos (*mvc*, do inglês *maximum value composition*) [57].

2.1.5 Classificação de imagens de satélite

A classificação de imagens de satélite vem concentrando esforços desde de que se tornaram de uso comum e fácil acesso. As imagens AVHRR/NOAA vêm sendo disponibilizadas pela *NASA* desde 1995 para uso do Cepagri/Unicamp, e o INPE vem disponibilizando imagens Landsat em seu site desde 2001. Existem trabalhos que atuam no problema de classificação fazendo uso destes acervos, aplicando técnicas de DM [7, 89, 96], mas nenhum deles atua sobre uma perspectiva de otimização.

A classificação de imagens de satélite é a associação de pontos de uma imagem a uma classe ou grupo de classes. Essas classes representam as feições e alvos terrestres, tais como: água, lavouras, área urbana, reflorestamento e outros. A classificação de imagens é um processo de reconhecimento de classes ou grupos cujos membros exibem características comuns. Uma classe poderia ser, por exemplo, a lavoura de milho e um grupo de classes poderia ser composto pelas áreas cultivadas com milho, soja ou café. Ao se classificar uma imagem, assume-se que os alvos diferentes apresentam propriedades espectrais específicas e que cada ponto pertence a uma única classe (abordagem *crisp*), ou a múltiplas classes com intensidades distintas (abordagem *fuzzy*) [73].

Pontos representativos de uma certa classe devem possuir padrões próximos de tonalidade, cor e textura. A classificação pode ser por inspeção visual, em que o analista interpreta visualmente os elementos da imagem a fim de identificar o cenário ali registrado. Essa interpretação é fortemente dependente do detalhamento espacial da imagem, da qualidade discriminativa do observador e de outros aspectos subjetivos atrelados à sua competência e personalidade. A abordagem subordinada à intervenção humana é comumente dispendiosa, pois exige a presença de um especialista a cada nova imagem, daí a necessidade de abordagens algorítmicas automáticas ou semiautomáticas.

A abordagem algorítmica pode ser dividida em segmentação e classificação. Na segmentação, propõe-se particionar a imagem em regiões, definidas como um conjunto de pixels contíguos, com espalhamento bidimensional, que se assemelhem sob algum critério. Nessa abordagem, os algoritmos usualmente adotam estratégias baseadas em crescimento de regiões ou detecção de bordas [41]. A seleção de uma destas estratégias, ou da combinação delas, depende fortemente dos tipos de dados usados na análise e da área de aplicação.

A abordagem baseada em classificação pode ser dividida em supervisionada, não supervisionada e semissupervisionada. A supervisionada é utilizada quando se tem algum conhecimento prévio sobre as classes na imagem, de modo a ter um conjunto de treinamento definido

por amostras das classes. Estes pontos (áreas amostrais) são utilizados pelos algoritmos de classificação para identificar na imagem os pontos semelhantes às classes do conjunto de treinamento e classificá-los de acordo com essa semelhança. Um método supervisionado relativamente simples é o método do paralelepípedo, que usa o conjunto de treinamento para determinar um intervalo de valores de níveis de cinza das bandas espectrais para cada classe. Por exemplo, quando se utiliza uma imagem com três bandas, a determinação dos intervalos nestas bandas, pelo conjunto de treinamento, define um paralelepípedo tridimensional, em que qualquer ponto da imagem que pertencer a essa região é considerado como pertencente à classe que gerou o paralelepípedo. Esse método de classificação é simples e de rápido processamento computacional. Entretanto, apresenta o inconveniente de aproximar de forma grosseira a assinatura espectral real dos alvos e de sobrepor as classes. Além disso, é fato que as classes, na realidade, não se enquadram em padrões geométricos perfeitos [30].

A abordagem semissupervisionada, assim como a supervisionada, exige algum conhecimento prévio sobre as classes na imagem, como um conjunto de treinamento (classificação semissupervisionada) ou um conjunto de restrições (*clusterização* semissupervisionada). A diferença entre as abordagens supervisionada e semissupervisionada se dá nas metodologias envolvidas e principalmente no tamanho da amostra das classes, usualmente bem menor na semissupervisionada.

A classificação não supervisionada, também conhecida como *clusterização* ou agrupamento, é útil quando não se tem informações relativas às classes de interesse. As classes são definidas automaticamente pelo próprio algoritmo da classificação, a partir de suas características estatísticas ou de distribuição.

Em qualquer destes paradigmas, assume-se que os níveis de cinza podem ser entendidos como variáveis aleatórias z. A ideia é que exista um intervalo de máxima confiança, onde $p(a < z < b) > l_z$, e l_z é um limiar mínimo. O algoritmo mais utilizado neste tipo de classificação é o *k-means*.

2.2 Análise de grupos

Indiscutivelmente, estamos cercados por um grande volume de dados. Técnicas e algoritmos de agrupamento (clusterização) assumem um papel central neste cenário pois são capazes de dar sentido aos dados e de fazer emergir padrões ocultos na pletora de informações disponíveis diariamente. O desenvolvimento de algoritmos de agrupamento é provavelmente o problema mais estudado em mineração de dados [2]. Além de ser escopo de pesquisa em áreas como aprendizagem de máquina e métodos não paramétricos, esses algoritmos têm aplicações em apoio à decisões, por permitir a obtenção automática ou semi-automática de informações a partir de extensos bancos de dados [53].

Métodos para agrupamento de dados são tradicionalmente aplicados para abordar diversos problemas práticos, tais como: segmentação de mercado, bioinformática, processamento de imagens, reconhecimento automático de caracteres (*OCR*, do inglês *Optical Character Recognition*) e busca na internet. É importante notar que algoritmos para agrupamento de dados têm sido estudados por décadas, mas continuam constituindo uma área de pesquisa relevante nos dias atuais, especialmente em áreas do conhecimento que necessitam processar grandes quantidades de dados. De acordo com *Microsoft Academic Research*[1], o número de publicações e citações de trabalhos relacionados ao termo *cluster algorithm* até 2012 foram respectivamente de 13.700 e 101.534, sendo que praticamente metade destes valores foi gerada na última meia década, como podemos verificar na Figura 2.3.

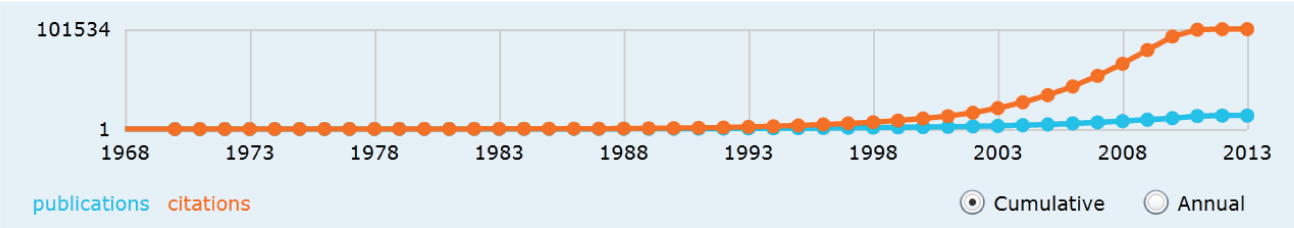


Fig. 2.3: Gráfico do *Microsoft Academic Research* em setembro de 2013, mostrando o crescimento no número de citações e publicações envolvendo o problema de clusterização.

2.2.1 Métodos de clusterização

Múltiplas abordagens do problema de clusterização se originaram em domínios distintos. Por exemplo, técnicas baseadas em árvores (*tree-based techniques*) ou teoria dos grafos são populares na comunidade de aprendizado de máquina (*machine learning*), enquanto métodos de agrupamentos orientados por funções objetivo, ou baseados em protótipos, como o k-médias e a modelagem por mistura Gaussiana, têm sido amplamente estudados e utilizados pela comunidade de reconhecimento de padrões e estatística.

É difícil estabelecer uma linha bem definida de divisão entre os métodos de clusterização, até porque alguns deles possuem características comuns. De forma geral, uma divisão bem aceita e recorrente na literatura é a seguinte:

[1] `academic.research.microsoft.com`.

Métodos partitivos são coleções de técnicas que, a partir de um particionamento inicial, de forma iterativa, mudam os elementos dos grupos buscando a melhoria de alguma medida de qualidade. Pela forma como usualmente estabelecem a dinâmica de atualização das partições, esses métodos são adequados para encontrar agrupamentos inscritíveis em formas esféricas[2], tendo dificuldade para identificar estruturas complexas de agrupamentos, como aquelas em que os envoltórios convexos de vários grupos são sobrepostos. Os algoritmos partitivos usualmente apelam para o conceito de protótipo, em que o conjunto é representado por um elemento que capture bem os aspectos gerais do grupo. Esses protótipos, também conhecidos como centroides, podem ser definidos como o ponto gerado pela média aritmética das coordenadas dos elementos do grupo, ou pelo elemento mais próximo a ela. Os representantes mais citados na literatura são o método das k-médias [79], cuja forma mais básica de funcionamento está descrita na Figura 2.4, e o *k-medoids* [71], em que a escolha dos centroides deve ser feita entre os elementos do grupo.

Vale citar duas contrapartidas *fuzzy* bem conhecidas dos métodos particionais *crisp*, que são os algoritmos *fuzzy c-means* e o *fuzzy c-medoids*.

Métodos hierárquicos consistem em construir grupos a partir de uma abordagem divisiva ou aglomerativa. A abordagem divisiva, também conhecida como *top-down*, admite que todos os elementos pertençam inicialmente a um único grupo. Em cada iteração os grupos existentes vão sendo progressivamente divididos até que cada elemento constitua um único grupo ou que algum critério de parada seja atendido. Já na abordagem aglomerativa, também conhecida como *bottom-up*, adota-se a estratégia inversa, ou seja, todos os elementos são inicialmente considerados grupos e são progressivamente fundidos para, ao final, formar um único grupo, ou atender algum critério de parada. Em ambas as abordagens, representadas na Figura 2.5, é comum usar um dendrograma para exibir a divisão alcançada.

No paradigma hierárquico, os dados são divididos de forma irrevogável, ou seja, não existe revisita de uma solução. Isso pode tornar o método menos adequado, no sentido de avaliar poucas possibilidades, mas pode ser mais adequado em estruturas de dados em que se busca uma taxonomia. O agrupamento gerado por esse paradigma é representado por um dendrograma e não produz uma classificação e sim $n-1$ possibilidades de classificação, pois o número de grupos é definido a *posteriori*.

Algoritmos hierárquicos funcionam bem, apesar de não terem uma justificativa teórica para isso, constituindo uma técnica *ad hoc* de alta efetividade. Representantes clássicos dessa metodologia são o *AGNES* (*AGlomerative NESting*) e *DIANA* (*DIvisive ANAlysis*) [112].

Uma linha de uso de métodos hierárquicos que atenua suas deficiências e melhora a qualidade dos agrupamentos obtidos são as alternativas híbridas que resultam em algoritmos multi-fases como o BIRCH (*Balanced Iterative Reducing and Clustering Using Hierarchies*) [118] e *Chameleon* [70].

[2] Isso significa que existe algum grau de separabilidade entre os conjuntos de forma que uma esfera que inscreva o envoltório convexo de um cluster não se intersectará, ou se intersectará pouco, com as esferas que inscrevam outros clûsteres.

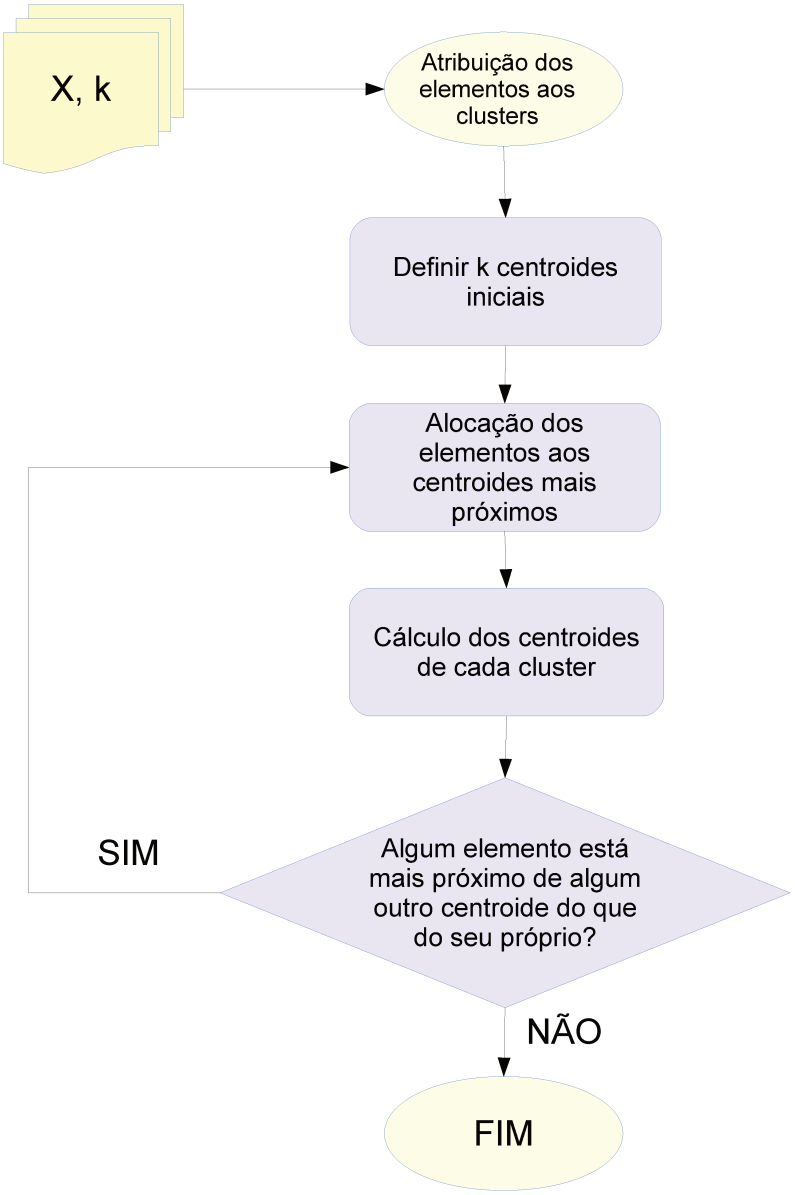


Fig. 2.4: Fluxogramas simplificado do funcionamento do algoritmo k-médias.

Métodos baseados em densidade são aqueles que incorporam elementos a um *cluster* desde que isto não reduza a densidade[3] do grupo para um valor abaixo de um determinado limiar. Algoritmos baseados em densidade admitem que um grupo é uma zona de alta densidade rodeada por uma zona de baixa densidade.

Esse paradigma tem a vantagem de lidar bem com valores discrepantes (*outliers*), reconhecer agrupamentos de formas arbitrárias e não necessariamente atribuir todos os elementos a algum grupo, ao contrário dos métodos partitivos e hierárquicos. Entretanto,

[3]Na abordagem baseada em centro, a densidade de um ponto é definida como o número de elementos, incluindo ele próprio, contidos na hiperesfera de raio ϵ centrada neste elemento.

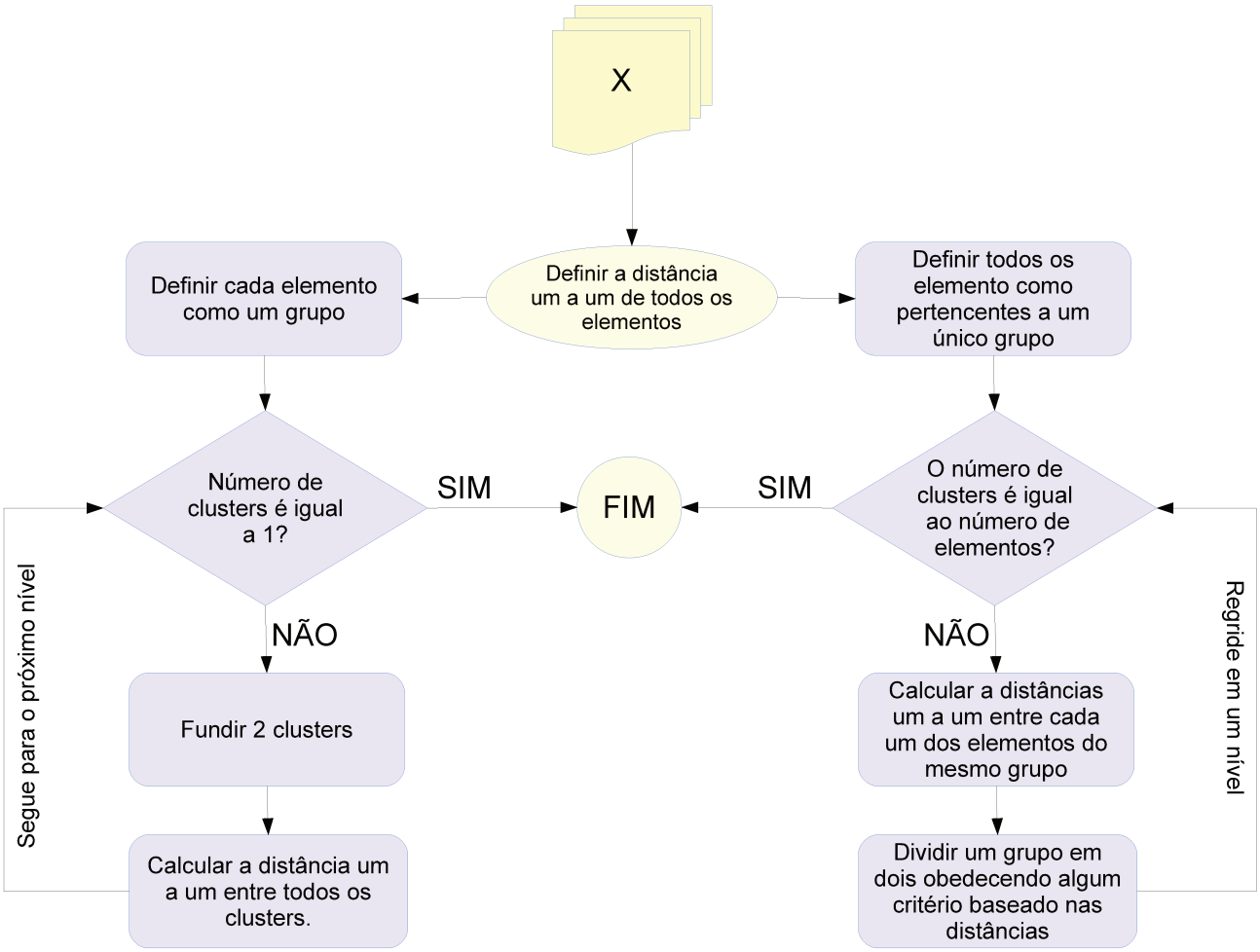


Fig. 2.5: Fluxogramas simplificado do funcionamento dos algoritmos aglomerativos (a esquerda) e divisivos (a direita).

esses métodos enfrentam dificuldades se o conjunto de dados apresenta densidades muito variadas ou alta dimensionalidade, e são computacionalmente caros.

O *DBSCAN* [37], representado na Figura 2.6, classifica como pontos centrais aqueles que têm em sua vizinhança ϵ uma quantidade miníma de pontos ($MinPts$), enquanto os pontos de limite são aqueles que não satisfazem essa condição mas estão na vizinhança ϵ de um ponto central. Os pontos de ruído são os pontos que não satisfazem a condição de vizinhança mínima e não estão na vizinhança de um ponto central.

Os algoritmos *DBSCAN* (*Density-Based Saptial Clustering of Aplications with Noise*) [37], *OPTICS* (*Ordering PoinTs to Identify the Clustering Structure*) [6] e *DENCLUE* (*DENsity-based CLUstering*) [56] são alguns dos representantes mais conhecidos de métodos baseados em densidade.

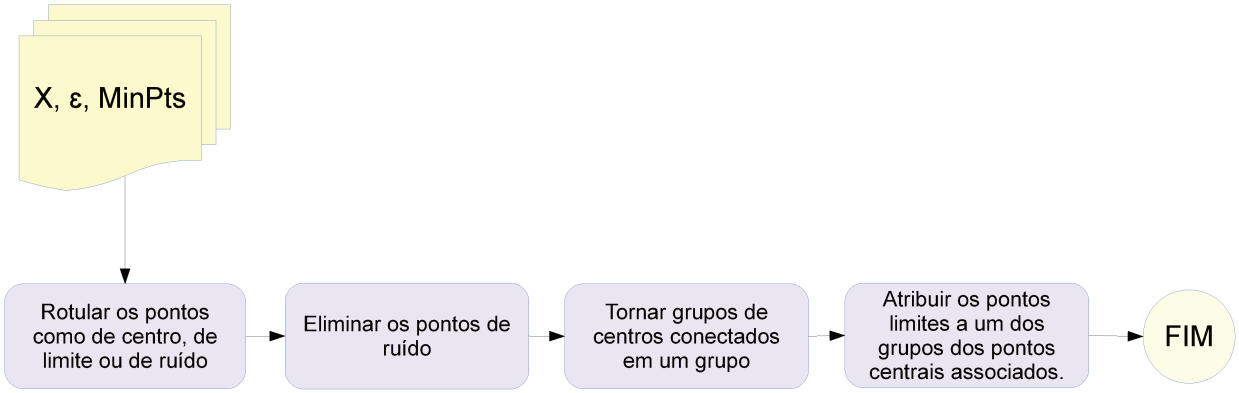


Fig. 2.6: Fluxograma simplificado do funcionamento do *DBSCAN*.

2.2.2 Clusterização de séries temporais

As técnicas de agrupamento podem ser divididas em agrupamentos de dados estáticos e de dados dinâmicos. Diz-se que um dado é estático se o valor de seus atributos são invariantes no tempo. A área de análise de agrupamentos dinâmicos é relativamente pequena quando comparada à análise de agrupamentos estáticos, mas tem se mostrado uma área de crescente interesse. Parte da atração que existe pela área de agrupamento de séries temporais (AST) deriva dos esforços que se realizam em mineração de dados temporais (*temporal data mining*) ou mineração de dados complexos, que compreende a análise de uma variedade grande de estilos de dados, como séries temporais, sequências simbólicas ou sequências biológicas.

O AST é uma tarefa descritiva e não preditiva. Portanto, não se deve compará-lo com metodologias como *long-memory time series modeling*, autoregressão e *ARIMA* (*AutoRegressive Integrated Moving Average*), apesar de existirem trabalhos que introduzem como métrica a distância Euclidiana entre as correspondentes expansões autorregressivas [92]. O objetivo do AST é encontrar características que se destaquem, a fim de gerar grupos de altas similaridades nestas características, e não estimar um valor para extrapolação.

As metodologias mais antigas são baseadas em dados brutos, que agrupam diretamente os dados, tendo suas principais diferenças na modificação das métricas estáticas (medidas de similaridade/dissimilaridade) por métricas adequadas a séries temporais [69, 76, 86].

Metodologias mais recentes se baseiam na transformação dos dados brutos em um vetor de características de dimensão reduzida [47, 103, 110, 115] ou um conjunto de parâmetros [80, 92]. A ideia é aplicar métodos clássicos de agrupamentos estáticos em dados obtidos pela extração das características ou parâmetros das distribuições dos dados originais. Por isso, essas abordagens são comumente chamadas de métodos baseados em características (*feature-based aproach*) e métodos baseados em modelos.

Ao extrair-se as características pretende-se fazer uma transformação dos dados, não só para redução dimensional, mas para que sua nova representação seja mais descritiva das similaridades entre elementos de mesma classe e das dissimilaridades entre elementos de classes distintas. Essa é uma estratégia comum entre as propostas de AST que convertem os dados dinâmicos em dados estáticos e, a partir desses dados transformados, utilizam metodologias convencionais.

As abordagens não paramétricas mais recorrentes da literatura usam, de forma implícita, o conceito de descritor, ou seja, elas podem ser interpretadas como uma combinação da escolha de uma métrica e de uma transformação. Descritor é um termo recorrente em análise de imagens e recuperação de conteúdo, e pode ser definido como uma entidade binária $< E, d >$, onde E é uma função do tipo $E : \mathbb{R}^t \rightarrow \mathbb{R}^m$, com $m \leq t$, que extrai um vetor de características x de uma série temporal, e d refere-se a uma métrica sobre o espaço vetorial no qual x pertence. Disso, temos que uma coleção de séries temporais $S = \{s(t_1), s(t_2), \cdots, s(t_n)\}$ pode ser transformada em uma coleção de vetores de características $\mathbf{X} = \{x_1, x_2, \cdots, x_n\}$, de forma que a distância entre um par (x_i, x_j) qualquer de $\mathbf{X}$ pode ser dada pela função $d(x_i, x_j)$.

Nessa perspectiva existem dois tipos de descritores mais recorrentes:

$< \mathbb{I}, d >$, baseado nos dados brutos, em que a única modificação recai sobre a função de distância, para a qual comumente se adota uma métrica adequada a séries temporais, como por exemplo *DTW* (*Dynamic Time Warping*) [102].

$< E, d >$, baseado em extração de características, onde se usa a métrica d para avaliar as distâncias entre os vetores característicos. É comum adotar a métrica euclidiana (L_2) ou outra métrica induzida pela norma para quantizar as distâncias entre os vetores de características [48].

O descritor $< E, d >$ é tal que E é uma transformação que gera vetores onde suas coordenadas são quantificações mais significativas para o desenho da curva temporal que os valores originais. O descritor pode ter seu foco em caracterizar a oscilação dos valores $x(t)$ e criar conceitos de distinção baseados nessas características, em vez de usar os dados brutos.

2.2.3 Paradigma semissupervisionado

Aprendizado semissupervisionado é o nome que se dá a duas abordagens de aprendizado chamadas classificação e clusterização semissupervisionada, sendo essa segunda também conhecida como clusterização com restrições.

A classificação é comumente entendida como um paradigma de aprendizado supervisionado, um processo de indução lógica em que se produz um modelo a partir de um conjunto de treinamento, para rotular registros não classificados. Entre os principais algoritmos de classificação presentes na literatura pode-se citar: *Decision Trees*, *Naive Bayes classifier*, *SVM*, *KNN*, *Logistic Regression*, *Neural Networks* e *Linear Discriminant Analysis* [50]. Resultados recentes mostram que, na maioria das vezes, o desempenho do classificador supervisionado pode ser melhorado por meio da inclusão de dados não rotulados no processo de geração do modelo. A esse paradigma se dá o nome de classificação semissupervisionada. Os principais algoritmos propostos são:

1. máxima expectativa semissupervisionado (*Semi-supervised EM*) [90],
2. treinamento conjunto (*Co-training*) [19],
3. máquinas de vetor de suporte transdutiva (*Transdutive SVM's*) [42],

4. algoritmos baseados em grafos (*Graph based algorithms*) [100].

A clusterização, ao contrário da classificação, obedece um paradigma não supervisionado. Trata-se de um processo de particionamento de um conjunto de objetos sem rótulos em k *clûsteres*, maximizando a similaridade intra-cluster e minimizando a similaridade inter-cluster. Entre as principais formas de clusterização, encontram-se o algoritmo k-médias, a clusterização hierárquica, a clusterização baseada em densidade e a clusterização espectral [78]. Entretanto, como existem múltiplas formas de agrupamento, o algoritmo pode gerar *clûsteres* que não são adequados.

A clusterização semissupervisionada ou clusterização com restrições se propõe a melhorar o desempenho alcançado pela clusterização não supervisionada adicionando conhecimento de uma pequena porção de dados nomeados. A ideia é adicionar ao processo de clusterização a busca por uma alta consistência entre a partição e o conhecimento do domínio. Essa forma de agrupar não é uma nova forma de classificação, pois admite-se que a quantidade de dados rotulados seja insuficiente para isso. Os principais algoritmos da atualidade utilizam as seguintes estratégias:

1. modificam a função objetivo para premiar rotulações que coincidam com os rótulos dos dados supervisionados, *ex.: Constrained k-means* [23].

2. reforçam as restrições *must-link* (*must be in same cluster*) e *cannot-link* (*cannot be in same cluster*) sobre os dados rotulados, *ex.: COP k-means* [111] .

3. usam os dados rotulados para inicializar a clusterização em algoritmos iterativos (k-means), *ex.: Seeded k-means* [16].

No algoritmo *Seeded k-means*, usa-se os dados rotulados apenas para iniciar o algoritmo, e não nos passos subsequentes. Já o algoritmo *Constrained k-means*, além de usar os dados rotulados para a inicialização, força esses dados a não variarem de rótulo durante o processo iterativo, i.e., somente os dados não rotulados são iterativamente rerrotulados.

No algoritmo *COP k-means*, a inicialização é feita de forma aleatória, mas obedecendo as restrições *must-link*, de forma que registros que devem pertencer ao mesmo *cluster* não podem ser centroides de *clûsteres* diferentes. Durante o processo de rerrotulação, o registro é atribuído ao *cluster* mais próximo, desde que isso não viole nenhuma restrição. Se tal forma de atribuição não existe, o algoritmo para. Tanto o *Constrained k-means* como o *COP k-means* requerem que todas as restrições sejam satisfeitas e podem não ser efetivos quando os pontos de inicialização contêm ruído. O *Seeded k-means* é menos sensível a ruído, mas usa o conhecimento apenas para iniciar o algoritmo. Os experimentos mostram que as variantes semi-supervisionadas do *k-means* superam o *k-means* tradicional.

2.3 Validação de *clûsteres*

Quando um algoritmo produz um agrupamento, devemos nos perguntar se existe alguma forma de avaliar a qualidade do resultado obtido. Agrupamentos distintos podem ser obtidos a partir de algoritmos distintos, e deve-se ser capaz de escolher a melhor resposta. Entretanto, não existe uma

forma inequívoca de dizer qual o melhor algoritmo a partir de um índice universal de qualidade, já que o conceito de qualidade pode variar, dependendo do que se julgar mais importante em um agrupamento.

2.3.1 Validação por visualização

"The greatest value of a picture is when it forces us to notice what we never expected to see."
- John W. Tukey, Exploratory Data Analysis - 1977.

A visualização permite que se percebam padrões e conexões entre números que, de outra forma, estariam dispersos entre vários atributos. Daí, uma forma de avaliar grupos é usando técnicas de visualização para dados multi-dimensionais que reforcem a coerência da partição obtida[4]. Essa forma de avaliar os dados é conhecida como *visual data mining*.

Representação em coordenadas paralelas [62] é uma estratégia que vem ganhando popularidade. Nela, desenha-se d-eixos paralelos e igualmente espaçados e representa-se cada ponto de uma d-upla como uma linha poligonal que liga os valores de seus atributos em cada eixo. Pretende-se, dessa forma, que cada grupo tenha um desenho característico que permita distinguí-lo. Isso se dá pela hipótese de que existe proximidade nos valores dos atributos de membros do mesmo grupo e distinção dos valores de atributo entre membros de grupos distintos.

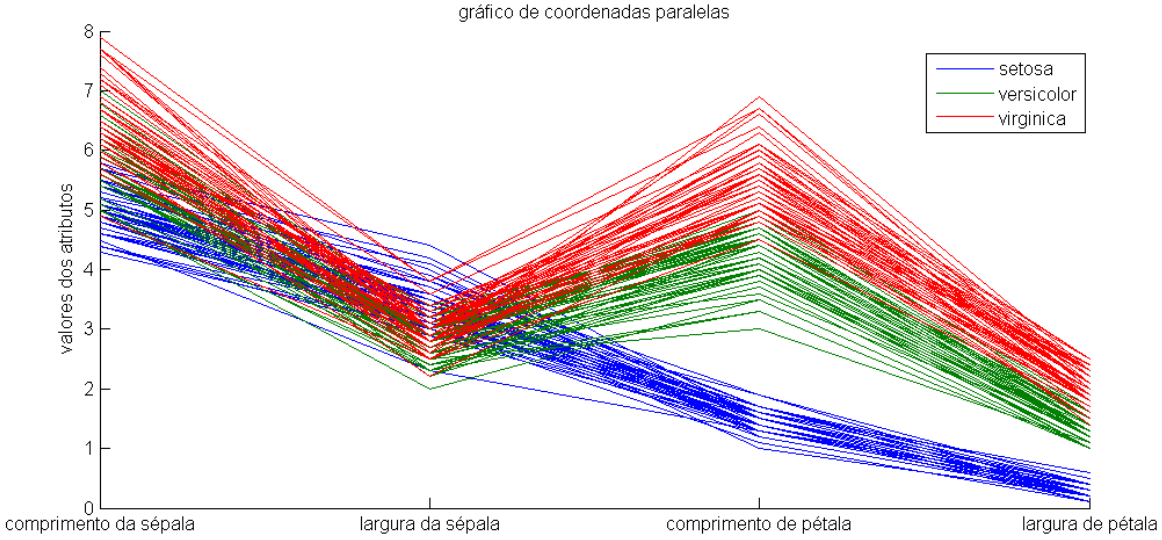


Fig. 2.7: Exemplo de gráfico de coordenadas paralelas para a base de dados Íris de Fisher.

Apesar de existirem esforços no sentido de se aprimorar a técnica [87], há desvantagem no uso de coordenadas paralelas para a representação de dados em bases grandes, pois não existe uma forma eficiente de representar todos os dados de uma única vez para muitas instâncias, já que o gráfico pode ficar poluído e ilegível.

Gráfico de silhueta [99] é outra técnica que vêm se popularizando na comunidade científica[5].

[4] Na hipótese de partição como equivalente a agrupamento, assume-se agrupamentos *crisps* em que cada elemento necessariamente pertence a um *cluster*, o que não é verdade para modelagens *fuzzy* ou agrupamentos por densidade.

[5] O artigo seminal passou de 53 citações em 2004 para 471 em 2012 de acordo com o Microsoft Academic Search.

De uma forma geral, o coeficiente de silhueta de um *cluster* avalia o quão coesos os elementos do mesmo grupo são e o quão separados os elementos de grupos distintos estão. Seja um conjunto de pontos X, divididos em k *clûsteres* $C_1, \cdots, C_k$. Para calcular o coeficiente de silhueta $S(x)$, com $x \in C_i$ $(1 \leq i \leq k)$, usa-se:

$$a(x) = \frac{\sum_{x' \in C_i - \{x\}} ||x - x'||_2}{|C_i| - 1}, \text{ onde } |C_i| \text{ é a quantidade de elementos de } C_i. \tag{2.2}$$

$$b(x) = \min_{j \in \{1,\ldots,k\} - \{i\}} \frac{\sum_{x' \in C_j} ||x - x'||_2}{|C_j|} \tag{2.3}$$

e define-se

$$S(x_i) = \frac{b(x_i) - a(x_i)}{\max\{b(x_i), a(x_i)\}}. \tag{2.4}$$

O valor $a(x_i)$ é uma estimativa da coesão do grupo de x_i dada pela média da distância entre x_i e os demais elementos do seu grupo. O desejado é que ele seja próximo de zero. Por sua vez $b(x_i)$ é a menor média das distâncias de x_i e os demais grupos. Portanto, é desejado que b seja grande.

O valor S é definido entre -1 e 1 e expressa uma ponderação entre esses dois atributos, de forma que *clûsteres* que tenham seus elementos com valores de coeficiente perto de 1 estão bem agrupados. Uma vez calculada a silhueta de cada elemento, o gráfico de silhueta é obtido colocando-se em paralelo barras de tamanho proporcional ao coeficientes de silhueta dos pontos do grupo.

Existem várias outras técnicas de representação de dados [50], como: técnicas de visualização de pixel (*pixel-orietented visualization*), segmentação circular (*circle segment*), visualização de projeção geométrica (*geometric projection visualization*) e matriz de gráficos de dispersão (*scatter-plot matrix*). Existem também técnicas de visualização baseadas em ícones (*icon based visualization*), com representações pouco usuais, como as faces de Chernoff (*Chernoff faces*) ou faces assimétricas de Chernoff (*assymetric Chernoff faces*).

Todas essas técnicas de visualização podem ser ajustadas para avaliar a qualidade dos *clûsteres* obtidos. Cada uma possui pontos fortes e fracos, a depender do tipo de uso. Optou-se por usar as técnicas mais comumente encontradas na literatura: gráfico de silhueta e coordenadas paralelas.

2.3.2 Índices extrínsecos

Formas mais tradicionais de validação de *cluster* remetem ao uso índices intrínsecos e extrínsecos, também chamados respectivamente de índices não supervisionados e supervisionados. O método intrínseco já está, de certa forma, embutido no algoritmo quando se busca, por exemplo, a minimização da soma dos quadrados dos erros (*SSE*, do inglês *sum of squared errors*), além de

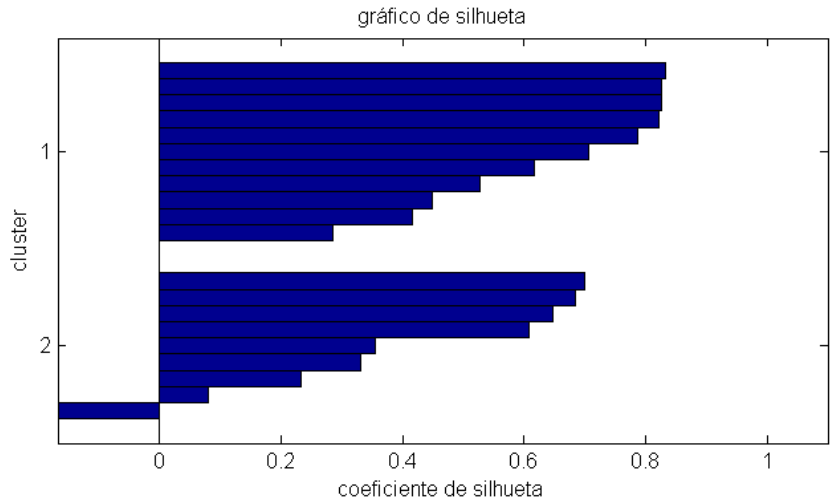


Fig. 2.8: Exemplo de gráfico de silhueta para 20 pontos gerados aleatoriamente em torno dos pontos $(1,1)$ e $(-1,-1)$ e agrupados pelo algoritmo *kmeans*.

estar implícito no gráfico de silhueta. Daí, um contraponto é o uso do método extrínseco, onde se pressupõe conhecimento prévio das classes dos pontos.

As medidas de qualidade supervisionadas são comumente heranças de metodologias aceitas no processo de avaliação de classificadores. Desenvolve-se uma métrica Q que permita comparar o resultado $\mathscr{C}$ de um algoritmo de agrupamento com um agrupamento supervisionado $\mathscr{C}^*$, ou seja, Q gera uma medida de qualidade dos agrupamentos de forma que $Q(\mathscr{C}) \geq Q(\mathscr{C}^*)$ ou $Q(\mathscr{C}) \leq Q(\mathscr{C}^*)$.

É razoável admitir que existem diversas formas de formular Q. Sendo assim, a prática mais usual é definir restrições às métricas de validação dos agrupamentos. Existem quatro restrições que são frequentemente citadas na literatura [5]:

Homogeneidade é a característica responsável por dizer o quanto dos elementos de um grupo pertencem de forma exclusiva a uma única classe, de acordo com $\mathscr{C}^*$. Ou seja, é a restrição que penaliza a confusão intragrupo.

A métrica Q obedece ao critério de homogeneidade se faz grupos de menor confusão intragrupo serem melhor classificados quando comparados com grupos de maior confusão intragrupo.

Completude é uma contrapartida da homogeneidade. Serve para restringir a fragmentação das classes, ou seja, atua no sentido de penalizar agrupamentos que gerem grupos distintos a partir de elementos das mesmas classes.

A métrica Q obedece ao critério de completude se penaliza grupos que separam elementos de um mesmo conjunto.

Grupo de outros (*rag bag*). Em várias situações práticas, é comum ter um grupo que aglomere todos os elementos que não possam ser agrupados com as demais classes (ao menos as dominantes). Essa restrição serve para penalizar a desordem em grupos homogêneos de forma diferente da penalização para grupos altamente heterogêneos.

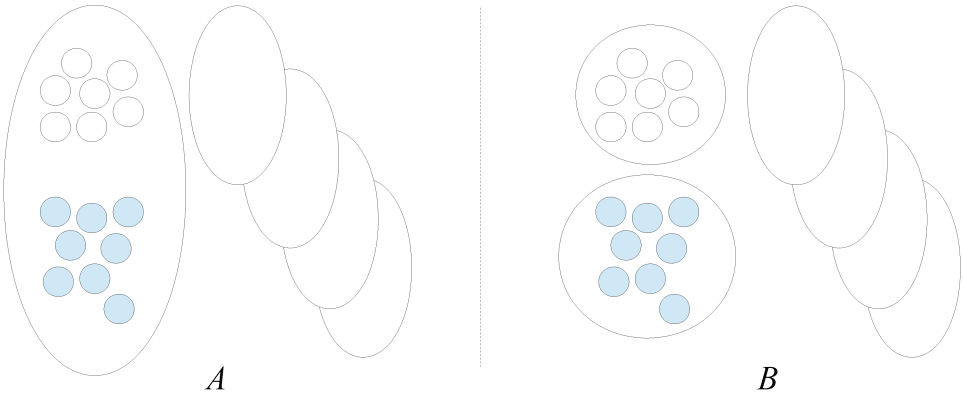


Fig. 2.9: O agrupamento $\mathscr{A}$ tem elementos de dois grupos disjuntos em um mesmo *cluster*, ao contrário do agrupamento $\mathscr{B}$. Uma métrica Q que obedece o critério de homogeneidade terá $Q(\mathscr{A}) < Q(\mathscr{B})$.

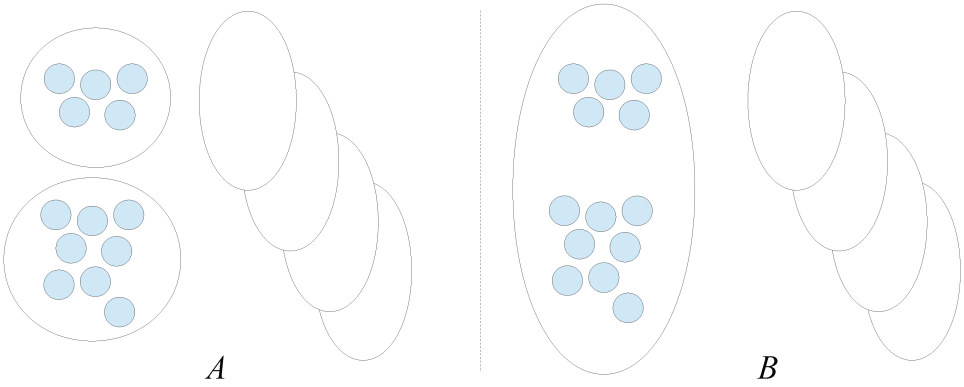


Fig. 2.10: O agrupamento $\mathscr{A}$ separa elementos de um mesmo grupo, ao contrário do agrupamento $\mathscr{B}$. Uma métrica Q que obedece o critério de completude terá $Q(\mathscr{A}) < Q(\mathscr{B})$.

Preservação de *clûsteres* pequenos. Grupos pequenos, quando desmembrados, geram uma penalização maior que quando se desmembra grupos grandes. A ideia é que o impacto de se remover uma certa quantidade de elementos de um grupo pequeno seja maior que remover a mesma quantidade de um grupo grande.

Essas restrições, que emergem de características que julgamos razoáveis de serem encontradas em bons agrupamentos, não são únicas, existindo outras como as restrições de Dom [33] e as restrições de Meila [83]. Entretanto, restringimos-nos neste texto aos critérios mais frequentemente encontrados na literatura para elaboração de métricas.

Várias medidas de qualidade atendem algumas dessas restrições [107], tais como:

- **métricas baseadas em correspondência de conjuntos**, como pureza e pureza invertida;
- **métricas baseadas em contagem de pares**, como estatística aleatória e coeficiente de *Jaccard*;
- **métricas baseadas em entropias**.

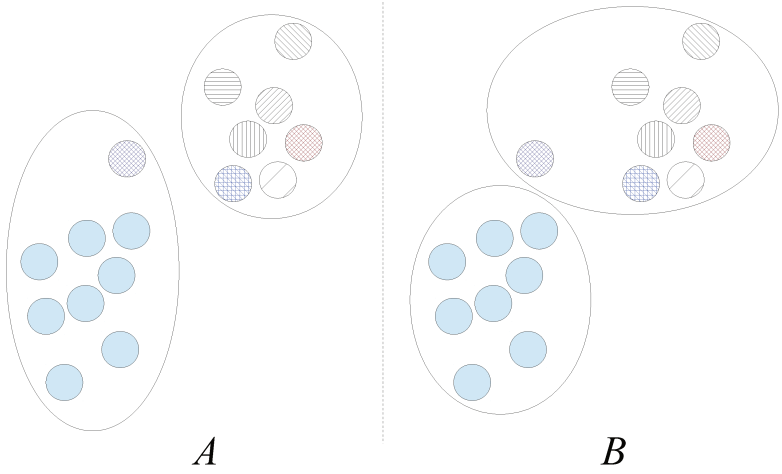


Fig. 2.11: O agrupamento $\mathscr{A}$ incorpora um elemento heterogêneo a um grupo homogêneo. O agrupamento $\mathscr{B}$ cria um grupo exclusivo de termos heterogêneos. Uma métrica Q que cria grupo para elementos não dominantes terá $Q(\mathscr{A}) < Q(\mathscr{B})$.

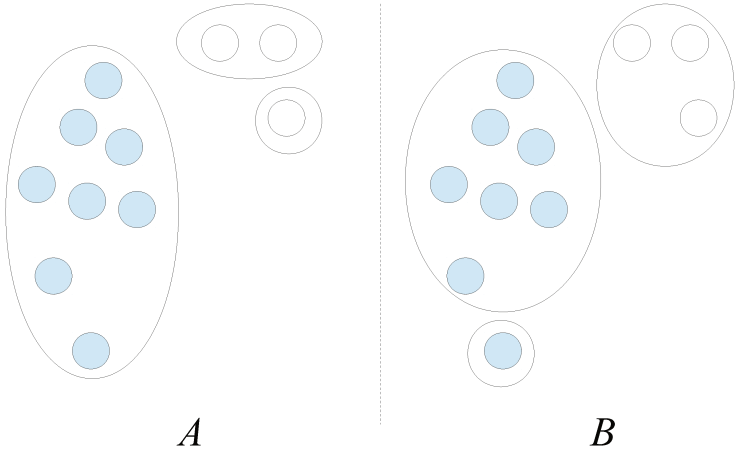


Fig. 2.12: O agrupamento $\mathscr{A}$ quebra o grupo menor em dois, enquanto o agrupamento $\mathscr{B}$ quebra o grupo maior em dois e preserva o grupo menor. Uma métrica Q que obedece o critério de preservação de grupos pequenos terá $Q(\mathscr{A}) < Q(\mathscr{B})$.

Entretanto, as métricas *B-cubed* [11] são as mais satisfatórias por atenderem todos os quatro critérios, ao contrário de todas as outras famílias de métricas [5, 50].

Nesse trabalho, consideramos três métricas da família *B-cubed*: as métricas lembrança (B3R, do inglês *B-cubed recall*), precisão (B3P, do inglês *B-cubed precision*) e F de Rijsbergen (B3F, do inglês *B-cubed Rijsbergen's F*). Essas métricas decompõem o processo de validação na avaliação de cada um dos grupos e classes a partir do princípio da *exatidão*. A ideia é que para cada elemento, a precisão (p) avalie quanto dos demais elementos pertencem à mesma classe e a lembrança (r) diga quantos membros da classe do elemento estão contidos no grupo a que ele pertence.

Uma abordagem formal desses conceitos pode ser feita da seguinte forma: Seja $c(x)$ o grupo, ou classe, ao qual x foi associado, e $g(x)$ a classe à qual ele deveria pertencer de acordo com a verdade terrestre (gabarito). Seja dado um outro elemento x' tal que $x \neq x'$, então

$$exatidão(x, x') = \begin{cases} 1, & \text{se } c(x) = c(x') \text{ e } g(x) = g(x') \\ 0, & \text{caso contrário.} \end{cases}$$

Definindo $p(x_i) = \frac{\sum_{x_j \in C_i} exatidão(x_i, x_j)}{|C_i| - 1}$, em que $C_i = x_i\{x \in \mathbf{X} | c(x) = c(x_i)\}$ é o conjunto dos elementos do mesmo grupo de x_i e $|C_i|$ é a quantidade de elementos de C_i, então a precisão (B3P) de um agrupamento é dada por:

$$P(\mathscr{C}) = \frac{1}{n} \sum_{i=1}^{n} p(x_i), \tag{2.5}$$

Da mesma forma, definindo $r(x_i) = \frac{\sum_{x_j \in G_i} exatidão(x_i, x_j)}{|G_i| - 1}$, sendo $G_i = x_i\{x \in \mathbf{X} | g(x) = g(x_i)\}$ o conjunto dos elementos de mesma classe de acordo com o gabarito e sendo $|G_i|$ a quantidade de elementos de G_i, temos que a lembrança (B3R) de um agrupamento é dada por:

$$R(\mathscr{C}) = \frac{1}{n} \sum_{i=1}^{n} r(x_i) \tag{2.6}$$

Uma forma comum de combinar índices consiste em usar o coeficiente F de Van Rijsbergen, que é obtido da seguinte forma:

$$F(R, P) = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}}, \tag{2.7}$$

sendo R e P as métricas avaliadas e α e $(1 - \alpha)$ os seus pesos relativos. A medida conhecida como *B-cubed* F é definida usando-se $\alpha = 0,5$, o que a torna na média harmônica entre precisão e lembrança, i.e. $2PR/(P + R)$.

Apesar de não atenderem todas as restrições declaradas, as métricas baseadas em entropia também têm sido amplamente adotadas em validação de classificadores. A entropia total de um agrupamento [105] é a média ponderada das entropias E_j de cada grupo, uma medida da confusão do grupo j dada por:

$$E_j = \sum_{i=1}^{k} \frac{\eta_{ij}}{|G_j|} \log_2 \left(\frac{\eta_{ij}}{|G_j|} \right), \tag{2.8}$$

sendo η_{ij} o número de elementos da classe i no grupo j, de acordo com o gabarito.
Dessa forma, temos que a entropia total de um agrupamento é definida como:

$$entropia(\mathscr{C}) = \sum_{j=1}^{k} \frac{|G_j|}{n} E_j, \tag{2.9}$$

A entropia é uma medida negativa (ou seja, quanto maior, pior) que avalia como os elementos de uma classe se distribuem pelos grupos. Além desta, existem outras métricas baseadas em entropia, como entropia de classe (*class entropy*) [12], variação de informação (*variation of information*) [117] ou medida$-V$ [97].

Capítulo 3

Clusterização por Meta-Heurística

"It is common sense to take a method and try it. If it fails, admit it frankly and try another. But above all, try something."
- Franklin D. Roosevelt

Uma meta-heurística é um procedimento de alto nível, não determinístico e de uso extensivo (ou seja, não é dependente do problema), que se caracteriza por guiar o processo de busca para encontrar soluções aproximativas sub-ótimas. Existem milhares de opções baseadas nas visões tradicionais de clusterização, mas alternativas baseadas em clusterização por meta-heurística são relativamente reduzidas[1]. Neste capítulo, apresenta-se os fundamentos da clusterização por meta-heurística, em específico a VNS, como arquitetura para resolver o problema de agrupamento geral.

3.1 Formalização do problema

Um passo importante no uso de meta-heurísticas para clusterização é a formulação do problema de agrupamento. O problema aqui tratado pode ser inicialmente apresentado como um problema específico de agrupamento k em que os elementos a serem agrupados são séries temporais discretas e univariadas ($\gamma(t) \in \mathbb{R}$) em períodos de tempo equidistantes, ou seja, é um problema de agrupamento de d-uplas reais. Disso, pode-se apresentar este problema da seguinte forma:

Definição 5 (problema de agrupamento k)**.** *Seja* $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, *onde* $\mathbf{x}_i = (x_{i1}, \cdots, x_{id}) \in \mathbb{R}^d$. *O problema de agrupamento consiste em obter uma partição* $\mathbf{P}^*$ *de* $\mathbf{X}$ *em k subconjuntos que atenda um determinado critério de qualidade Q, de forma que* $Q(\mathbf{P}^*) \geq Q(\mathbf{P})$, $\forall \mathbf{P} \in \mathscr{P}$.

Usou-se $\mathscr{P}(\mathbf{X}, k) = \mathscr{P}$ para representar o espaço de soluções possíveis, i.e. $\mathscr{P}$ é o conjunto de todas as possíveis partições de $\mathbf{X}$ em k subconjuntos, onde se define uma partição $\mathbf{P} =$

[1] O termo "`clustering techniques`" retornou 84.800 ocorrências em consulta realizada em 27 de novembro de 2013 pelo google acadêmico (`scholar.google.com`), enquanto o termo "`metaheuristic clustering`", na mesma data, retornou 110 ocorrências. Isso equivale a uma relação inferior a 0,14%, o que atesta a incipiência da área de clusterização por meta-heurísticas em relação à área de clusterização.

$\{C_1, C_2, \cdots, C_k\}$ como um conjunto de k subconjuntos C_j de X que satisfaçam a seguinte propriedade:

Propriedade 1 (regras para conjuntos formadores de partição)**.** *Os conjuntos formadores de uma partição devem necessariamente obedecer às seguintes regras:*

i. $C_i \neq \emptyset$,

ii. $\bigcup_{i \in [k]} C_i = \mathbf{X}$,

iii. $C_i \cap C_j = \emptyset, \forall\, i \neq j$.

O problema de dividir n objetos em k grupos é um problema combinatorial. Sendo assim, uma abordagem baseada em uma busca exaustiva é infactível, pois a cardinalidade $|\mathscr{P}|$ do conjunto de partições de $\mathbf{X}$ em k clusters é dada por um número de segunda ordem de Stirling , i.e.

$$|\mathscr{P}| = \left\{ \begin{array}{c} n \\ k \end{array} \right\} = \frac{1}{k!} \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} j^n.$$

Para um simples exemplo de agrupamento binário, em uma base de 100 registros, um método baseado em força bruta teria que avaliar um espaço da ordem de 10^{30}. Um computador capaz de realizar uma avaliação a cada 10^{-10} segundos gastaria mais de 400 vezes a atual idade da Terra para esgotar todas as possibilidades. Pela Figura 3.1, fica claro o crescimento explosivo do número de partições à medida que se aumenta o número de elementos n. Por isso, há a necessidade de estabelecer uma heurística que encontre uma solução próxima à ótima, sem visitar todas as soluções do espaço de busca.

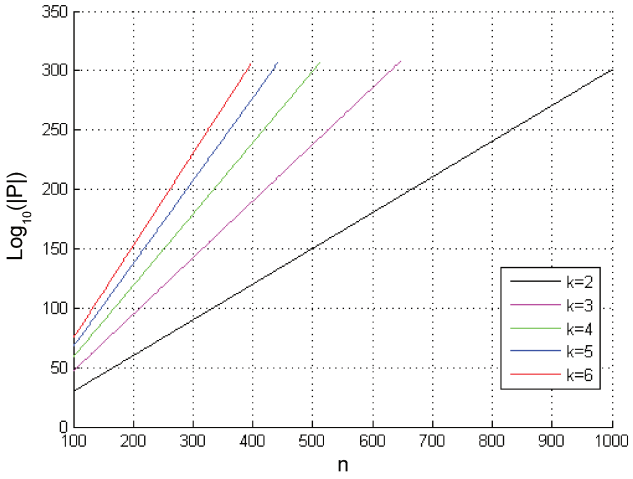


Fig. 3.1: Variação de $|\mathscr{P}|$, quantidade de partições, com k entre 2 e 6.

3.1.1 Centroides

Uma forma comum de clusterização consiste em agrupar os elementos por sua similaridade a um modelo que seja uma representação sinóptica dos grupos (*prototype based clustering*). Em outras palavras, usamos um único vetor $\mathbf{c}_j$ como um protótipo que represente o *cluster* $\mathbf{C}_j = \{\mathbf{x}_{j_1}, \ldots, \mathbf{x}_{j_p}\}$. Um candidato comum para protótipos são os **centroides** de agrupamentos. Em uma visão geométrica, um centroide pode ser entendido como um ponto ao qual os elementos a ele associados estão mais próximos, sendo portanto, semelhantes. Assim, cada grupo pode ser definido por seu centroide $\mathbf{c}_j \in \mathbb{R}^d$, sendo cada instância associada ao centroide mais próximo. Daí, o agrupamento é definido pela matriz $C = [\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_k]^T$ de ordem $k \times d$.

Para apresentarmos a equação geral dos centroides vamos definir primeiramente as funções pertinência (m) e peso (w). A função $m(\mathbf{c}_j|\mathbf{x}_i)$ é uma medida de pertinência de $\mathbf{x}_i$ ao cluster C_j, de modo que $m(\mathbf{c}_j|\mathbf{x}_i) \geq 0$ e $\sum_{j=1}^{k} m(\mathbf{c}_j|\mathbf{x}_i) = 1$. Essa função define se a abordagem será *fuzzy*, com $m(\mathbf{c}_j|\mathbf{x}_i)$ assumindo qualquer valor no intervalo $[0,1]$, ou *crisp*, caso em que $m(\mathbf{c}_j|\mathbf{x}_i)$ assuma necessariamente o valor 0 ou 1.

A função $w(\mathbf{x}_i)$ é uma medida do impacto que $\mathbf{x}_i$ tem no cálculo do centroide $\mathbf{c}_j$ e serve para contornar o efeito negativo de valores discrepantes. Como exemplo, no algoritmo das k-médias harmônicas (*k-harmonic means*), $w(\mathbf{x}_i)$ é baseada no inverso da distância em uma função objetivo baseada na média harmônica. Essa mudança na forma de calcular o centroide torna o *k-harmonic means* menos sensível aos pontos de inicialização e a *outliers*.

Disso, temos que os centroides podem, de forma geral, ser definidos pela expressão:

$$c_j = \frac{\sum_{i=1}^{n} m(\mathbf{c}_j|\mathbf{x}_i)\, w(\mathbf{x}_i)\mathbf{x}_i}{\sum_{i=1}^{n} m(\mathbf{c}_j|\mathbf{x}_i)\, w(\mathbf{x}_i)}. \tag{3.1}$$

Para o método usual do k-médias, temos $w(\mathbf{x}_i) = 1$ e $m(\mathbf{c}_i|\mathbf{x}_j) = \delta_{jl}$, onde δ_{jl} é o delta de Kronecker e $l = \underset{1 \leq j \leq k}{\operatorname{argmin}} ||\mathbf{x}_i - \mathbf{c}_j||$. Para o algoritmo *Expectation-Maximization* [32], temos $w(\mathbf{x}_i) = 1$ e

$$m(\mathbf{c}_j|\mathbf{x}_i) = \frac{p(\mathbf{x}_i|\mathbf{c}_j)\, p(\mathbf{c}_j)}{p(\mathbf{x}_i)},$$

onde $p(\mathbf{x}_i|\mathbf{c}_j)$ é a probabilidade de $\mathbf{x}_i$ ter sido gerada por uma distribuição Gaussiana de centro $\mathbf{c}_j$, e $p(\mathbf{c}_j)$ é a probabilidade a priori do centro $\mathbf{c}_j$ [49]. Já para o *k-harmonic means*, temos:

$$m(\mathbf{c}_j|\mathbf{x}_i) = \frac{||\mathbf{x}_i - \mathbf{c}_j||^{-p-2}}{\sum_{j=1}^{k} ||\mathbf{x}_i - \mathbf{c}_j||^{-p-2}}, \tag{3.2}$$

$$w(\mathbf{x}_i) = \frac{\sum_{j=1}^{k} ||\mathbf{x}_i - \mathbf{c}_j||^{-p-2}}{\left(\sum_{j=1}^{k} ||\mathbf{x}_i - \mathbf{c}_j||^{-p}\right)^2}, \tag{3.3}$$

onde p é um parâmetro de entrada, usualmente maior ou igual a 2.

A forma como se define os centroides altera a função critério que orienta a qualidade do *cluster*. Independentemente das métricas adotadas, os métodos baseados em centroides compartilham a seguinte propriedade:

Propriedade 2 (critério de associação)**.** *Sejam* $\mathbf{c}_i$ *e* $\mathbf{c}_j$*, com* $i \neq j$*, dois centroides no espaço* $\mathbb{R}^n$*, respectivamente associados aos clusters* C_i *e* C_j*. Nesse caso, temos* $d(\mathbf{x}, \mathbf{c}_i) < d(\mathbf{x}, \mathbf{c}_j) \iff \mathbf{x} \in C_i$*.*

Mesmo usando o conceito informal de centroide como um ponto qualquer (não necessariamente pertencente a X) que minimize os desvios entre os elementos do grupo e seu representante, acabamos chegando à forma 3.1. Como exemplo, o centroide do espaço usual (espaço Euclidiano) acaba por ser o vetor de médias dos elementos de cada grupo, que equivale a $m(\mathbf{c}_j|\mathbf{x}_i) = 1$ e $w(\mathbf{x}_i) = 1$.

Teorema 1 (Centroide do espaço Euclidiano)**.** *O centroide* $\mathbf{c}^*$ *do grupo* $\mathbf{C_i} \subset \mathbb{R}^d$*, de norma* $||\cdot||_2$*, dado por*

$$\mathbf{c}^* = \frac{1}{|\mathbf{C_i}|} \sum_{\mathbf{x} \in \mathbf{C_i}} \mathbf{x},$$

minimiza o somatório dos desvios quadráticos $SSE(\mathbf{c}) = \sum_{\forall \mathbf{x} \in \mathbf{C_i}} ||\mathbf{x} - \mathbf{c}||_2^2$*. Ou seja, o ponto do envoltório convexo de* $\mathbf{C_i}$ *que minimiza a variância intracluster é o centroide definido pelas médias dos elementos de* $\mathbf{C_i}$*.*

Demonstração. A função $SSE(\mathbf{c})$ é convexa. Portanto, é suficiente resolver a equação $\nabla SSE(\mathbf{c}) = \mathbf{0}$ para assegurar o resultado acima. Mesmo para normas arbitrárias, a função $SSE(\mathbf{c})$ é convexa e, com algumas exceções, tem um único minimizador. □

Por todos estes aspectos, o centroide é o ponto do envoltório convexo que melhor sintetiza as propriedades e características do grupo. Baseado nessa visão, o problema de dividir n pontos em um espaço real d-dimensional $\mathbb{R}^d$ em k grupos pode ser formulado da seguinte maneira:

Definição 6 (Formulação do problema de agrupamento baseado em centroides)**.** *Determinar um conjunto* $C^* = \{\mathbf{c}_1, \cdots, \mathbf{c}_k\}$ *de centroides que formem k subconjuntos do tipo*

$$C_j = \left\{ \forall x \in X \middle| \operatorname*{argmin}_{1 \leq l \leq k} d(\mathbf{x}, \mathbf{c}_l) = j \right\},$$

de forma que os conjuntos C_j *satisfaçam a propriedade 1 e que* C^* *atenda um critério de qualidade Q, de forma que* $Q(\mathbf{C}^*) \geq Q(\mathbf{C})$ *para todo C, onde C é um conjunto de centroides qualquer.*

É importante observar que nem todo conjunto de pontos do envoltório convexo de $\mathbf{X}$ é um conjunto de centroides. Isto decorre do fato de que pontos arbitrários do envoltório convexo podem induzir a um agrupamento degenerado que não estabeleça uma partição. Por exemplo, os pontos $C = \{2, 5, 9\}$ não induzem a uma partição válida para o conjunto $X = \{1, 3, 10\}$. Esse conjunto de não centroides, ou conjunto de centroides degenerados, pode ser considerado uma solução infactível para o problema de agrupamento. Além disso, acrescenta-se que não existe unicidade entre partições e conjunto de centroides, ou seja, os centroides que induzem uma determinada partição não necessariamente são únicos. Por exemplo, para o conjunto $X = \{1, 3, 10\}$, os centroides $C = \{1, 10\}$ e $C' = \{2, 9\}$ induzem a mesma partição $\{\{1, 3\}, \{10\}\}$.

Vale citar que o problema de agrupamento baseado em centroides possui correspondência com problemas já tratados na área de otimização, como: localização de instalações, diagramas de *Voronoi*, árvore de extensão mínima, triangularização de *Delaunay* e problema de *Weber*. Muitos resultados obtidos para esses problemas podem ser diretamente aplicados ao problema de agrupamento.

3.1.2 Caracterização dos espaços de busca

Algoritmos de clusterização fazem suposições sobre o conjunto de dados. O método hierárquico, por exemplo, assume que exista uma hierarquia na organização dos dados, e os métodos baseados em densidade admitem que existem faixas de baixa densidade entre as classes. Para compreender as implicações do uso de centroides, é preciso caracterizar as particularidades e os tipos de soluções que se pode obter a partir do seu uso. O problema geral de busca por uma partição (Definição 5) se propõe a encontrar um agrupamento qualquer, desde que respeite as propriedades de partição (Propriedade 1). Entretanto, a formulação baseada em centroides (Definição 6) busca por agrupamentos que derivem de partições que são induzíveis por centroides, o que implicitamente impõe hipóteses adicionais.

Considerando $\mathscr{P}'$ o conjunto de todas as partições possíveis de se obter a partir de um conjunto de centroides e $\mathscr{P}$ o conjunto de todas as partições possíveis, então $|\mathscr{P}'| \leq |\mathscr{P}|$. Seja $X = \{1,2,3\}$ e $k = 2$, não existe nenhum conjunto de centroides que induza [2] a partição $P = \{\{1,3\}\{2\}\}$, i.e. $\nexists C \Rightarrow (C \rightarrow P)$. Portanto, o uso de centroides leva a uma busca sobre um espaço reduzido, que claramente despreza partições que não sejam induzíveis por centroides. De fato, os grupos gerados por centroides são tais que os politopos formados pelos envoltórios convexos de cada *cluster* não se interseccionam. Essa característica é equivalente a dizer que os conjuntos formados são dois a dois linearmente separáveis.

Definição 1 (grupos linearmente separáveis)**.** *Dois grupos C_i e C_j são ditos linearmente separáveis se existe ao menos um hiperplano $\mathscr{H} = \{\mathbf{x} \in \mathbb{R}^d | \mathbf{w}^t\mathbf{x} + b = 0\}$, de forma que $\forall \mathbf{x} \in C_i$ e $\forall \mathbf{y} \in C_j$ temos $(\mathbf{w}^t\mathbf{x}) \cdot (\mathbf{w}^t\mathbf{y}) < 0$.*

Decorre que o conjunto $\mathscr{P}'$ é composto exclusivamente por partições que sejam linearmente separáveis, também conhecidas como partições de Voronoi.

O algoritmo das k-médias e suas variantes exploram o espaço solução por meio de um processo de indução alternada entre P e C. Pelo que foi dito acima, sabe-se que essa forma de gerar agrupamentos nunca irá encontrar alternativas que não sejam linearmente separáveis.

Entretanto, para problemas onde existe pouca confusão nas fronteiras de decisão entre uma classe e outra, essa hipótese não é um inconveniente. Além do mais, existem soluções baseadas em lógica *fuzzy* que flexibilizam a característica rígida das fronteiras de decisão.

[2] Por brevidade $C \rightarrow P$ significa que a partição P é induzida pelo conjunto de centroides $C = \{c_1, \ldots, c_k\}$ e $P \rightarrow C$ o contrário, quando P induz C. Nem sempre existe uma autoindução, $C \rightarrow P \rightarrow C$ ou $P \rightarrow C \rightarrow P$ e, na verdade, este é o caso de convergência para alguns algoritmos baseados em centroides.

3.2 Meta-heurísticas

A raiz da palavra *heurística* é comum à palavra *eureka* e significa descobrir. *Heurística*, dentro da área de otimização, é o nome dado ao conjunto de estratégias não determinísticas de busca de mínimos locais ou globais. Os métodos heurísticos são usados quando não é viável caracterizar o domínio de aplicação para uso de métodos determinísticos ou quando se quer acelerar a obtenção de uma solução, ainda que subótima.

As meta-heurísticas, em sua definição original, são métodos de busca de solução que combinam procedimentos de melhorias locais, como as heurísticas, e estratégias de alto nível, para criar um processo capaz de escapar de ótimos locais e realizar uma busca robusta no espaço de solução. Ao longo do tempo, esses métodos também têm sido usados para se referir a todos os procedimentos que utilizam estratégias para fuga de ótimos locais em espaços de soluções complexas.

Um grande número de ferramentas e mecanismos que surgiram a partir da criação de métodos meta-heurísticos provaram ser extraordinariamente eficazes, tanto que as meta-heurísticas têm-se tornado a linha preferencial de ataque para resolver problemas complexos, em especial os de natureza combinatorial. Embora as meta-heurísticas sejam incapazes de certificar que as soluções que encontram sejam ótimas, os procedimentos exatos, quando aplicáveis, muitas vezes mostram-se incapazes de encontrar soluções cuja qualidade é comparável às obtidas pelas principais meta-heurísticas, particularmente para os problemas do mundo real, que muitas vezes atingem elevados níveis de complexidade. Além disso, algumas aplicações bem sucedidas vêm incorporando estratégias de meta-heurísticas a métodos exatos [93] por meio de abordagens híbridas.

As meta-heurísticas e as heurísticas são usualmente simples de implementar e compreender, tendo baixo custo computacional, embora nem sempre garantam soluções globalmente ótimas. Além disso, por sua natureza experimental, seus resultados teóricos quase sempre se resumem à análise da convergência local [20].

As meta-heurísticas têm crescido em interesse e variedade, de modo que, hoje, têm-se um grande acervo de métodos divididos em múltiplos paradigmas de busca, que podem variar de conceitos mais simples, como o da busca local iterada (*iterated local search*), até conceitos mais complexos, como as estratégias bio-inspiradas. As principais meta-heurísticas da atualidade surgiram em trabalhos publicados principalmente a partir da década de 70, como se pode ver na Figura 3.2.

3.3 Busca em vizinhança variável

A busca em vizinhança variável, também conhecida como VNS (do inglês *variable neighborhood search*), foi proposta em 1997 por Nenad Mladenović e Pierre Hansen no artigo seminal *Variable neighborhood search* [85]. A VNS é uma meta-heurística simples, coerente, eficiente e altamente precisa em vários domínios de aplicação, sendo frequentemente usada par resolver problemas da forma:

$$\min\{f(x)|x \in \Omega, \Omega \subset S\}, \tag{3.4}$$

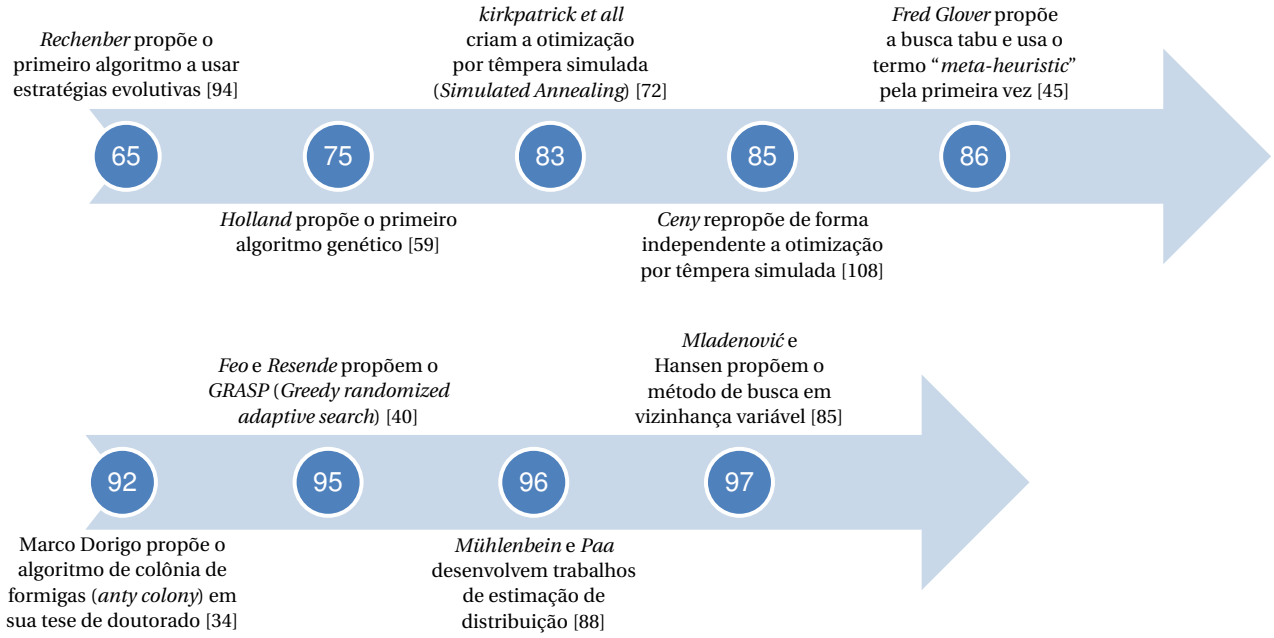


Fig. 3.2: Cronologia das meta-heurísticas que têm sido as linhas mais comuns de pesquisa nos últimos 50 anos.

em que $f : S \subset \mathbb{R}^d \to \mathbb{R}$, S é normalmente um espaço grande, mas finito, e Ω é o conjunto das soluções factíveis.

Todas as meta-heurísticas têm seus pontos fortes e fracos, mas a VNS é a que melhor explora o conceito de vizinhança sem apelar para uma sofisticação muitas vezes desnecessária. Ela se baseia na mudança sistemática de vizinhanças, alternando uma fase de descida, em busca de um ótimo local, e uma fase de pertubação para fuga de vales.

3.3.1 Elementos de VNS

Para compreender a meta-heurística VNS é imprescindível que se entenda seus princípios e suas estruturas, e um elemento central da VNS é o conceito de vizinhança. A estrutura de vizinhança de $\mathbf{x}$ é uma coleção de pontos factíveis próximos a $\mathbf{x}$, e esse conceito de proximidade pode variar dependendo da métrica d ou quase-métrica escolhida. Um exemplo de estrutura de vizinhança aninhada, onde $N_i(\bar{\mathbf{x}}) \subseteq N_{i+1}(\bar{\mathbf{x}})$, é o conjunto $N_i(\bar{\mathbf{x}}) = \{\forall \mathbf{x} \in \Omega \,|\, d(\mathbf{x}, \bar{\mathbf{x}}) \leq i\}$.

Os preceitos gerais que governam os esquemas derivados da VNS são:

Propriedade 3 (princípios da VNS)**.**

(i) Um mínimo local com relação a uma estrutura de vizinhança não é necessariamente um mínimo local com relação às outras estruturas de vizinhança;

(ii) Um mínimo global é um mínimo local com relação a quaisquer estruturas de vizinhança; e

(iii) Para um grande número de problemas, mínimos locais com relação a uma, ou a várias vizinhanças, são relativamente próximos.

Os dois primeiros princípios são relativamente intuitivos, enquanto o último é conhecido, em outras meta-heurísticas, como princípio de otimalidade próxima e se baseia na ideia empírica de que mínimos locais comumente fornecem algum tipo de informação sobre o mínimo global. A VNS parte de uma solução $\mathbf{x} \in \Omega$ e investiga sua vizinhança $N_i(\mathbf{x})$ afim de encontrar uma solução melhor. Em seguida, inicia-se uma busca centrada nesta nova solução, ou parte-se para a investigação de uma nova vizinhança (Figura 3.3).

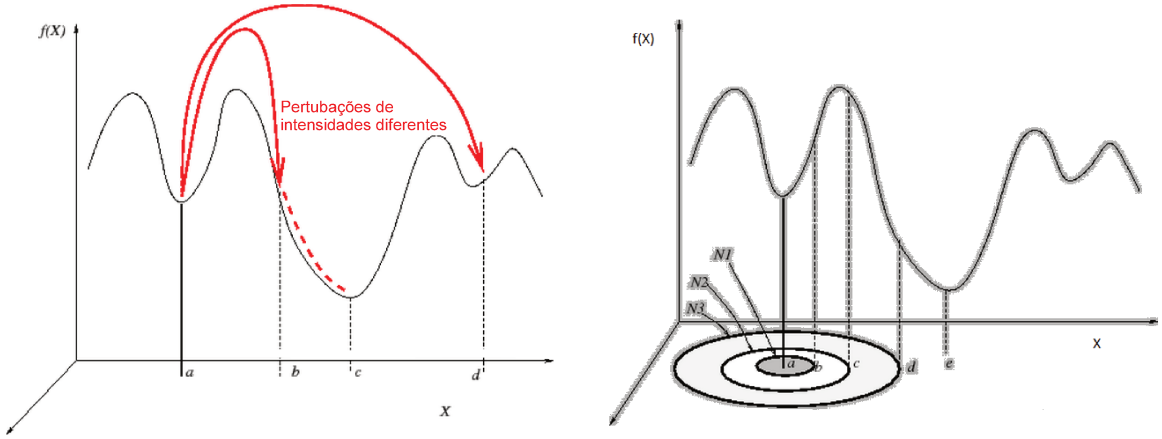


Fig. 3.3: Vizinhanças geradas a partir de pertubações incrementais permitem que se busque soluções além de um vale formado em torno de um mínimo local.
Adaptado de `lion.disi.unitn.it/reactive-search/thebook` em novembro de 2013.

Esquemas da VNS

A busca se dá por um procedimento iterativo (Algoritmo 1) que avalia soluções na vizinhança da solução corrente e substitui esta última sempre que uma melhor é encontrada. Essa busca pode ser feita de forma determinística, estocástica ou híbrida, e a escolha de como fazer a aplicação dos princípios do VNS pode definir esquemas completamente diferentes.

Algoritmo 1: esquema de descida em vizinhanças variáveis - *variable neighbourhood descent*

entrada:
- $\mathbf{x}$ uma solução inicial.
- i_{max} número máximo de vizinhanças a se investigar.

saída :
- $\mathbf{x}$ melhor solução obtida dentre as investigadas.

VND($\mathbf{x}, i_{max}$)
repeat
  $i \leftarrow 1$;
  repeat
    $\mathbf{x}' \leftarrow \underset{\bar{\mathbf{x}} \in N_i(\mathbf{x})}{\operatorname{argmin}} f(\bar{\mathbf{x}})$; // Melhor vizinho em $N_i(\mathbf{x})$
    $\mathbf{x}, i \leftarrow$ NChange($\mathbf{x}, \mathbf{x}', i$)
  until $i = i_{max}$;
until *nenhuma melhora é obtida*;

No esquema descrito no Algoritmo 1, comparamos a melhor solução da vizinhança com a solução corrente através da função NChange, dada no Algoritmo 2, que avalia se $\mathbf{x}'$ é melhor que $\mathbf{x}$. Quando $\mathbf{x}'$ melhor que $\mathbf{x}$, reinicia-se o processo centrado na nova solução $\mathbf{x}'$, caso contrário, a busca por um candidato melhor prossegue na próxima vizinhança.

Algoritmo 2: mudança de vizinhança - *neighbourhood change*

entrada:
- $\mathbf{x}$ solução na qual a vizinhança é centrada.
- $\mathbf{x}'$ solução pertencente a $N_i(\mathbf{x})$
- i parâmetro de identificação da vizinhança.

saída :
- $\mathbf{x}$ melhor solução.
- i próxima vizinhança a ser avaliada.

NChange($\mathbf{x}, \mathbf{x}', i$)
if $f(\mathbf{x}') < f(\mathbf{x})$ **then**
  $\mathbf{x} \leftarrow \mathbf{x}'$
  $i \leftarrow 1$
else
  $i \leftarrow i + 1$
end

A *VND* (Algoritmo 1) busca uma nova solução através de um procedimento determinístico: $\mathbf{x}' \leftarrow \underset{\bar{\mathbf{x}} \in N_i(\mathbf{x})}{\operatorname{argmin}} f(\bar{\mathbf{x}})$. Entretanto, para obter soluções melhores pode-se testar esquemas com abordagem estocástica, como o *Reduced VNS*, que usa uma função Shake($\mathbf{x}, i$) para escolher, de forma aleatória, um candidato $\mathbf{x}' \in N_i(\mathbf{x}) = \{\mathbf{x}^1, \ldots, \mathbf{x}^{|N_i(\mathbf{x})|}\}$. A função Shake adiciona uma pertubação aleatória, mas controlada, de $\mathbf{x}$.

O esquema básico de busca em vizinhança variável (*basic* VNS) tem sido um dos esquemas mais usados na literatura. Sua estrutura simples (ver Fig. 3.4 e Algoritmo 4) permite o uso em uma

Algoritmo 3: esquema de busca reduzida em vizinhança variável- *Reduced VNS*

entrada:
x solução inicial.
i_{max} número máximo de vizinhanças a se investigar.
saída :
x melhor solução obtida dentre as investigadas.

RVNS($\mathbf{x}, i_{max}, t_{max}$)
repeat
  $i \leftarrow 1$;
  repeat
    $\mathbf{x}' \leftarrow$ Shake($\mathbf{x}, i$);
    $\mathbf{x}, i \leftarrow$ NChange($\mathbf{x}, \mathbf{x}', i$)
  until $i = i_{max}$;
  $t \leftarrow$ tempo de processamento
until $t > t_{max}$;

grande variedade de problemas. Vale citar que a proposta de clusterização pela meta-heurística VNS, feita por *Pierre Hansen* e *Nenad Mladenović* [52] é a aplicação do *Basic* VNS com busca local por uma heurística criada por eles, chamada j-means.

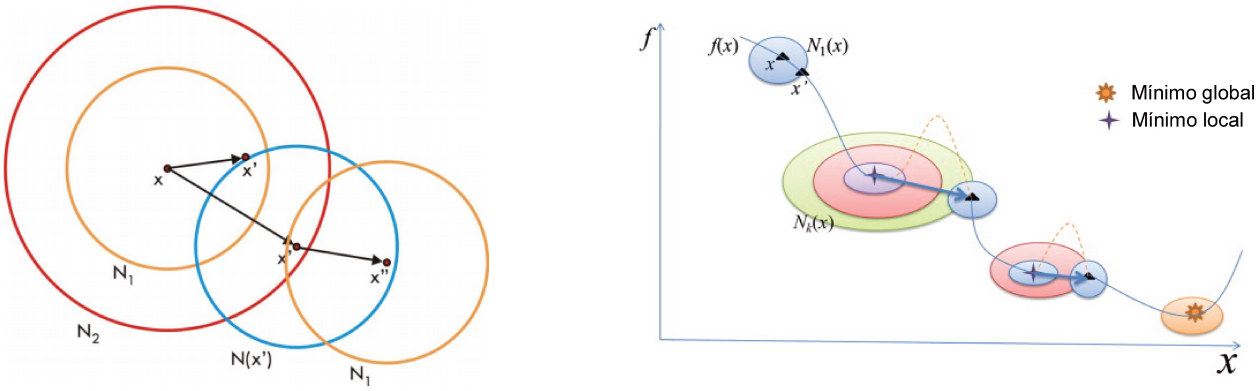


Fig. 3.4: No esquema básico, as vizinhanças são definidas de forma incremental e dentro de cada uma delas se gera uma busca local. A ideia é que as pertubações geradas pela mudança de vizinhança permitam encontrar mínimos locais melhores dentro de outros vales, ou mesmo o mínimo global.

A tendência é que, na medida em que uma meta-heurística se mostre bem sucedida, ela ganhe popularidade e cresça o número de esquemas que utilizem os seus princípios. Já existem, na atualidade, vários esquemas, além dos citados, derivados da VNS:

(i) busca em vizinhança variável genérica - *general VNS*,

(ii) busca em vizinhança variável enviesada - *skewed VNS*,

Algoritmo 4: esquema básico de busca em vizinhança variável - *Basic VNS*

entrada:
- **x** solução inicial.
- i_{max} número máximo de vizinhanças a se investigar.
- t_{max} tempo máximo de execução

saída :
- **x** melhor solução obtida dentre as investigadas.

BasicVNS($\mathbf{x}, i_{max}, t_{max}$)
repeat
  $i \leftarrow 1$;
  repeat
    $\mathbf{x}' \leftarrow$ Shake(x, i);
    $\mathbf{x}'' \leftarrow$ LocalSearch(x');
    $\mathbf{x}, i \leftarrow$ NChange$(\mathbf{x}, \mathbf{x}'', i)$
  until $i = i_{max}$;
  $t \leftarrow$ tempo de processamento
until $t > t_{max}$;

(iii) busca em vizinhança variável decomposta - *variable neighbourhood decomposition search*,

(iv) busca em vizinhança variável primal-dual - *primal-dual VNS*.

Na atualidade, a arquitetura da VNS tem se mostrado eficiente na geração de soluções factíveis para grandes problemas de programação mista e na geração de boas soluções factíveis para problemas contínuos de programação não linear. Além disso, alguns esquemas têm se destacado em campos específicos, como o caso da VNS primal-dual , que tem se mostrado bem sucedida na busca de soluções exatas para problemas de localização de grande porte [43].

3.3.2 Definindo uma estrutura de vizinhança

Para adequar as ideias do *Basic* VNS ao problema de clusterização, o conceito mais relevante a se definir é o da estrutura de vizinhança, que está intimamente relacionado à noção de pertubação. Nos algoritmos da família k-médias, o uso alternado da forma de representar os agrupamentos, ora como partição, ora como centroides, permite o uso de dois tipos de pertubações: as **pertubações sobre o conjunto de centroides** e as **pertubações por realocações**. Na primeira, aplica-se as pertubações sobre um ou mais vetores do conjunto $C = \{c_1, \cdots, c_k\}$, gerando-se uma variedade de soluções a partir de modificações controladas dos centroides. Como exemplo, uma vizinhança N_1 pode ser definida como sendo as k partições induzidas pelos centroides $C_i = \{c_1, \cdots, c_i + \delta_1, \cdots, c_k\}$ com $\delta_1 \in \mathbb{R}^d$, enquanto N_2 seria a vizinhança obtida pelas partições derivadas das $\binom{k}{2}$ variedades de centroides obtidas através da soma de δ_2 a dois centroides de C.

As pertubações por realocações são variações dos agrupamentos obtidas pela realocação de um ou mais pontos. As regras de realocação e o nível de vizinhança definem se as mudanças que serão causadas no agrupamento serão mais ou menos intensas. Como exemplo, pode-se usar a estrutura de vizinhança N_j definida pelas partições obtidas pela realocação dos j elementos mais

discrepantes dos centroides ao qual pertencem. Vale observar que a proposta de clusterização por VNS de Pierre Hansen e Nenad Mladenović [52] usa estruturas de vizinhança obtidas por pertubações por realocações.

3.3.3 Busca local

A busca local (LocalSearch) é uma instância da VNS frequentemente tratada como uma "caixa preta", ou seja, usa-se qualquer heurística já existente para o problema a ser resolvido. A única exigência é que o método seja capaz de melhorar uma solução dada. Pode-se, inclusive, usar outras meta-heurísticas como busca tabu, têmpera simulada e *VND*, ou explorar alternativas menos comuns como métodos de otimização sem derivadas.

Geralmente, quanto mais eficiente o algoritmo de busca, melhor. Entretanto, há casos em que um algoritmo simples, quando associado à VNS, tem potencial de encontrar boas soluções com baixo custo computacional. Nesta linha, entre as muitas heurísticas que resolvem o problema *MSSC* (do inglês *minimum sum-of-squares clustering*), a mais conhecida e, provavelmente, a mais utilizada, é a das k-médias [60, 66, 79]. Nessa heurística, descrita pelo Algoritmo 5, a função Inicialização($\mathbf{X}$, k) pode usar estratégias distintas para promover boas escolhas de pontos iniciais [24, 113]. As formas mais comumente usadas na literatura são:

O método ***forgy*****,** que usa a escolha aleatória de k pontos do conjunto $\mathbf{X}$ para servir como centroides iniciais;

A partição aleatória (*random partition***),** que faz uma escolha aleatória de partição para iniciar o algoritmo, ou seja, calcula os centroides depois de fazer uma atribuição aleatória dos pontos aos k grupos.

Observa-se que, para o algoritmo padrão das k-médias (*standart k-means*) e as variantes *EM*, o método *forgy* é preferível, por sua característica de espalhamento. Já a característica de centralidade da inicialização por partição aleatória é preferível para os algoritmos *k harmonic means* e *fuzzy k-means*.

Após a solução inicial ser gerada, todos os objetos são atribuídos ao centroide mais próximo. Normalmente, a métrica adotada é escolhida pelo usuário e determinada pela natureza dos dados que se pretende agrupar. Depois disso, o centroide é recalculado como a da média dos elementos associados a ele. O processo de atribuição dos objetos e o recálculo dos centroides são repetidos até que o processo convirja, como mostrado na Figura 3.5. É possível mostrar que o método das k-médias sempre converge em um número finito de iterações.

Esse procedimento iterativo, baseado em passos alternados de realocação e de recálculo, é também conhecido como algoritmo ISODATA [13, 14], algoritmo de Lloyd's [77], *hard c-means* [44] ou *h-means* [52]. Vários ajustes relativos à métrica adotada, à escolha inicial dos centroides e às formas de cálculo de centroides têm sido explorados, bem como o cálculo automático do número de grupos, o valor k. No entanto, o princípio fundamental permanece o mesmo.

Algoritmo 5: algoritmo k-médias (*k-means*)

entrada:

$\mathbf{X}$ conjunto dos pontos $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$.

k número de grupos a serem gerados.

saída :

$\mathbf{C}$ conjunto de centroides $\{\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_k\}$.

k-means(X, k)

$\mathbf{C} \leftarrow$ Inicialização$(\mathbf{X}, k)$; // Inicializa os centroides $\mathbf{C} = \{\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_k\}$

repeat

for all $i \in [n]$ **do**

$\mathbf{I}_i \leftarrow \underset{j \in [k]}{\text{argmin}} ||\mathbf{x}_i - \mathbf{c}_j||$

end for

for all $j \in [k]$ **do**

$\mathbf{c}_j \longleftarrow \frac{\sum_{i=1}^{n} \delta_{\mathbf{I}_i, j} \cdot \mathbf{x}_i}{\sum_{i=1}^{n} \delta_{\mathbf{I}_i, j}}$

end for

until convergir

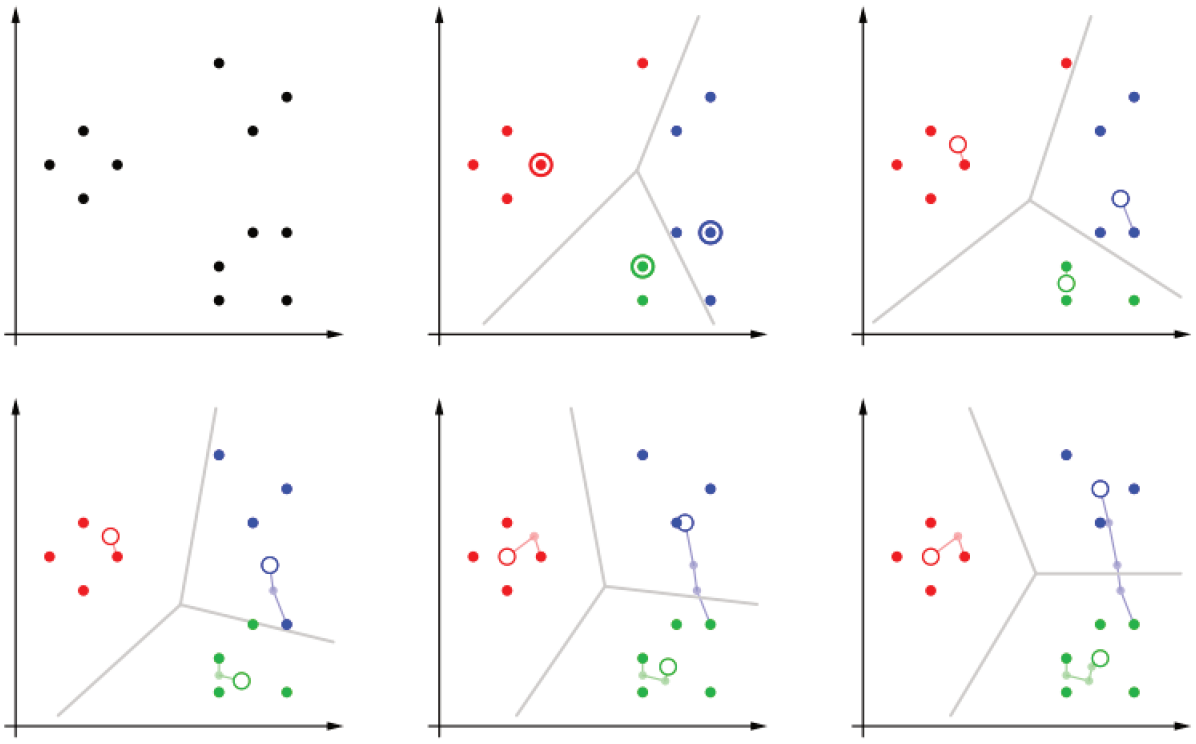


Fig. 3.5: Descreve-se, da esquerda para a direita, os estágios de agrupamento realizados pelo k-médias em um conjunto de dados bidimensionais. Inicialmente três pontos são escolhidos como centroides e depois repete-se o processo de atribuição dos objetos e o recálculo dos centroides até que haja convergência.

Capítulo 4

VNS Básico com busca local e restrições

"Ideas are the factors that lift civilization. They create revolutions.
There is more dynamite in an idea than in many bombs."
- Bishop Vincent Issues

Neste capítulo, apresentamos detalhes do método proposto para resolução do problema de agrupamento k com uso combinado da meta-heurística VNS com a busca local baseada em uma variante do k-médias que aproveita algum conhecimento prévio disponível, incorporado na forma de restrições espaciais. Por fim, apresentamos também uma sugestão de transformação de dados dinâmicos em estáticos, para permitir que se use o método proposto no agrupamento de séries temporais univariadas.

4.1 Formulação do problema de otimização

"If you have built castles in the air, your work need not be lost;
that is where they should be. Now put the foundations under them."
- Henry David Thoreau

Como dito anteriormente, a apresentação do problema de agrupamento em um formato mais adequado à programação matemática é um passo importante para a aplicação de meta-heurísticas ao problema de clusterização. Esse ajuste na representação do problema modifica a abordagem do mesmo e permite que as proposições e os teoremas inerentes à área de otimização sejam estendidos à área de agrupamento. O problema já foi formalizado (Seção 3.1), mas ainda carece de um ajuste detalhado para trabalhá-lo como um problema de otimização.

Os trabalhos seminais na interface de métodos de otimização e problemas de DM remontam a *Mangasarian*, que atacou o problema de separar duas classes através da formulação de um problema de programação linear [81]. Desde então, o interesse na interface de DM e de métodos de otimização tem crescido à medida que as técnicas de DM crescem em popularidade (e.g. [21, 25, 39, 46, 82, 106]).

Um aspecto chave da formulação do problema de agrupamento usando centroides é a minimização de uma função critério, conceito que também é um ponto central da teoria de otimização. Formular um problema de otimização significa escolher uma função objetivo a ser minimizada, ou maximizada, e definir um espaço de soluções factíveis.

Usualmente, representa-se os problemas de otimização na forma $\min\{f(x)|x \in \Omega, \Omega \subset S\}$, sendo $f : S \subset \mathbb{R}^d \to \mathbb{R}$. É comum admitir que, se S é um espaço grande, mas finito, trata-se de um problema de otimização combinatorial, enquanto $S = \mathbb{R}^d$ implica uma modelagem contínua. Para o subconjunto das soluções factíveis Ω, se $\Omega = S$, tem-se de um problema sem restrições.

Para abordar o problema de agrupamento por meio da teoria de otimização, deve-se estabelecer uma função objetivo e um domínio. A definição de uma função objetivo já é implicitamente usada em alguns métodos de agrupamentos. Entretanto, vale observar que definir a melhor forma de avaliar um *cluster* é simplesmente o tema mais controverso de toda a área de análise de agrupamentos [38].

Existe mais de uma maneira de formular o problema de agrupamento como um problema de otimização. Uma delas é a formulação inteira proposta por *Vinod* [109], em que as variáveis de decisão são indicadores de atribuição das instâncias aos grupos,

$$x_{ij} = \begin{cases} 1 & \text{se a } i\text{-ésima instância está associada ao } j\text{-ésimo grupo,} \\ 0 & \text{caso contrário,} \end{cases} \tag{4.1}$$

e o objetivo é minimizar o custo total de atribuições, onde w_{ij} é algum tipo de custo atribuído à associação entre a i-ésima instância e o j-ésimo grupo:

$$\begin{aligned} \min \ & \sum_{i=1}^{n} \sum_{j=1}^{k} w_{ij} x_{ij} \\ \text{s.a} \ & \sum_{j=1}^{k} x_{ij} = 1, \quad i = 1, 2, \ldots, n \\ & \sum_{i=1}^{n} x_{ij} \geq 1, \quad j = 1, 2, \ldots, k. \end{aligned} \tag{4.2}$$

Observa-se que as restrições propostas servem para obrigar cada *cluster* a ter ao menos um elemento e cada elemento a pertencer a exatamente um *cluster*.

Optando por uma abordagem clássica, escolheu-se a variação intracluster como função objetivo do problema de agrupamento (Definição 6). Para os fins de nossa proposta, a formulação usada é:

$$\min \ f(C) = \sum_{i=1}^{n} \min_{1 \leq j \leq k} ||x_i - c_j||^2, \tag{4.3}$$

$$\text{s.a} \ \ C \in \Omega \tag{4.4}$$

onde Ω é o conjunto de todos os grupos de k centroides que induzem uma partição de *Voronoi*, i.e.

(i) $C \in \Omega \Rightarrow (\forall \mathbf{c_j} \in C, \exists \mathbf{x} \in \mathbf{X}$ tal que $j = \underset{1 \leq l \leq k}{\operatorname{argmin}} ||x - c_l||_2^2)$.

(ii) Para quaisquer dois grupos C_i e C_j diferentes, têm-se que, se $\mathbf{x} \in C_i$ e $\mathbf{y} \in C_j$, então $\exists \mathbf{w}, \mathbf{b} \in \mathbb{R}^d$ tais que $[\mathbf{w}^t(\mathbf{x} - \mathbf{b})] \cdot [\mathbf{w}^t(\mathbf{y} - \mathbf{b})] < 0$.

A restrição (i) obriga a serem consideradas factíveis apenas as soluções C que não induzam a uma partição com $k' < k$ grupos. Uma solução C é infactível se possui algum centroide degenerado. Um centroide é considerado degenerado se não existe ao menos um elemento de $\mathbf{X}$ que esteja mais próximo a ele do que dos demais centroides. A restrição (ii) assegura que as partições formuladas sejam linearmente separáveis.

Escolhidos a função objetivo e o espaço factível, outros conceitos importantes a se definir são o de mínimo local e o mínimo global.

Definição 7 (Mínimo local)**.** *Um ponto $C' \in \Omega$ é dito mínimo local se $\exists \epsilon > 0$, de forma que $f(C') \leq f(C), \forall C \in B(C', \epsilon)$, com $B(C', \epsilon)$ sendo uma coleção de soluções de vizinhança ϵ de C'.*

Definição 8 (Mínimo global)**.** *Um ponto $C' \in \Omega$ é dito mínimo global se $f(C') \leq f(C), \forall C \in \Omega$.*

O objetivo de nosso problema é encontrar o mínimo global de f ou um mínimo local suficientemente bom. Entretanto, o problema de obter a soma mínima dos quadrados de um agrupamento é sabidamente um problema NP-difícil [27], ou seja, é um problema inerentemente complicado, independentemente do algoritmo adotado.

A função $f(C)$ (Equação 4.3) é complicada em vários sentidos. Ela é uma função com múltiplos mínimos locais, o que dificulta a obtenção do mínimo global. Além disso, apesar de ser diferenciável e contínua por partes, sob uma perspectiva de diferenciabilidade de funções com argumento matricial[1], não é possível definir os valores de mínimos locais por uma abordagem contínua, ou seja, não é possível definir para esta função um gradiente para todo o espaço de solução. Comumente, quando a função objetivo tem essas características, é usual adotar estratégias não determinísticas, como as meta-heurísticas, que fazem uso de um modelo de busca por uma solução ótima sem quase nenhum pressuposto sobre a função.

Além do mais, o domínio não é suficientemente simples a ponto de existir uma formulação matricial compacta para substituir Ω, daí o uso de regras para definir o espaço de soluções factíveis.

4.2 Estruturas de vizinhança

Propôs-se duas estruturas de vizinhanças que usam a pertubação por realocação, de forma que vizinhanças mais amplas são obtidas a partir da realocação de mais elementos. Para isso, define-se uma medida de aderência que quantifica a força com que um elemento está vinculado a um grupo e, em seguida, realoca-se os elementos com baixa aderência. Mais especificamente têm-se:

Estrutura 1: vizinhança baseada na realocação dos mais distantes Seja $(C_1 \rightarrow P_1) \bigwedge (C_2 \rightarrow P_2)$, em que C_1 e C_2 são conjuntos de centroides que induzem respectivamente as partições P_1 e P_2. Disso, definimos as funções $r(C_1, C_2)$ e $\mathbf{w}(C_1)$, sendo r o número de realocações feitas em P_1 para se obter P_2 e $\mathbf{w}$ o vetor das distâncias dos n elementos aos seus respectivos centroides C_1. Vale observar que $r : \Omega \times \Omega \rightarrow \mathbb{N}$ é positiva e simétrica, mas não necessariamente satisfaz a desigualdade triangular ou o axioma da coincidência, sendo, portanto, uma premétrica e não uma métrica.

[1] Existe mais de uma opção para definir diferenciabilidade de funções aplicadas sobre espaço de matrizes, como a derivada de Fréchet, que é definida sobre um espaço de Banach qualquer.

Se não fosse estabelecido um critério de ordenação para as coordenadas da imagem de $\mathbf{w}$, os valores de distâncias poderiam ser expressos por qualquer uma das $n!$ possíveis permutações. Portanto, por simplicidade, assumimos que a imagem de $\mathbf{w}(C_1)$ é sempre um vetor decrescente[2], de acordo com a Definição 9.

Definição 9 (vetor decrescente)**.** *Um vetor* $\mathbf{x} \in \mathbb{R}^n$ *é dito decrescente quando suas coordenadas estão arranjadas em ordem decrescente, i.e.,* $\forall\, i, j \in \{1, \cdots, n\}$*, se* $(i < j) \Rightarrow (x_i \geq x_j)$*, onde* x_j *é a* j*-ésima coordenada de* $\mathbf{x}$*.*

A partir destes conceitos, pode-se formular a estrutura de vizinhança $N^{(1)}$, tal que sua i-ésima instância, quando centrada na solução C_1, é dada por:

$$N_i^{(1)}(C_1) = \left\{ C \in \Omega \,\middle|\, (r(C_1, C) = i) \bigwedge \left(\mathbf{w}(C_1) \cdot \mathbf{e}_j \geq L, \forall\, j \in \{1, \cdots, i\} \right) \right\}, \tag{4.5}$$

onde L é um limiar mínimo de distância, que pode variar à medida que se muda i, e $\mathbf{e}_j$ é um vetor de zeros exceto pelo valor 1 na j-ésima posição.

Para os fins de nossa proposta, usou-se L como o i-ésimo maior valor de distância entre um ponto e seu centroide. Em outras palavras, a vizinhança $N_i^{(1)}$ é obtida a partir das partições geradas pelas realocações dos i elementos mais distantes de seus centroides.

Estrutura 2: vizinhanças baseadas em realocação dos mais discrepantes (*outliers*) Em lugar de considerar a distância de um elemento ao centroide, é possível considerar a distância relativa. Nesse caso, elementos mais distantes de um centroide podem ter maior coerência no contexto de seus grupos do que pontos de outros grupos que estejam mais próximos de seus centroides.

Dessa forma, pode-se definir a força de vínculo, ou aderência, de um ponto ao seu centroide como um valor dependente de todas as distâncias. A realocação de valores discrepantes (*outliers*) é uma forma de buscar soluções cujos *clûsteres* sejam homogêneos em seus próprios contextos.

Não existe uma definição rígida de valores discrepantes. Portanto, classificar um valor como *outlier* é um exercício subjetivo de definição de valores atípicos. É comum usar critérios baseados em distâncias interquatílicas e em valores normalizados ou estratégias para dados multidimensionais baseadas em distância [74] e em densidade [26].

A classificação de um valor como discrepante é comumente tratada como um problema binário. Pensando em uma visão não determinística de *outlier*, cria-se uma aplicação η sobre cada elemento $\mathbf{x}_j \in \mathbf{X}$ que quantifica o nível de discrepância deste elemento em relação ao seu grupo:

$$\eta(\mathbf{x}_j) = \frac{||\mathbf{x}_j - \mathbf{c}_l|| - \mu_l}{\sigma_l}, \tag{4.6}$$

[2]Assumindo a imagem de $\mathbf{w}(C_1)$ como um vetor decrescente, $\mathbf{w} : \Omega \to \mathbb{R}^n$ deixa de ser uma multifunção e passa a ser uma função.

onde

$$\mu_l = \frac{\sum_{\mathbf{x}\in\mathbf{C}_l} ||\mathbf{x}-\mathbf{c}_l||}{|\mathbf{C}_l|}, \qquad \text{(média das distorções do grupo } l\text{)} \qquad (4.7)$$

$$\sigma_l = \left(\frac{1}{|\mathbf{C}_l|}\sum_{\mathbf{x}\in\mathbf{C}_l}\left|||\mathbf{x}_j-\mathbf{c}_l|| - \mu_l\right|^2\right)^{\frac{1}{2}}. \qquad \text{(desvio padrão das distorções do grupo } l\text{)} \qquad (4.8)$$

A ideia é que a realocação leve em conta o quão discrepante um ponto é em relação ao seu grupo. Dessa forma, a medida de aderência de $\mathbf{x}_j$ é inversamente proporcional a $\eta(\mathbf{x}_j)$.

Definindo-se a imagem $\Theta(C_1)$ como o vetor decrescente dos valores η dos elementos de C_1, a i-ésima vizinhança de $N^{(2)}$, centrada na solução C_1, é dada por:

$$N_i^{(2)}(C_1) = \left\{C \in \Omega \middle| (r(C_1,C) = i) \bigwedge \left(\Theta(C_1)\cdot\mathbf{e}_j \geq L, \forall\, j \in \{1,\cdots,i\}\right)\right\}, \qquad (4.9)$$

onde L é o i-ésimo maior valor de μ entre os elementos da solução C_1.

Em outras palavras, a vizinhança $N_i^{(2)}$ é definida pelos centroides obtidos nas realocações dos i elementos mais discrepantes no contexto de seus respectivos *clûsteres*.

As propostas de vizinhança $N^{(1)}$ e $N^{(2)}$ vão gradativamente aumentando o nível de pertubação ao realocar um número maior de elementos. Entretanto, muitas vezes, são necessárias estruturas que explorem pertubações mais vigorosas, que usem soluções mais distantes da solução corrente. Para isso, pode-se usar diferentes estratégias, como a escolha de pontos aleatoriamente distantes, embora essa escolha possa não ser uma boa opção, pois se assemelha muito ao procedimento de reinicialização por recomeços. Como opção, pode-se usar esquemas diferentes do básico, como o *skewed* VNS, ou criar uma estrutura de vizinhanças que favoreça um maior número de realocações.

Para qualquer estrutura entre as duas vizinhanças propostas, a função Shake de alguns esquemas da VNS incorpora o aspecto estocástico ao algoritmo quando realiza, de forma aleatória, a escolha dos *clûsteres* que receberão os pontos a serem realocados.

4.3 Incorporando conhecimento

Incorporar conhecimento, mesmo em pequena quantidade, pode beneficiar a qualidade dos grupos formados [15, 18, 111, 116]. Desse princípio, surgiu a clusterização semissupervisionada, também conhecida como clusterização com restrições (*constrained clustering*). A clusterização com restrições se propõe a melhorar o desempenho alcançado pela clusterização não supervisionada, adicionando conhecimento prévio de uma pequena porção de dados. Experimentos têm mostrado que as variantes semissupervisionadas do *k-means*, como o *Seeded k-means*, o *Constrained k-means* e o *COP k-means*, superam o algoritmo tradicional.

Como visto na Seção 2.2.3, as formas como se incorpora conhecimento prévio no processo de agrupamento e de classificação são diferentes. A classificação usa uma indução lógica do grupo

de treinamento para definir uma fronteira espacial entre as classes, enquanto a clusterização semissupervisionada usa o conhecimento prévio para guiar o processo de agrupamento pela inclusão de restrições no espaço de busca. Além disso, a diferença entre as duas se dá pela quantidade de conhecimento *a priori* que se usa. Na classificação, o conjunto de treinamento deve conter informações sobre todas as classes e em volume suficiente para caracterizá-las, enquanto, na clusterização, as informações podem ser deficientes, descrevendo apenas algumas relações ou classes em uma quantidade restrita.

A forma das restrições pode variar dependendo da natureza do conhecimento e da estratégia para incorporá-lo, e este conhecimento *a priori* normalmente se apresenta em uma das seguintes formas:

Relações de paridade, que são um conjunto de restrições do tipo paridade obrigatória (*must-link*) e disparidade obrigatória (*cannot-link*). A paridade obrigatória é usada para especificar que duas instâncias devem necessariamente ser associadas ao mesmo grupo, enquanto a disparidade obrigatória é usada para especificar que as duas instâncias não devem ser associadas ao mesmo *cluster*.

Amostra de classes, que, de forma semelhante a relações de paridade, serve para definir quais elementos pertencem à mesma classe e quais elementos não devem ser postos no mesmo grupo. Sua diferença com relação às relações de paridade dá-se ao definir as relações de vínculo, ou de desvínculo, para grupos de elementos e não para pares. E, ao contrário das amostras usadas em classificação, nem todas as classes são representadas.

Pontos de inicialização, é um conjunto de pontos iniciais (*Seeds*). Um conjunto de relações de paridade, ou amostras, pode ser convertido em um conjunto de *seeds*. A hipótese é que estes pontos iniciais estão próximos aos centroides reais das classes, fazendo deles uma informação privilegiada para inicialização do método de agrupamento.

Para os fins propostos, a clusterização clássica, na perspectiva de classificação não supervisionada, pode ser entendida como um caso degenerado de clusterização semissupervisionada. Ou seja, a clusterização usa informações na medida de sua disponibilidade, na esperança de que, quanto maior a quantidade de informação, mais confiáveis os resultados obtidos serão.

Quando se acentua ou se incorpora restrições novas, ocorre uma redução do espaço de soluções factíveis. Admite-se que essa redução não afeta, ao menos de forma significativa, o valor da função objetivo. Mas, mesmo quando há uma perda nominal no valor da função objetivo ou de outro estimador de qualidade, essa penalização ocorre para ajustar o agrupamento ao conhecimento prévio incorporado. Assim, deve-se ter cautela ao afirmar que um agrupamento com menor valor de função objetivo é melhor do que um agrupamento com um valor maior. Isso porque esse aumento pode ser decorrente da incorporação das restrições e, por mais que a função objetivo seja o guia de qualidade do agrupamento, essa qualidade só pode ser realmente atestada pela perspectiva subjetiva de um especialista no domínio das informações.

4.3.1 Restrições de caixas

A incorporação do conhecimento ao algoritmo depende da forma como se transforma a informação *a priori* em restrições. O algoritmo *Seeded k-means* usa dados rotulados apenas para

inicializar o algoritmo. Já a inicialização do algoritmo *COP k-means* é feita de forma aleatória, mas durante seu processo de rerotulação, ele nunca desfaz uma restrição *must-link* ou força a pertencer ao mesmo grupo elementos que tenham uma restrição *cannot-link*.

Grande parte dos algoritmos semissupervisionados presentes na literatura força a associação de instâncias uma a outra e termina o processo quando não há agrupamentos que satisfaçam essas restrições. Pode-se também relaxar o conceito de infactibilidade, aceitando soluções que minimizam a quantidade de violações.

Uma desvantagem do processo de incorpoção das restrições é que ele pode não ser efetivo quando os *seeds* contêm ruído ou estão enviesados. Os algoritmos que usam *seeds* para inicialização são menos sensíveis a ruídos, mas se restringem a usar o conhecimento prévio para inicializar o algoritmo.

Sobre a hipótese de que é mais importante trabalhar as características decorrentes de restrições espaciais do que de restrições de vínculo, propôs-se o conceito de restrições por caixas.

Definição 10 (Restrições de caixas)**.** *As restrições por caixas confinam os centroides às regiões de confiança, caracterizadas por uma caixa. Esta caixa $\mathscr{H}_i$ pode ser formalmente expressa como:*

$$\mathscr{H}_i = \left\{ \mathbf{x} \in \mathbb{R}^d \middle| a_j \leq x_j \leq b_j \right\}, \forall\, j \in \{1, \cdots, d\}, \tag{4.10}$$

sendo a_j e b_j valores derivados da amostra de classes ou das restrições must-link.

O uso de restrições de caixas é uma abordagem inédita e seu objetivo é transformar as restrições de instâncias em restrições espaciais. Usa-se amostras de classes ou restrições de paridade (*must-link*) para gerar uma caixa que limite o movimento dos centroides. A ideia é que essas restrições funcionem como um guia para o algoritmo de agrupamento a fim de encontrar agrupamentos melhores.

4.4 Método proposto

Quando uma estratégia supervisionada é impraticável, pode-se usar estratégias não supervisionadas ou semissupervisionadas com o uso de algoritmos de busca local que são comumente baratos computacionalmente e de bom desempenho. Entretanto, estes algoritmos são frequentemente dependentes dos pontos iniciais, o que pode fazê-los estancar em mínimos locais ruins, mesmo quando adotadas estratégias de recomeços. Neste contexto, propôs-se o Algoritmo 6, um esquema geral que tenta reduzir a sensibilidade do algoritmo aos pontos iniciais, através da incorporação do conhecimento prévio e do uso combinado da busca local com a VNS.

Função Shake

A função Shake serve para ampliar a diversidade da exploração, ao escolher a solução que será usada para inicializar a busca local (LocalSearch). Seus propósito é escolher, de forma aleatória, uma solução $\mathbf{C}'$ da i-ésima vizinhança da solução corrente $\mathbf{C}$.

O valor de i_{max}, nível máximo de vizinhanças, pode gerar o inconveniente da degeneração caso seja maior ou igual ao número de elementos do menor grupo, dada a possibilidade de que esse grupo seja completamente esvaziado. Para contornar esse problema , uma regra heurística

Algoritmo 6: esquema geral do algoritmo proposto

entrada:

$\mathbf{X}$	conjunto dos pontos $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ a serem agrupados
$\mathbf{C}$	conjunto de centroides iniciais $\{\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_k\}$
Ω_c	conjunto de restrições de caixa $\{\mathscr{H}_1, \mathscr{H}_2, \cdots, \mathscr{H}_{k'}\}$
k	número de grupos a serem gerados
t_{max}	tempo máximo de execução

saída :

$\mathbf{C}$	melhor solução (conjunto de centroides) obtida dentre as investigadas.

```
BasicVNS(X, C, Ω_c, k, t_max)
i_max ⟵ n/10
repeat
    i ← 1 ;
    repeat
        C' ← Shake(C, i);
        C'' ← LocalSearch(X, k, C', Ω_c);
        C, i ← NChange(C, C'', i)
    until i = i_max;
    t ← tempo de processamento
until t > t_max;
```

que tem se mostrado consistente é assumir que a pertubação por realocação seja limitada definindo $i_{max} = \frac{n}{2k}$ ou $i_{max} = \frac{n}{10}$, i.e., o número de elementos a serem realocados não deve ultrapassar 50% do número médio de elementos por cluster ou 10% do número total de elementos de X.

Função LocalSearch

Para algoritmos de busca local, é possível usar informações sobre os grupos. A nossa proposta leva em conta restrições de caixa definidas da seguinte maneira:

Definição 11 (Caixa Ω_c)**.** *Sejam μ_j e σ_j, respectivamente, a média e o desvio padrão, na dimensão j, da amostra $A_i = \{\mathbf{x}_1^{(i)}, \cdots, \mathbf{x}_{n_i}^{(i)}\}$ de uma classe i. Disso tem-se*

$$\mathscr{H}_i = \left\{ \mathbf{x} \in \mathbb{R}^d \middle| \mu_j - \frac{3\sigma_j}{\sqrt{n_i}} \leq x_j \leq \mu_j - \frac{3\sigma_j}{\sqrt{n_i}}, \forall\, j \in \{1, \cdots, d\} \right\}. \tag{4.11}$$

Onde $\Omega_c = \{\mathscr{H}_1, \cdots, \mathscr{H}_{k'}\}$, sendo k' o número de restrições definido pelo número de classes distintas que tenham uma amostra.

Sabe-se que um centroide não pode estar fora de certos limites definidos por uma amostra de seu grupo. Decorre então que o princípio das restrições por caixas é o de usar o conhecimento prévio de uma amostra da classe para definir uma porção do espaço onde se confia que o centroide esteja.

Diremos que um centroide $\mathbf{c}_i$ é factível se satisfaz as restrições de caixa $\mathscr{H}_i$. Caso seja infactível, deve-se trocar $\mathbf{c}_i$ por suas projeção em $\mathscr{H}_i$. Para isso definimos uma projeção da seguinte maneira:

Definição 12 (Projeção do centroide sobre a caixa)**.** *Seja* $\mathbf{c}_{new} \in C'$ *a atualização infactível de* $\mathbf{c}_{old} \in C$ *e* $\mathscr{H}_i = \{\mathbf{x} \in \mathbb{R}^d | a_j \leq x_j \leq b_j, \forall\, j \in \{1, \cdot, d\}\}$ *a restrição de caixa que contém* $\mathbf{c}_{old}$*. Então, a projeção de* $\mathbf{c}_{new}$ *sobre a caixa* $\mathscr{H}_i$ *é dada por* $\mathbf{c}^* = \mathbf{c}_{old} + \alpha\mathbf{d}$ *com* $\mathbf{d} = \mathbf{c}_{new} - \mathbf{c}_{old}$ *e*

$$\alpha = \min\left\{1, \min_{j|d_j>0}\left\{\frac{b_j - c_j}{d_j}\right\}, \min_{j|d_j<0}\left\{\frac{a_j - c_j}{d_j}\right\}\right\},$$

Sendo a função de projeção definida pelo Algoritmo 7.

Algoritmo 7: projeção sobre a caixa

$C' \leftarrow$ Projection(C', C, Ω_c)

passo 1 (verificação): se todas as caixas de Ω_c possuem ao menos um centroide, parar

passo 2 (seleção): selecionar os centroides de C' que violam as restrições

passo 3 (projeção): calcular a projeção do centroides selecionados de acordo com a regra de projeção em caixa (ver definição 12)

passo 4 (atualização): trocar, em C', os centroides selecionados pelos centroides projetados

Por fim, usando os conceitos de restrições por caixa e projeção propomos a função de busca local como descrito no Algoritmo 8.

Algoritmo 8: *projected k-means*: Variante k-médias com restrição de caixa

Descrição por passos da busca local com restrições por caixas

$\mathbf{C} \leftarrow$ LocalSearch$(\mathbf{X}, k, \mathbf{C}, \Omega_c)$

passo 1 (inicialização): usar C como solução inicial

passo 2 (atribuição): associar cada elemento x_j, com $j \in \{1, \cdots, n\}$, ao centroide mais próximo ($C \rightarrow P$)

passo 3 (teste de otimalidade local): se não houve alguma mudança no passo anterior, parar aqui

passo 4 (projeção): $C' \longleftarrow$ Projection(C', C, Ω_c)

passo 5 (atualização): trocar os centroides C da solução corrente por C' e voltar para o **passo 2**

Prevenindo degeneração na busca local

O algoritmo de busca local proposto, assim como várias variantes do k-médias padrão, pode convergir para soluções degeneradas [104, p. 68]. Ou seja, ele pode convergir para soluções onde

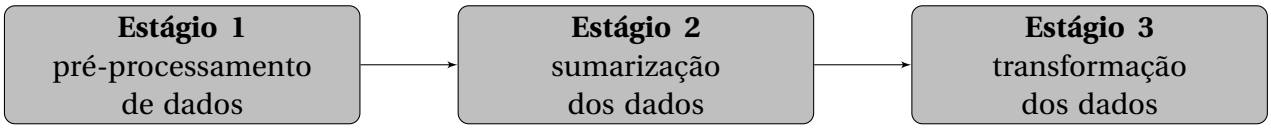
um ou mais conjuntos são vazios, gerando um particionamento final com uma quantidade de grupos menor que k. Entretanto, uma solução final degenerada pode ser facilmente corrigida por uma estratégia de inclusão, como a heurística de permutação de *Cooper* [104]. Dada uma solução degenerada, então $\exists k_d > 0$, tal que o número de *clûsteres* da solução corrente é $k - k_d$. Daí, transforma-se cada um dos k_d pontos mais distantes de seus respectivos centroides, e portanto os que mais impactam na função objetivo, em k_d grupos de um único elemento. É fácil demonstrar que a nova solução é melhor que a anterior, apesar de poder ser melhorada. Por isso, para estas soluções, o processo de busca local recomeça a partir da solução modificada. Essa variante da busca local, que previne degenerações, pode ser bem definida com a substituição da instrução do **teste de otimalidade local** do Algoritmo 8 pela seguinte:

passo 3 (teste de otimalidade local)
Se não houve alguma mudança no passo anterior, então
Se a solução corrente não é degenerada, parar,
caso contrário, selecionar k_d pontos como centroides e voltar ao **passo 2**

4.5 Representação de séries temporais

Além do domínio usual de dados estáticos, pretende-se usar o método proposto em uma aplicação mais específica: a classificação de séries temporais de imagens de satélite. Entretanto, isso é inviável sem um módulo de preparação dos dados. Mais que o simples ajustamento dos dados ao método proposto, objetiva-se com isso transformar a classificação de imagens de satélite, problema de domínio específico, em um problema universal. A ideia é que esses dados, após as etapas de ajustamento, sejam tratáveis por qualquer método ou algoritmo clássico de clusterização.

Uma série temporal de imagens de satélite pode ser entendida, de forma simplificada, como uma coleção de pixels, em que a cada pixel está associada uma série temporal das refletâncias de faixas específicas do espectro eletromagnético de uma dada região de interesse. Disso decorre que uma sequência multitemporal de imagens de satélite é uma coleção de séries temporais multivariadas. Propôs-se três etapas de ajustamento dos dados para se usar o método proposto neste tipo específico de objeto:



As fases 1 e 2 são etapas de preparação dos dados. Na fase 1, faz-se o georreferenciamento, a correção radiométrica, a correção atmosférica e a correção geométrica das imagens. Na fase 2, converte-se a série multivariada em univariada, transformando os valores de refletância, de duas ou mais bandas, em um índice sumarizante que tenha alta correlação com a biomassa vegetal das culturas agrícolas a serem identificadas e que mitigue os ruídos decorrentes do processo de captação e das condições metereológicas e climáticas.

Para essas fases, empregou-se uma solução já presente na literatura [10, 36], usando o sistema *NAVPRO*, que processa os dados brutos e os converte em composições de máximo valor (*mvc*) de *NDVI* (ver Seção 2.1.3). O sistema NAVPRO [8] foi criado pela Embrapa Informática Agropecuária em parceria com a Universidade Estadual de Campinas (Unicamp) e contou com o pacote computacional NAV (*NAVigation*), desenvolvido pelo *Colorado Center for Astrodynamics Research (CCAR)*, da Universidade do Colorado, EUA.

4.5.1 Transformação dos dados

Os algoritmos de clusterização, em geral, dependem implicitamente da imposição de certas hipóteses a respeito da forma dos *clûsteres* ou da configuração dos múltiplos *clûsteres*. Os dados dificilmente estão estruturados de forma ideal, ou seja, não formam configurações hiperesféricas, hiperelipsoidais, lineares, etc., de modo que cada novo algoritmo de clusterização pode apresentar um comportamento superior aos já existentes para uma dada conformação específica dos dados.

Dessa dificuldade, usa-se a fase de transformação (Estágio 3), na qual a análise de dados funcionais (*Functional Data Analysis*) é empregada para explorar a característica funcional dos dados. Aqui, o termo funcional se refere à estrutura dos dados e não à sua forma explícita, pois, na prática, os dados são observados de maneira discreta.

Abordagens funcionais têm sido observadas com frequência cada vez maior em diversos campos. Isso se justifica porque, em muitos casos, o interesse está na estimação não somente das curvas, mas também de outros funcionais, como derivadas e integrais destas curvas. Por exemplo, no problema de crescimento de crianças, pode-se estar interessado não somente em estimar a curva de crescimento, mas simultaneamente em estimar a velocidade de crescimento ou a aceleração como função do tempo para cada indivíduo.

De forma mais específica, o objetivo é converter um conjunto de n séries discretas, sendo cada uma expressa pelas medidas $y_{i1}, \cdots, y_{id}$, em uma função x_i com valores $x_i(t)$ para todos os valores de t. Se é admitido que essas observações não contêm erros, esse processo é conhecido como interpolação (*interpolation*). Caso contrário, se as medições contêm algum erro observacional, então a conversão deste conjunto finito para uma função pode envolver uma suavização (*smoothing*).

Para os dados agrícolas longitudinais supomos que suas principais informações são derivadas de sua curva. Por isso propôs-se a transformação $\mathscr{F}$, que substitui as séries originais por vetores de coeficientes obtidos pela transformada de Fourier.

O descritor $< \mathscr{F}, L_2 >$ caracteriza a oscilação dos valores $x(t)$ usando a métrica usual de espaço Euclidiano. A ideia por trás deste descritor não é verdadeiramente nova, tendo equivalentes teóricos em abordagens como a *HANTS* (*Harmonic ANalysis of Time Series*) [119].

A transformação $\mathscr{F}$ decompõe um sinal em um número infinito de componentes (harmônicos), em que cada componente é formado por ondas senoidais e cossenoidais de mesma frequência. Descrevemos uma abordagem contínua para, posteriormente, usar as ideias desenvolvidas para formular a teoria de transformação para sequências discretas.

Seja $f:[0,L]\to\mathbb{R}$ contínua, então:

$$f(x)=\frac{1}{2}a_0+\sum_{n=1}^{\infty}\left(a_n\cos\frac{2\pi nx}{L}+b_n\,\mathrm{sen}\frac{2\pi nx}{L}\right). \tag{4.12}$$

O lado direito da Equação 4.12 é a representação da função $f(x)$ pela série de Fourier. Os coeficientes de Fourier podem ser obtidos pela manipulação (multiplicando por seno ou cosseno e integrando) da expressão acima, de modo a obtermos a_n e b_n na forma:

$$a_n=\frac{2}{L}\int_0^L f(x)\cos\frac{2\pi nx}{L}dx \quad \text{e} \quad b_n=\frac{2}{L}\int_0^L f(x)\,\mathrm{sen}\frac{2\pi nx}{L}dx, \quad \text{com } n\geq 1. \tag{4.13}$$

Note que $\frac{1}{2}a_0=\frac{1}{L}\int_0^L f(x)dx$ é o valor médio de $f(x)$. Da forma como os coeficientes foram definidos acima, a série de Fourier é única. Definimos o j-ésimo harmônico como sendo o j-ésimo termo da série de Fourier (para $j\geq 1$), dado por

$$a_j\cos\frac{2\pi jx}{L}+b_j\,\mathrm{sen}\frac{2\pi jx}{L}.$$

Convertemos o j-ésimo termo em um único termo de cosseno da seguinte forma:

$$\begin{aligned}a_j\cos\frac{2\pi jx}{L}+b_j\,\mathrm{sen}\frac{2\pi jx}{L} &= \sqrt{a_j^2+b_j^2}\left(\frac{a_j}{\sqrt{a_j^2+b_j^2}}\cos\frac{2\pi jx}{L}+\frac{b_j}{\sqrt{a_j^2+b_j^2}}\,\mathrm{sen}\frac{2\pi jx}{L}\right)\\ &= \sqrt{a_j^2+b_j^2}\left(\cos\phi_j\cos\frac{2\pi jx}{L}+\,\mathrm{sen}\phi_j\,\mathrm{sen}\frac{2\pi jx}{L}\right)=c_j\cos\left(\frac{2\pi jx}{L}-\phi_j\right),\end{aligned}$$

onde $c_j=\sqrt{a_j^2+b_j^2}$ e geralmente se define $\phi=tg^{-1}\left(\frac{b_j}{a_j}\right)$. Como essa definição de ϕ produz valores no intervalo $[-\frac{\pi}{2},\frac{\pi}{2}]$, adotamos a forma modificada $\phi=tg^{-1}\left(\frac{b_j}{a_j}\right)+\pi$, sempre que $a_j<0$. Disso decorre que $\phi\in[-\frac{\pi}{2},\frac{3\pi}{2}]$. Usando $c_0=\frac{1}{2}a_0$, temos:

$$f(x)=c_0+\sum_{n=1}^{\infty}c_n\cos\left(\frac{2\pi nx}{L}-\phi_n\right),$$

onde c_n é a amplitude e ϕ_n é o ângulo de fase do n-ésimo termo (n-ésimo harmônico).

Para um conjunto de dados finito $y(k)$, com $k\in\{1,2,3,\cdots,n\}$, usamos as ideias anteriores para desenvolver $\mathscr{F}$ através de uma técnica finita em que se substitui as integrais de Riemann, em (4.13), por aproximações trapezoidais. Qualquer série de n pontos pode ser representada exatamente pela expressão:

$$y_t = \bar{y}+\sum_{k=1}^{\lfloor\frac{n}{2}\rfloor}\left[C_k\cos\left(\frac{2\pi kt}{n}-\Phi_k\right)\right]= \tag{4.14}$$

$$= \bar{y}+\sum_{k=1}^{\lfloor\frac{n}{2}\rfloor}\left[A_k\cos\left(\frac{2\pi kt}{n}\right)+B_k\,\mathrm{sen}\left(\frac{2\pi kt}{n}\right)\right], \tag{4.15}$$

onde $\bar{y}$ é a média aritmética dos dados. Os termos A_k e B_k, para uma série de dados igualmente espaçados no tempo e sem valores faltantes, assumem a forma:

$$A_k = \frac{2}{n}\sum_{t=1}^{n} y_t \cos\left(\frac{2\pi k t}{n}\right) \qquad \text{e} \qquad B_k = \frac{2}{n}\sum_{t=1}^{n} y_t \operatorname{sen}\left(\frac{2\pi k t}{n}\right). \tag{4.16}$$

Disso decorre que $\mathscr{F}(\mathbf{y}_t)$ substitui os pares (k, y_k) por C_k, Φ_k, onde $C_k = \sqrt{A_k^2 + B_k^2}$ e $\Phi_k = tg^{-1}\left(\frac{B_j}{A_j}\right)$.

Vale atentar que para se aproximar da série, pode-se utilizar qualquer base, não sendo obrigatório o uso da base de senos e cossenos.

Capítulo 5

Experimentos Computacionais

"It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts."
- Sherlock Holmes - (Sir Arthur Conan Doyle)

Neste capítulo, é feita a comparação de desempenho entre os algoritmos propostos na tese e métodos mais tradicionais, como o k-médias e a sua variante semissupervisionada mais conhecida, o *COP* k-médias. A comparação é feita pela análise dos valores das métricas *b-cubed* geradas pelos algoritmos quando aplicados a conjunto de dados icônicos, i.e. quando usados para agrupar bases de dados com as características mais frequentes em dados de testes e dados agrícolas. O objetivo é usar as estatísticas de desempenho para inferir o comportamento dos algoritmos em bases que sejam similares às testadas. A hipótese norteadora dessa forma empírica de avaliar algoritmos é que as estatísticas de desempenho são significativas para grupos com distribuição e dimensão semelhantes. Essa maneira de testar algoritmos não é exatamente nova. Além de existirem diversos artigos que usam o desempenho em testes[1] para sugerir a força de seus métodos, existe uma tendência crescente nas publicações das ciências formais de se oferecer reprodutibilidade empírica para os resultados de experimentos computacionais. Um indicativo disso é o uso crescente de avaliações de *solvers* através de gráficos de perfil de desempenho (*performance profile*) em bases de problemas como o CUTEr[2].

Todos os resultados e análises apresentados podem ser reproduzidos a partir da versão virtual da tese, um repositório *on-line* com todos os programas desenvolvidos (*MatLab*, *R* e python) e arquivos de dados planos (*.csv*), disponível em `wandersonluiz.github.io/vnsconstrainedkmeans/`. Todos os códigos fontes estão disponíveis sobre a licença *Gnu LeSSEr General Public Licence*[3].

[1] Como exemplo pode-se citar o repositório de problemas para aprendizado de máquina UCI em `http://archive.ics.uci.edu/ml/`, visitado em 21/09/2015.

[2] `http://www.cuter.rl.ac.uk/`, visitado em 01/01/2015.

[3] `gnu.org/licenses/lgpl`

5.1 Metodologia

As análises comparativas consideraram a qualidade das soluções sem se preocupar com o tempo consumido para alcançá-las. Isso foi feito porque o tempo computacional foi usado como um critério de parada que assegura que todos os algoritmos avaliados tenham seu tempo de execução limitado a sessenta segundos. Este valor de tempo foi considerado um limite razoável pois tempos menores não são considerados melhores a não ser que tragam soluções melhores. Acrescido a isso, o tempo computacional de execução dos algoritmos testados não tem distinção estatística, pois a intersecção do intervalo de confiança das médias de tempo, de cada um dos métodos, é não nula. Disso, temos que as comparações de desempenho foram feitas a partir das métricas: precisão (**B3P**), lembrança (**B3R**), F de *Rijsbergen* (**B3F**) e o total de variação intracluster (**SSE**). As métricas B3 (ver Seção 2.3.2) consistem em estimativas do quanto as relações de paridade presentes no agrupamento de fato correspondem a relações de paridade existentes no padrão ouro. Dessa forma, cada escore serve para descrever o comportamento médio de associações verdadeiras no grupo gerado e quanto das associações originais são preservadas.

Algoritmos testados

- **KM:** Método k-médias com múltiplos reinícios (*multi-start k-means*), em que se usa o agrupamento com melhor resultado de SSE das 10 rodadas do k-médias, sendo os pontos iniciais de cada rodada gerados pelo padrão de inicialização *forgy*.
- **PKM-SS:** k-médias por projeção (*projected k-means*) com informações iniciais restritas a a alguns exemplares de algumas classes (*Supervised Seed*). É essencialmente o algoritmo PKM, sem o controle de uma barreira de factibilidade.
- **PKM:** k-médias por projeção com informação inicial suficiente para gerar uma região de confiança que restrinja os centroides (ver Algoritmo 8).
- **CKM:** COP k-médias (ver Seção 2.2.3).
- **VNS-KM-N1:** VNS k-médias com estruturas de vizinhança do tipo 1 (ver Seção 4.2).
- **VNS-KM-N2:** VNS k-médias com estruturas de vizinhança do tipo 2 (ver Seção 4.2).
- **VNS-PKM-N1:** VNS Projetado com k-médias em estruturas de vizinhança do tipo 1.
- **VNS-PKM-N2:** VNS Projetado com k-médias em estruturas de vizinhança do tipo 2.
- **VNS-CKM-N1:** VNS COP k-médias com estruturas de vizinhança do tipo 1.
- **VNS-CKM-N2:** VNS COP k-médias com estruturas de vizinhança do tipo 2.

Calibração dos algoritmos

Os métodos de busca, quando necessário, usaram os parâmetros de interrupção $t = 60s$ e $i_{max} = 10\%|X|$, sendo $|X|$ o número de elementos de X. O número $i_{max} = 10\%|X|$ assegura que as variedades de solução obtidas pelas realocações dos elementos mais discrepantes, nas variantes VNS, nunca ocorram pela realocação de mais de um décimo do tamanho da base.

As variantes que fazem uso de restrições tiveram as restrições induzidas a partir de amostras aleatórias S, sendo $S = \bigcup_{i=1}^{k'} S_i$, tal que e $|S_i| \leq 20\%|G_i|$. Ou seja, as restrições, quando usadas, foram induzidas por amostras de tamanho não superior a 20% do tamanho total do grupo por classe G_i. Para o algoritmo PKM-SS, em que se fez uso de uma quantidade de dados pequena ($|S_i| \leq 1$), o método usa o representante da classe como ponto inicial (*seed*), sem geração de caixa.

A escolha dos valores acima tem o objetivo de produzir uma variação do método de **validação cruzada**, que usa parte do conjunto para induzir um critério de separação e testa este critério no restante do conjunto. O tamanho das amostras utilizadas para gerar as restrições é pequeno quando comparado ao tamanho de amostras necessárias em algoritmos de classificação supervisionada. O propósito é simular situações em que existe disponibilidade de dados, mas em quantidade insuficiente para o paradigma supervisionado de classificação.

Análise estatística

Como o desempenho dos algoritmos é dependente de fatores aleatórios decorrente de suas condições iniciais (coleção de pontos conhecidos previamente e pontos de inicialização), estabeleceu-se um conjunto destas condições para avaliar o desempenho dos algoritmos. Para cada base de dados estáticos, formulou-se 100 conjuntos de pontos iniciais e conjuntos de pontos cuja a classe se conhece a priori. Para os dados dinâmicos estabeleceu-se 20 destas condições iniciais. As cinco bases testadas forneceram uma variedade de 340 condições iniciais e, para cada uma delas, os algoritmos fornecem quatro valores de métricas de desempenho. Esses valores são analisados sob uma perspectiva tabular e gráfica.

A análise tabular conta com uma tabela de médias e desvio padrão, como a Tabela 5.1, e outra de ranking dos algoritmos, como a Tabela 5.2. A primeira dispõe a medida de tendência central e dispersão de cada algoritmo, em cada métrica, nas linhas da tabela. A segunda auxilia na interpretação da primeira tabela, estabelecendo uma hierarquia entre os resultados apresentados. Em ambas as tabelas, enfatiza-se a coluna com o escore mais significativo para a base de dados analisada. Isso decorre do fato de que, embora a métrica F seja comumente considerada o valor universal para estabelecer comparações, existem bases de dados que distorcem a lembrança e, consequentemente, a métrica F. Para estes casos, da-se ênfase à precisão.

As médias dos escores são estimadores pontuais do desempenho real do algoritmo. Como existe uma certa variabilidade no desempenho dos algoritmos, considerou-se para análise comparativa, além das tabelas de médias, os gráficos de intervalos de confiança das médias e os gráficos de perfil de desempenho.

Os gráficos de intervalo de confiança foram gerados com o propósito de criar uma análise intervalar de desempenho. Para tanto, avaliou-se a normalidade dos escores, ou seja, se os dados apresentados vêm de uma distribuição normal. Como muitos procedimentos estatísticos fazem a suposição de que uma distribuição subjacente é normal, testes de normalidade podem fornecer alguma garantia de que a suposição é justificada ou indicar que essa suposição pode ser falsa. Para as métricas avaliadas fica claro que uma quantidade significativa das suas distribuições não é normal (ver Fig. 5.1).

Frente à não normalidade dos dados, usou-se uma estratégia de reamostragem chamada *bias-corrected and accelerated interval* (*BCa interval*) para inferir o intervalo de confiança. A reamostragem consiste em amostrar n observações, com reposição dos dados observados; calcular a média do conjunto simulado e repetir este teste B vezes, obtendo B médias de simulação. As médias obtidas fazem um desenho aproximado da distribuição amostral de tamanho n, a partir do qual podemos construir o histograma das reamostragens, calcular o desvio padrão da média e estimar quartis da média.

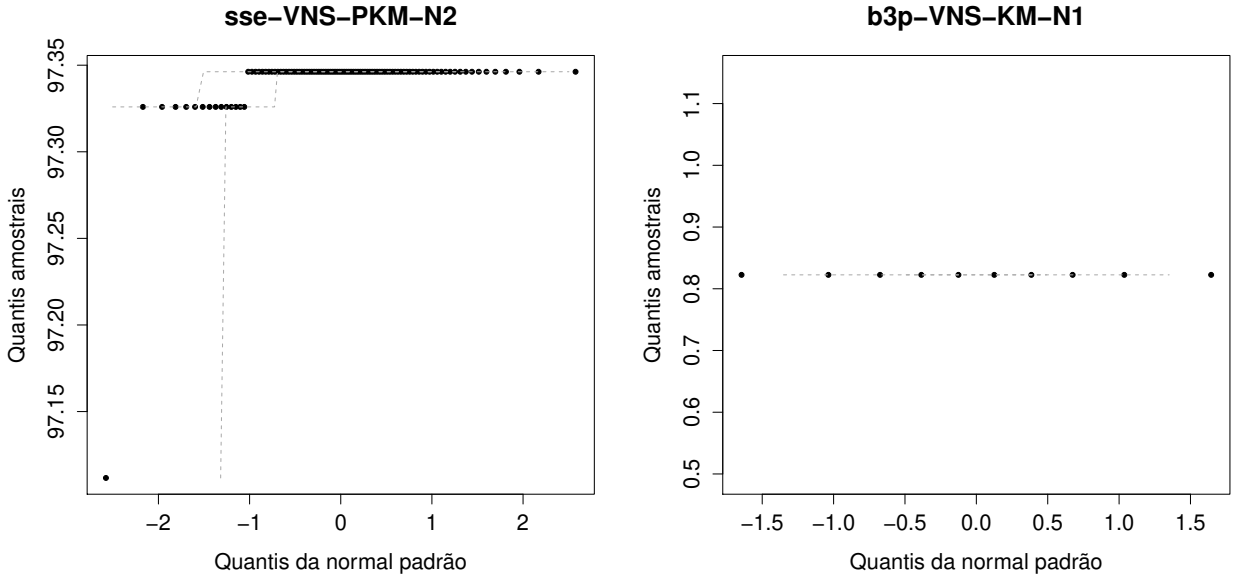


Fig. 5.1: Nos gráficos quantil-quantil, fica claro que os pontos estão fora do entorno da diagonal principal, indicando, assim, a não normalidade.

Por fim, construiu-se um intervalo de confiança com confiabilidade de 95% para os valores médios de desempenho, usando-se reamostragens do tamanho da população e $B = 10.000$.

A análise feita com os gráficos de perfil de desempenho permitem inferir a probabilidade de um determinado algoritmo se distanciar de até uma fração τ do melhor desempenho obtido. Seja o conjunto das 340 condições iniciais P_0 e o conjunto S dos 10 algoritmos testados, em que, para cada condição inicial $p \in P_0$ e $s \in S$, têm-se $\overline{B3F} = 1,1 - B3F$ e $\overline{B3P} = 1,1 - B3P$. Para um algoritmo s, a comparação de seu desempenho na resolução de um problema p, em relação ao desempenho do melhor algoritmo, é feita utilizando uma das seguintes razões de desempenho[4]:

$$r_{p,s} = \frac{1,1 - B3P}{\min\{1,1 - B3P : \forall s \in S\}} = \frac{\overline{B3P}}{\min\{\overline{B3P} : \forall s \in S\}}, \tag{5.1}$$

$$r_{p,s} = \frac{1,1 - B3F}{\min\{1,1 - B3F : \forall s \in S\}} = \frac{\overline{B3F}}{\min\{\overline{B3F} : \forall s \in S\}}. \tag{5.2}$$

O valor 1.1 das métricas $\overline{B3F}$ e $\overline{B3P}$ foi arbitrado para evitar que os termos $\overline{B3P}$ e $\overline{B3F}$ anulassem o denominador da razão de desempenho (r_{ps}).

A razão de desempenho mostra o comportamento do algoritmo na resolução de um problema com uma dada condição inicial. Entretanto, como o desejo é de avaliar o desempenho geral do algoritmo, definimos a função perfil de desempenho dada por:

$$\rho_s(\tau) = \frac{1}{340} \cdot \left| p \in P_0 : r_{p,s} \leq \tau \right|, \tag{5.3}$$

[4]As formulações $\overline{B3F}$ e $\overline{B3P}$ foram feitas porque o perfil de desempenho depende de uma métrica em que os valores menores indicam melhor desempenho.

em que $|\cdot|$ representa o número de elementos do conjunto. Perceba que o perfil de desempenho é a probabilidade da razão $r_{p,s}$ estar dentro de um fator τ da melhor razão. Dessa maneira, o perfil de desempenho pode ser considerado uma função de distribuição acumulada para a razão de desempenho dada por

$$\rho_s(\tau) \quad = P(r_{p,s} \leq \tau : 1 \leq s \leq 10). \tag{5.4}$$

5.2 Dados estáticos (*Benchmark*)

Os dados estáticos são estruturas de informação invariantes no tempo. Para iniciar a comparação entre os algoritmos, avaliou-se primeiramente seu desempenho em dados sintéticos estáticos. Esses dados foram sintetizados para avaliar o desempenho dos diferentes métodos em estruturas particulares de distribuição e dimensão, como os dados bidimensionais não globulares ou dados hiperdimensionais ($d >> n$) e globulares. A vantagem de testes em estruturas de dados bidimensionais é que se pode utilizar o discernimento visual para verificar a qualidade dos algoritmos a partir de gráficos como diagrama de Voronoi, envoltório convexo e trajetória central. E o motivo de se realizar testes em estruturas de dados hiperdimensionais é estimar o impacto que a alta dimensionalidade tem no critério de busca [4], ou seja, quão capaz o método é de mitigar os efeitos do fenômeno de Hughes[5].

5.2.1 Base de dados sintéticos

Em estruturas perfeitamente globulares, linearmente separáveis e de baixa dimensionalidade, todos os algoritmos têm um comportamento ótimo, até mesmo para um conjunto mal distribuído de pontos iniciais (Figura 5.2). Para esse tipo de dado, mesmo para métodos mais simples, como o k-médias sem reinicialização, as métricas *B-cubed* (precisão, lembrança e F de Rijsbergen) são quase sempre unitárias.

Entretanto, para estruturas não globulares com desequilíbrio severo, ou mesmo dados globulares hiperdimensionais, os algoritmos não necessariamente mantêm seu perfil de desempenho. Testar os métodos em dados com estas características é uma maneira de identificar os domínios de especialidade e de fraqueza dos algoritmos.

Estruturas não globulares de baixa dimensionalidade

As estruturas não globulares são formadas por dados cuja distribuição em relação a um ponto central não possui a mesma amplitude em todas as dimensões (*ex.:* Figura 5.3).

Essa base de dados é bem definida pelas seguintes características: **baixa dimensionalidade**, **distribuição não globular** e **desequilíbrio severo** entre as quantidades de pontos de cada grupo. Essa estrutura de dados, ao contrário da estrutura globular, é sensivelmente mais difícil de ser tratada pelos algoritmos tradicionais, como sugere a Tabela 5.1.

Para avaliar os benefícios de se usar a meta-heurística VNS vamos comparar as variações de desempenhos em três categorias: KM e suas variantes VNS-N1 e VNS-N2; CKM e suas variantes

[5] O fenômeno de Hughes é também conhecido como maldição da dimensionalidade (*curse of dimensionality*).

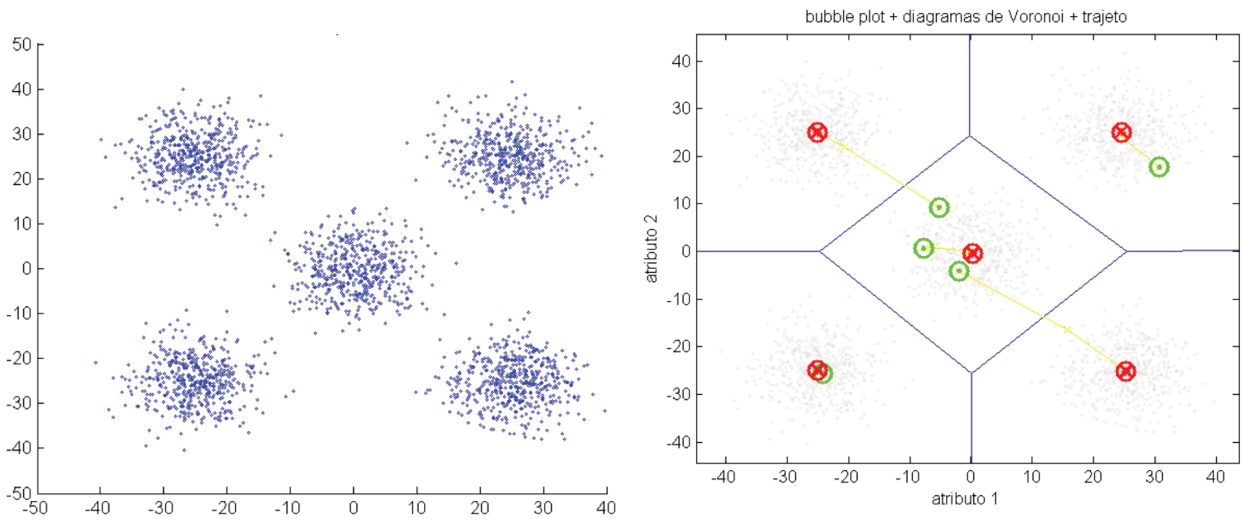


Fig. 5.2: A trajetória dos centroides obtidos durante as iterações inicia-se na marcação verde e finda-se na marca vermelha. Observa-se que, em poucas iterações, o algoritmo padrão *k*-médias já é capaz de criar agrupamentos ideais, mesmo partindo de pontos iniciais mal distribuídos.

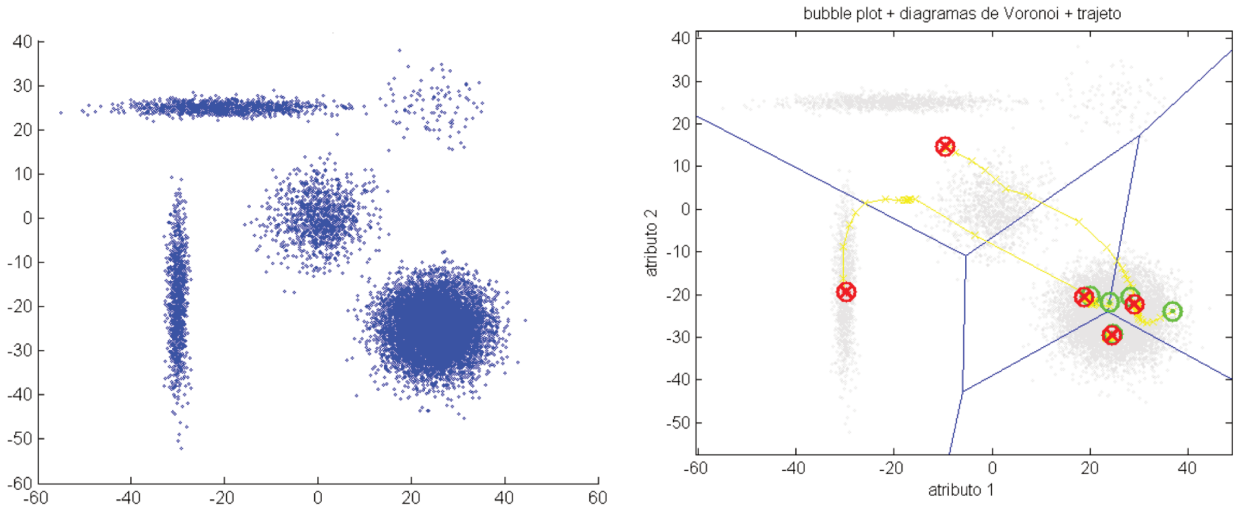


Fig. 5.3: A aplicação do *k*-médias padrão nesta estrutura de dados (desbalanceada e não globular) não foi capaz de gerar um agrupamento com coerência espacial.

VNS-N1 e VNS-N2; PKM e suas variantes VNS-N1 e VNS-N2. Os experimentos mostram que a VNS foi capaz de melhorar a variância intracluster em pelo menos 50% da vezes para o método PKM e quase em 90% para os métodos CKM e KM, lembrando que a comparação de desempenho não foi feita com o algoritmo *k-means* padrão e sim com sua variante com múltiplas inicializações. Além disso, para as variantes VNS-N1 dos algoritmos CKM e KM, o acoplamento foi capaz de gerar ganho aproximado em 80% das vezes nas métricas supervisionadas. Entretanto, no método PKM, as variantes VNS não foram capazes de superar os escores supervisionadas do método puro (ver Tabela 5.2). Provavelmente, isso decorre do fato de que a coerência entre redução de SSE e aumento de métricas B3 tem um limite, a partir do qual a redução do SSE implica um declínio da

Tab. 5.1: Valores de média (μ) e desvio padrão (σ) dos algoritmos aplicados a base de dados bidimensional não globular.

	métodos	B3P		B3R		B3F		SSE ($\times 10^3$)	
		μ	σ	μ	σ	μ	σ	μ	σ
1	KM	0,92	0,04	0,52	0,10	0,66	0,08	8,67	0,47
2	PKM-SS	1,00	0,00	1,00	0,00	1,00	0,00	8,64	0,00
3	PKM	1,00	0,00	1,00	0,00	1,00	0,00	8,64	0,00
4	CKM	0,92	0,04	0,52	0,07	0,66	0,07	8,64	0,49
5	VNS - KM - N1	0,98	0,00	0,62	0,00	0,76	0,00	7,73	0,01
6	VNS - KM - N2	0,91	0,03	0,50	0,05	0,65	0,05	8,71	0,42
7	VNS - CKM - N1	0,98	0,00	0,62	0,00	0,76	0,00	7,73	0,01
8	VNS - CKM - N2	0,92	0,04	0,51	0,06	0,66	0,06	8,60	0,50
9	VNS - PKM - N1	0,99	0,01	0,88	0,17	0,92	0,11	8,38	0,39
10	VNS - PKM - N2	0,99	0,01	0,90	0,16	0,93	0,10	8,42	0,37

qualidade do cluster, segundo as métricas B3.

Tab. 5.2: Lista em ordem decrescente de desempenho dos algoritmos para base de dados não globulares de baixa dimensionalidade

	B3F	B3P	B3R	SSE
1	PKM-SS	PKM-SS	PKM-SS	VNS-KM-N1
2	PKM	PKM	PKM	VNS-CKM-N1
3	VNS-PKM-N2	VNS-PKM-N2	VNS-PKM-N2	VNS-PKM-N1
4	VNS-PKM-N1	VNS-PKM-N1	VNS-PKM-N1	VNS-PKM-N2
5	VNS-KM-N1	VNS-KM-N1	VNS-KM-N1	VNS-CKM-N2
6	VNS-CKM-N1	VNS-CKM-N1	VNS-CKM-N1	PKM-SS
7	KM	VNS-CKM-N2	KM	PKM
8	CKM	CKM	CKM	CKM
9	VNS-CKM-N2	KM	VNS-CKM-N2	KM
10	VNS-KM-N2	VNS-KM-N2	VNS-KM-N2	VNS-KM-N2

Ainda sim, como pode ser visto pela comparação dos gráficos de intervalo de confiança das métricas B3F e SSE (Fig. 5.4), as variantes VNS-N1 têm o melhor desempenho em SSE ao mesmo tempo que alcançam a segunda melhor categoria de desempenho na métrica B3F.

Aparentemente os algoritmos sem restrições (KM), ou que não tolerem violações de restrições de paridade (CKM), forçam a descida do SSE sem avaliar consistências adicionais, como a distribuição e densidade, que acabam sendo tratadas de forma indireta pelos métodos PKM. Para estes métodos, o VNS-N1 não só foi capaz de reduzir a SSE como também ampliou os indicadores supervisionados (B3).

Em uma comparação global, as estatísticas observadas sugerem que os métodos PKM-SS e PKM sobrepõe-se a todos os demais algoritmos nos escores de métricas $B3$. Entretanto, os

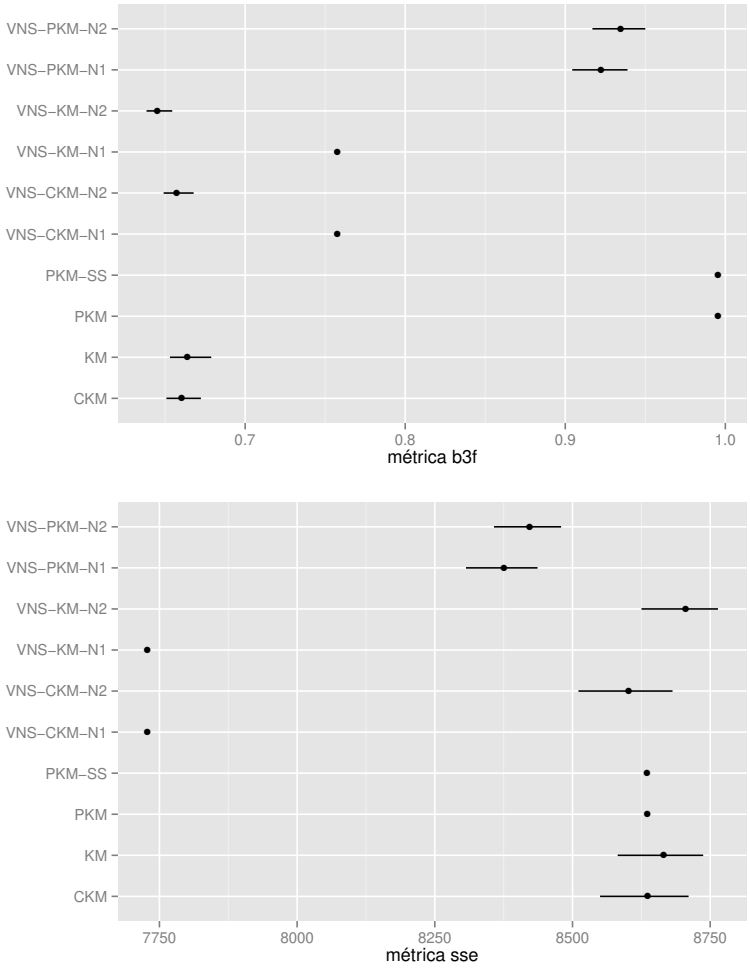


Fig. 5.4: Intervalos de confiança para a medida supervisionada B3F e desempenhos da variação intracluster para dados globulares.

melhores desempenhos SSE foram dos algoritmos acoplados com VNS na estrutura de vizinhança N1 (ver Tabela 5.2).

Utilizando o gráfico de perfil de desempenho (ver Fig. 5.5) fica claro que existe dominância das variantes PKM sobre os demais algoritmos. Do gráfico observa-se que $P(r_{ps} < 3) < 5\%$ para as variantes de KM e CKM. Considerando a relação proposta pela Equação 5.2 têm-se que

$$3 > r_{p,s} \quad = \quad \frac{\overline{B3F}}{\min\{\overline{B3F} : \forall s \in S\}} \quad = \quad \frac{1,1 - B3F}{\min\{1,1 - B3F : \forall s \in S\}} \Leftrightarrow \frac{1,1 - B3F}{0,1} < 3 \Leftrightarrow B3F > 0,8.$$

Portanto a chance de $B3F > 0,8$ é menor que 5% para as variantes KM e CKM, enquanto que para as variantes PKM essa probabilidade é de pelo menos 65%. Disso, é razoável afirmar que a busca pelo menor SSE é uma boa forma de orientar os métodos, mas estratégias de diversificação e restrições espaciais podem gerar grupos que além de coesos também são ótimos em uma perspectica supervisionada.

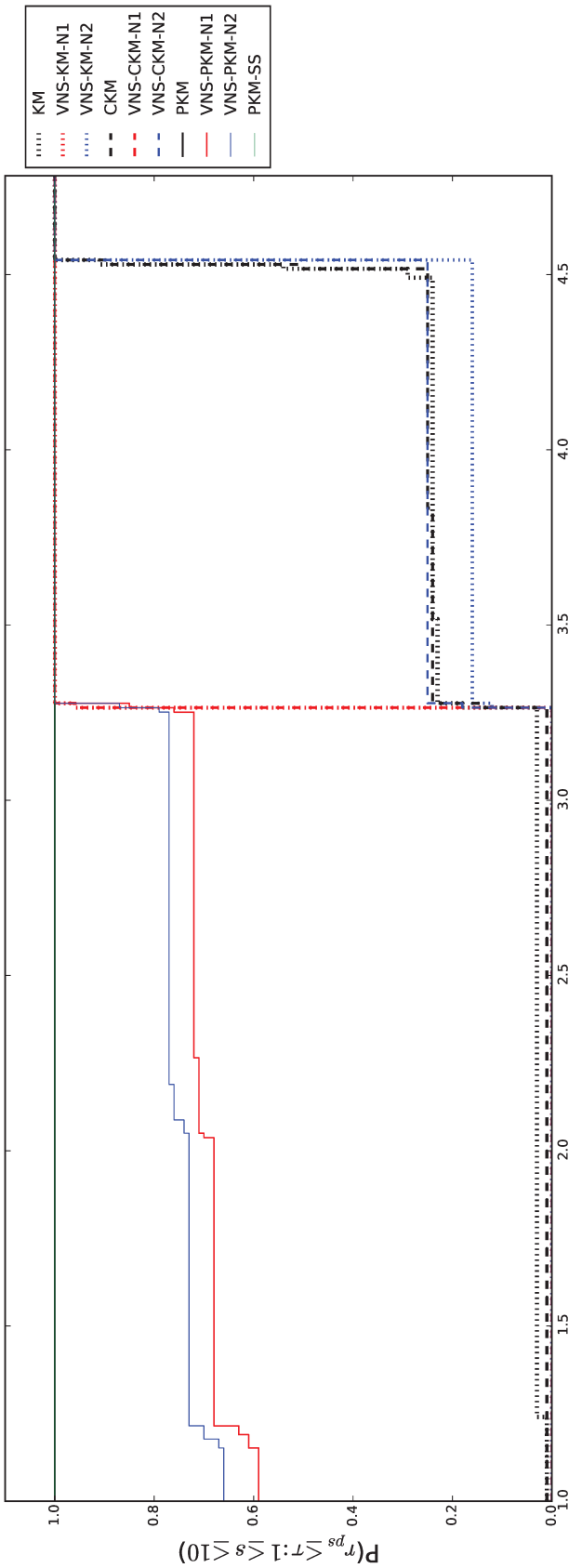


Fig. 5.5: Perfil de desempenho da métrica $\overline{B3F}$ na base de dados não globulares para uma variedade de 100 condições iniciais.

Estruturas hiperdimensionais

As estruturas de dados com alta dimensionalidade são interessantes pois, em teoria de classificadores e agrupamentos, sabe-se que o desempenho de todo algoritmo tende a degradar a partir de um certo número de atributos. Isso é intuitivo quando admitimos que atributos que não são relevantes para a determinação da classe, como atributos correlacionados entre si ou descorrelacionados com a classe, são uma espécie de ruído para o processo de classificação. Por isso é importante avaliar o comportamento do algoritmo em estruturas hiperdimensionais, sendo assim, usou-se 500 pontos, de dimensão 1.000, distribuídos igualmente e normalmente entre 5 centros com alto nível de separação, i.e., nos quais os elementos são distribuídos de forma globular, com classes de mesmo tamanho e linearmente separáveis. Apesar de os dados serem convenientemente distribuídos, existe uma degradação severa da capacidade dos algoritmos tradicionais de identificarem os grupos de forma correta (ver Tabela 5.3), em virtude da alta dimensionalidade.

Tab. 5.3: Valores de média (μ) e desvio padrão (σ) dos algoritmos aplicados a base de dados globulares de alta dimensionalidade.

	métodos	B3P		B3R		B3F		SSE ($\times 10^3$)	
		μ	σ	μ	σ	μ	σ	μ	σ
1	KM	0,47	0,18	0,59	0,17	0,52	0,18	79,04	0,17
2	PKM-SS	0,62	0,19	0,76	0,13	0,68	0,16	78,91	0,17
3	PKM	0,96	0,01	0,96	0,01	0,96	0,01	78,63	0,00
4	CKM	0,46	0,16	0,60	0,17	0,52	0,16	79,05	0,15
5	VNS - KM - N1	0,55	0,17	0,67	0,15	0,60	0,16	78,97	0,15
6	VNS - KM - N2	0,57	0,18	0,69	0,16	0,62	0,17	78,95	0,16
7	VNS - CKM - N1	0,56	0,16	0,68	0,15	0,61	0,15	78,95	0,14
8	VNS - CKM - N2	0,58	0,16	0,69	0,15	0,63	0,15	78,94	0,14
9	VNS - PKM - N1	0,97	0,01	0,97	0,01	0,97	0,01	78,62	0,00
10	VNS - PKM - N2	0,97	0,01	0,97	0,01	0,97	0,01	78,62	0,00

Os experimentos mostram que a VNS foi capaz de melhorar a variância intracluster em mais de 50% da vezes para o método KM e em mais de 95% das vezes para os métodos CKM e PKM. Além disso, as variantes VNS foram capazes de melhorar os desempenhos da métrica B3F de forma coerente com a SSE. Ou seja, para as variantes KM, as variantes VNS-N1 e VNS-N2 aumentaram o indicador B3F, respectivamente em 76% e 71% das vezes, enquanto as variantes VNS-CKM e VNS-PKM foram capazes de melhorar o indicador B3F mais de 95% das vezes, independente da estrutura de vizinhança. As versões acopladas, como visto na Tabela 5.4, foram superiores em todas as métricas.

Em estruturas globulares, não importando a dimensão, os melhores resultados de constituição de grupo derivam dos agrupamentos com menor variação intracluster, ou seja, para estes dados, diminuir o valor de SSE implica na melhoria dos indicadores B3. As estimativas intervalares do desempenho dos algoritmos apresentadas no Gráfico 5.6 sugerem que todos os algoritmos acoplados à VNS se sobrepuseram, em desempenho extrínseco e intrínseco, às suas versões sem

Tab. 5.4: Lista em ordem decrescente de desempenho dos algoritmos para base de dados globulares de alta dimensionalidade.

	B3F	B3P	B3R	SSE
1	VNS-PKM-N1	VNS-PKM-N1	VNS-PKM-N1	VNS-PKM-N1
2	VNS-PKM-N2	VNS-PKM-N2	VNS-PKM-N2	VNS-PKM-N2
3	PKM	PKM	PKM	PKM
4	PKM-SS	PKM-SS	PKM-SS	PKM-SS
5	VNS-CKM-N2	VNS-CKM-N2	VNS-KM-N2	VNS-CKM-N2
6	VNS-KM-N2	VNS-KM-N2	VNS-CKM-N2	VNS-KM-N2
7	VNS-CKM-N1	VNS-CKM-N1	VNS-CKM-N1	VNS-CKM-N1
8	VNS-KM-N1	VNS-KM-N1	VNS-KM-N1	VNS-KM-N1
9	KM	KM	CKM	KM
10	CKM	CKM	KM	CKM

VNS. E não existe nenhuma evidência estatística de que as vizinhanças N1 e N2 afetam de forma distinta o desempenho dos algoritmos em estruturas de distribuição com os perfis apresentados.

Avaliando o desempenho global, as heurísticas PKM-SS e PKM tiveram um desempenho de destaque. Aquela com pontos de inicialização (PKM-SS) superou inclusive a concorrente semissuperviosionada CKM, que contava com mais informações para sua inicialização. Entretanto, as versões PKM com barreira (VNS-PKM-N1 e VNS-PKM-N2) tiveram qualidade melhor de agrupamento em todos os indicadores, como demonstra o gráfico de perfil de desempenho da Fig. 5.7.

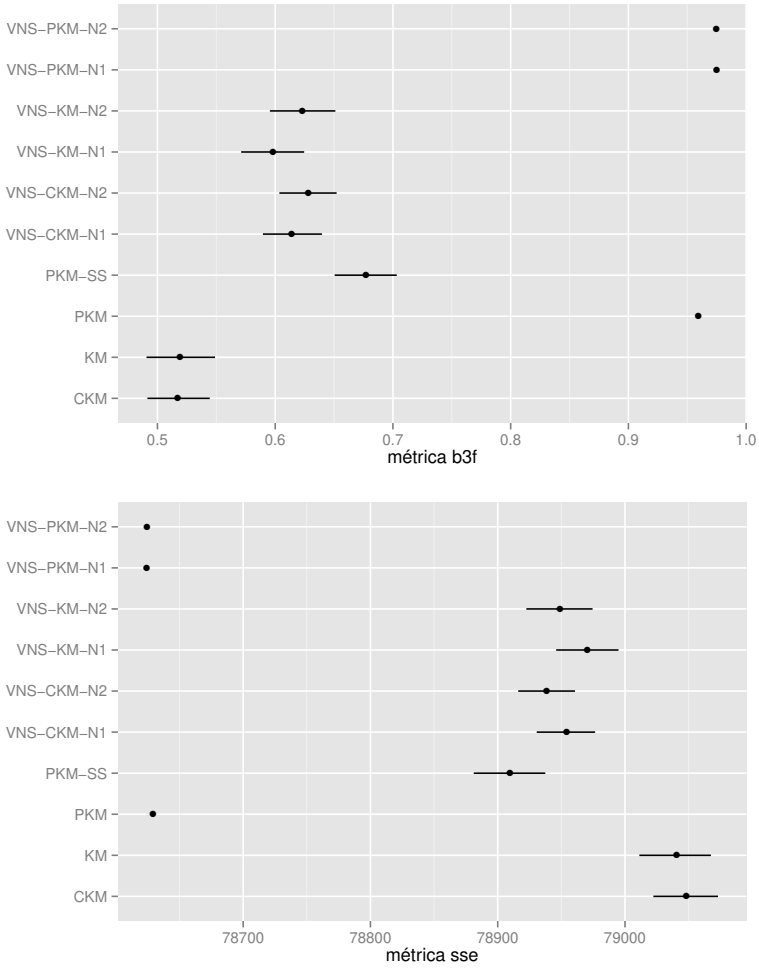


Fig. 5.6: Nos dados hiperdimensionais, estabeleceu-se uma clara coerência entre a otimização da *SSE* e a melhoria do indicador *B3F*.

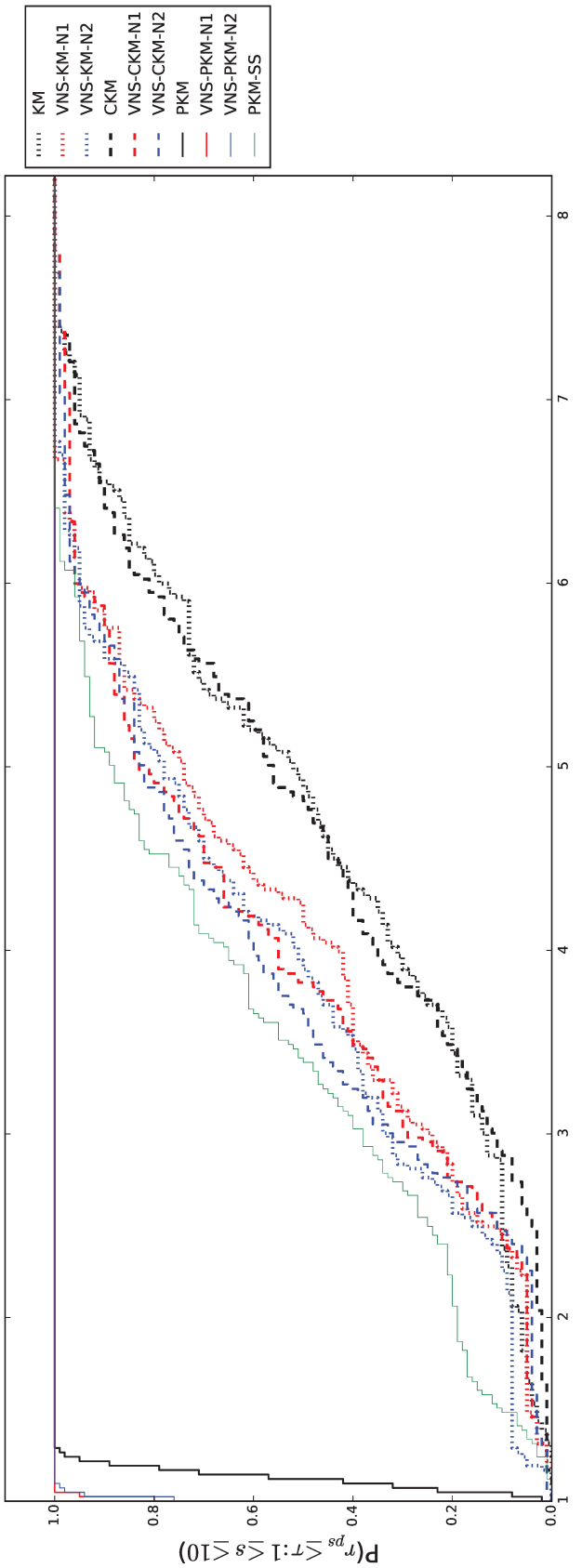


Fig. 5.7: Perfil de desempenho da métrica *B3F* na base de dados hiperdimensionais para uma variedade de 100 condições iniciais.

5.2.2 Base de dados Íris - *Iris Plants*

A base de dados Íris[6] (ver Figura 5.8) provavelmente é a mais conhecida na literatura de reconhecimento de padrões. É uma base de dados encontrada em artigos de classificação e agrupamentos e, por ser frequentemente usada em testes, torna-se um meio de comparação entre os diversos métodos. Das três classes presentes na base, uma é linearmente separável das demais, enquanto as outras duas não são linearmente separáveis entre si.

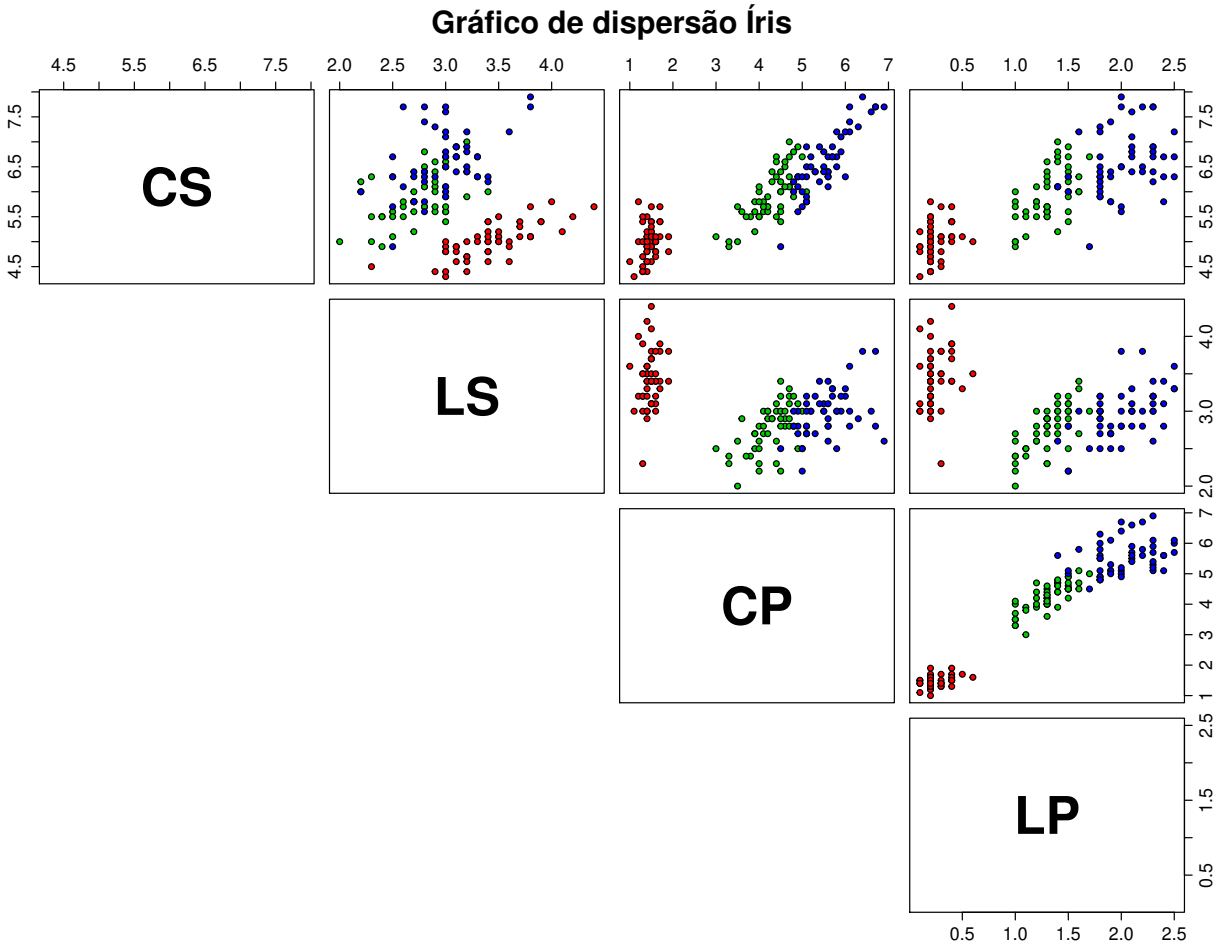


Fig. 5.8: A base de dados Íris contém 150 amostras de 4 variáveis constituídas de medidas (em cm) do comprimento das sépalas (CS), da largura das sépalas (LS), comprimento das pétalas (CP) e largura das pétalas (LP) de três espécies de flor íris: *Setosa*, *Virginica* e *Versicolor*.

As variantes VNS dos métodos CKM e PKM sempre preservam ou reduzem a variação intracluster, pois é intrínseco à meta-heurística VNS a busca de soluções apenas se existe redução de SSE. Observando o desempenho dos algoritmos nos dados Íris, observou-se que a estrutura de vizinhança N1 foi sensivelmente melhor que a N2. Quando acoplada a CKM, ela foi capaz de

[6]Essa base é disponibilizada gratuitamente pela *UCI Machine learning Repository* pelo site `http://archive.ics.uci.edu/ml/`, acessado em 03/03/2015.

reduzir a SSE em mais da metade das vezes e, quando acoplada no método PKM, reduziu a SSE em 80% das vezes.

No que se refere às variante KM, a comparação é feita entre o KM com múltiplos reinícios (*multi-start K-means*) e as variantes VNS. Nesta comparação, as variantes VNS foram capazes de melhorar o resultado em 38% e mantêm sua qualidade em 40% das vezes. E todos os casos em que houve superação do KM sem uso da VNS, isso ocorreu com o uso da estrutura de vizinhança N2. Entretanto, a Tabela 5.5 indica que a média de desempenho da VNS é muito superior às versões concorrentes, independentemente da estrutura de vizinhança.

Tab. 5.5: Valores de média (μ) e desvio padrão (σ) dos algoritmos aplicados a base de dados Íris.

	métodos	B3P		B3R		B3F		SSE	
		μ	σ	μ	σ	μ	σ	μ	σ
1	KM	0,77	0,09	0,82	0,02	0,79	0,06	104,80	11,73
2	PKM-SS	0,82	0,01	0,83	0,01	0,82	0,01	97,34	0,01
3	PKM	0,82	0,00	0,83	0,00	0,82	0,00	97,34	0,01
4	CKM	0,78	0,08	0,82	0,02	0,80	0,05	103,43	11,05
5	VNS - KM - N1	0,83	0,00	0,84	0,00	0,83	0,00	97,33	0,00
6	VNS - KM - N2	0,82	0,01	0,83	0,01	0,83	0,01	97,34	0,01
7	VNS - CKM - N1	0,83	0,00	0,84	0,00	0,83	0,00	97,33	0,00
8	VNS - CKM - N2	0,82	0,02	0,83	0,01	0,82	0,01	97,67	3,32
9	VNS - PKM - N1	0,83	0,00	0,84	0,00	0,83	0,00	97,33	0,00
10	VNS - PKM - N2	0,82	0,01	0,83	0,01	0,82	0,01	97,34	0,01

Apesar da Tabela 5.6 fornecer uma hierarquia dos métodos, vários deles apresentam valores muito próximos ou idênticos para as métricas. Assim, como visto na Figura 5.9, os desempenhos se distinguem em três categorias: melhores resultados, alcançados pela variantes VNS-N1; bons resultados, obtidos pelas variantes VNS-N2 e resultados inferiores, obtidos por KM e CKM.

Tab. 5.6: Lista em ordem decrescente de desempenho dos algoritmos para a base Íris.

	B3F	B3P	B3R	SSE
1	VNS-CKM-N1	VNS-CKM-N1	VNS-CKM-N1	VNS-CKM-N1
2	VNS-PKM-N1	VNS-PKM-N1	VNS-PKM-N1	VNS-PKM-N1
3	VNS-KM-N1	VNS-KM-N1	VNS-KM-N1	VNS-KM-N1
4	VNS-KM-N2	VNS-KM-N2	VNS-CKM-N2	VNS-KM-N2
5	PKM-SS	PKM-SS	VNS-KM-N2	PKM-SS
6	VNS-CKM-N2	VNS-PKM-N2	PKM-SS	VNS-PKM-N2
7	VNS-PKM-N2	VNS-CKM-N2	VNS-PKM-N2	PKM
8	PKM	PKM	PKM	VNS-CKM-N2
9	CKM	CKM	CKM	CKM
10	KM	KM	KM	KM

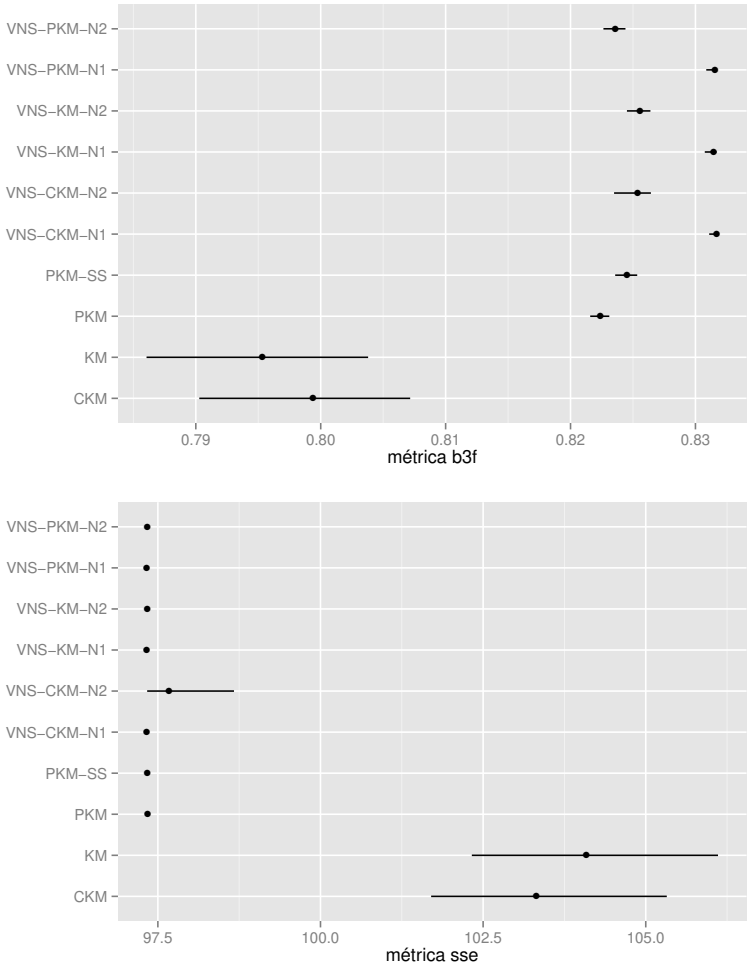


Fig. 5.9: Intervalos de confiança para a medida supervisionada B3F e desempenhos da variação intracluster para dados Íris.

Avaliando o gráfico de perfil de desempenho (Fig. 5.10), percebe-se que, globalmente, os algoritmos acoplados com VNS com estrutura de vizinhança do tipo $N1$ foram superiores aos demais métodos em todas as métricas (ver tabela 5.5), ou seja, para esta base de dados, não importando a heurística de busca local, o modelo acoplado com VNS com estrutura de vizinhança baseada em distância Euclidiana tem um desempenho melhor que os demais métodos.

Dentre os resultados obtidos, os métodos k-médias com múltiplos reinícios (KM) e o CKM se mostraram muito abaixo da média de desempenho dos demais algoritmos. E os métodos PKM e PKM-SS não acoplados à VNS conseguiram manter um desempenho competitivo com os métodos acoplados.

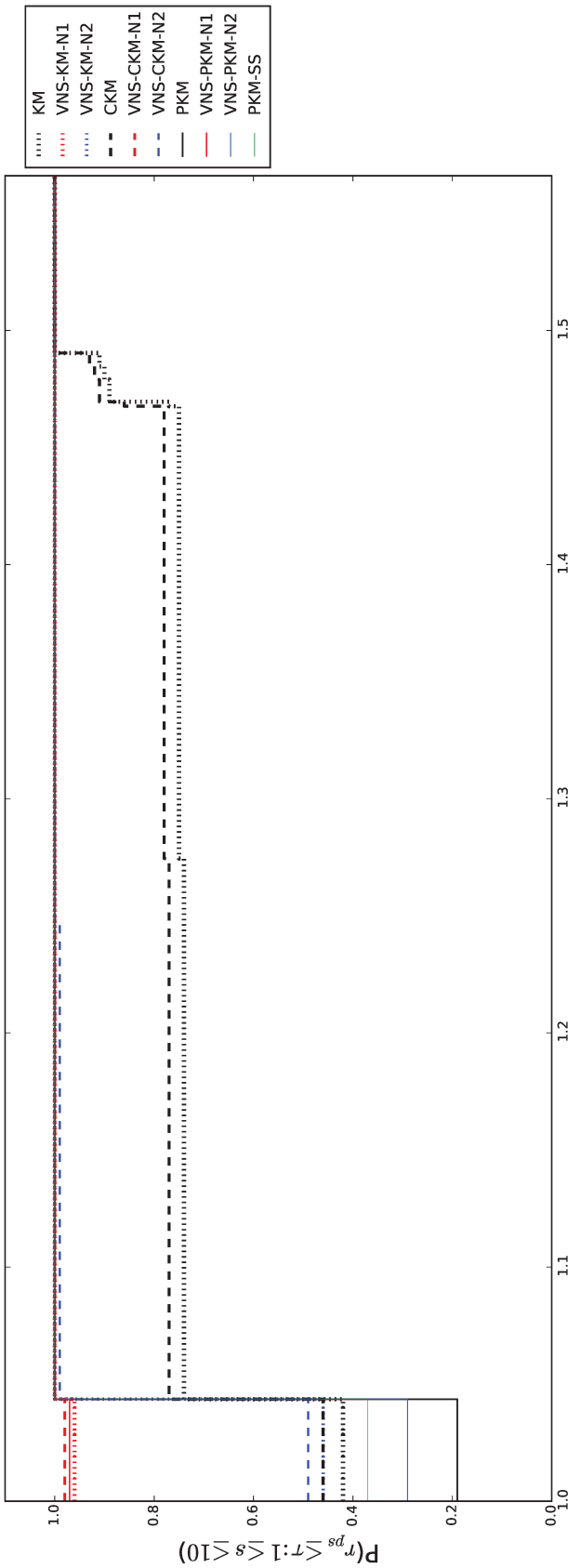


Fig. 5.10: Perfil de desempenho da métrica *B3F* na base de dados Íris para uma variedade de 100 condições iniciais.

5.3 Séries temporais de imagens *NDVI*

"I have not failed. I've just found 10,000 ways that won't work."
- Thomas Edison

5.3.1 Tratamento de séries temporais de imagens

Embora muitos métodos de classificação e agrupamento de séries temporais de imagens para a agricultura tenham sido desenvolvidos, tratar este problema é, mesmo nos dias atuais, particularmente difícil. Alguns métodos são melhores em uma determinada cultura e região, e piores em outras. A classificação digital é frequentemente dependente do resultado da geometria do imageamento e do ruído atmosférico. Além disso, a correção radiométrica pode gerar erros e, quando a dimensão dos dados é muito grande, torna-se necessário mitigar o fenômeno de Hughes da análise dos dados, também conhecido como maldição da dimensionalidade (*curse of dimensionality*).

As imagens de satélites aqui tratadas são referentes aos sensores Terra/Modis e AVHRR/NOAA, para as quais escolheu-se o NDVI como índice sumarizante das refletâncias registradas pelos canais espectrais (ver Seção 4.5). Todas as séries de NDVI foram obtidas de imagens processadas de acordo com os seguintes passos: conversão do formato bruto para um formato intermediário, calibração radiométrica, correção geométrica e atenuação do efeito das nuvens através da geração de composição de máximo valor. O processamento foi realizado pelo sistema NAVPRO [8], que garante que cada imagem tenha menos de 30% de pixels cobertos com nuvens e que não haja ruído.

Composição por máximos valores

A escolha do NDVI decorre de sua capacidade descritiva da cobertura da área visada, como visto na Seção 2.1.3. Entretanto, o valor real do NDVI de um pixel pode ser distorcido por condições atmosféricas, uma vez que nuvens podem diminuir seu valor ou mesmo torná-lo negativo. Para enfrentar este problema, usou-se o método de geração de composição por máximos valores (mvc, do inglês *maximum value composition*). Esse método consiste em usar o valor máximo de um pixel, em uma sequência de imagens, para construir uma única imagem final (ver Fig. 5.11). Imagens mvc de NDVI têm redução nos impactos da refletância direcional, nos efeitos de sombra e nos efeitos de partículas no ar [57].

Esse método, associado ao grande volume de imagens com georreferenciamento preciso, permite construir composições a partir das séries temporais de imagens. Dessa forma, *pixels* afetados por estes efeitos de distorção têm menos chance de fazer parte da composição final. Quanto maior a série, melhor será a qualidade da atenuação. Entretanto, deve-se observar que o uso de períodos muito longos pode comprometer a análise da dinâmica de cobertura da área investigada, pois séries muito longas contêm imagens de superfícies cuja cobertura varia no decorrer do imageamento.

Para os tratamentos de pré-processamento e a geração de composições de máximo valor de NDVI, fez-se uso do sistema *NAVPRO* [10, 36], que processa os dados brutos e os converte em composições de máximo valor (mvc) de NDVI. O sistema NAVPRO é capaz de

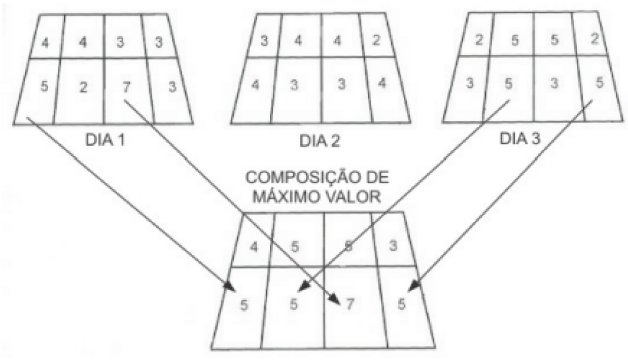


Fig. 5.11: Elaboração de uma composição de máximo valor a partir de dados diários.
Fonte: Embrapa Informática Agropecuária. Comunicado técnico 107 de Abril de 2012.

gerar automaticamente imagens com deslocamentos máximos de 1 pixel, valor aceitável para aplicações com dados de baixa resolução espacial como as imagens NOAA.

Extração de características

É razoável pensar em estratégias que transformem os dados dinâmicos em estruturas mais adequadas aos algoritmos, sem perda das características originais. Para as séries temporais, supôs-se que suas principais informações são derivadas de sua curva. Sendo assim, propôs-se o descritor $< \mathscr{F}, L_2 >$, apresentado na Seção 4.5.1, cuja transformação $\mathscr{F}$ substitui as séries originais por vetores de números que sejam descritivos do desenho da curva de ajuste dos dados.

Este descritor usa a métrica usual de espaço Euclidiano para lidar com os dados em um espaço de atributos reduzido, em que cada série temporal de NDVI é transformada por $\mathscr{F}$ em um vetor constituído dos coeficientes obtidos pelo ajuste do polinômio trigonométrico da série original. Portanto, faz-se a substituição da série temporal pelos coeficientes de sua representação na série discreta de Fourier. A esse vetor deu-se o nome de **vetor característico** e cada coordenada deste vetor chama-se **característica**. A seguir apresentamos o descritor $< \mathscr{F}, L_2 >$ permite que se mantenha, ou mesmo se amplie, a capacidade preditiva dos classificadores em um espaço significativamente menor, funcionando, desta forma, como um método de redução de atributos.

5.3.2 AVHRR/NOAA - Jaboticabal 2004/2005

A amostra analisada constitui-se de 5.000 instâncias, sem valores faltantes, das quais 492 referem-se a áreas cobertas com cana-de-açúcar (9,84%) e 4.508 a áreas com outras coberturas que não cana-de-açúcar (90,16%). Usou-se as composições de máximo valor de séries de imagens NDVI extraídas de imagens AVHRR/NOAA 17 da área agrícola referente à cidade de Jabotical, no estado de São Paulo, Brasil, entre os períodos de abril/2004 e março/2005. Para cada ano safra, as séries de NDVI variaram desde o início do plantio, passando pelo período de maior vigor vegetativo e indo até o final da colheita.

Embora já existam disponíveis sensores de maior qualidade espectral e espacial, como MODIS, SPOT/VEGETATION e WFI, as imagens AVHRR-NOAA continuam sendo de grande uso em estudos

envolvendo a análise de ecossistemas, em função da disponibilidade de longas séries temporais de imagens, pelo grande acervo histórico de dados NOAA. Como exemplo, o Centro de Pesquisas Meteorológicas e Climáticas Aplicadas à Agricultura (CEPAGRI) da Universidade Estadual de Campinas (UNICAMP) possui um banco crescente de imagens AVHRR/NOAA iniciado em abril de 1995. E ainda que as imagens AVHRR/NOAA apresentem resolução espacial e espectral menores em comparação a outros sensores, elas possuem características adequadas para o estudo de alvos com grande dinâmica espectral, tais como sua alta resolução temporal, cobertura diária, garantia de cobertura global e gratuidade das imagens.

Pixels de controle

Como pixels de controle (também referidos como *verdade terrestre* ou *padrão ouro*), tem-se uma amostra no estado de SP, Brasil, dividida em duas categorias: cana e não-cana. Os pixels de cana foram obtidos de uma máscara que elimina áreas urbanas, solo e outros tipos de vegetação. Essa máscara foi produzida a partir de especialistas humanos que classificaram imagens *LandSat* (projeto CANASAT). Em seguida, ajustou-se essas imagens para imagens AVHRR/NOAA e selecionou-se os pixels com valores de NDVI entre 0 e 0,75. Já os pixels de não cana-de-açúcar foram obtidos da máscara negativa dos pixels de cana-de-açúcar.

5.3.3 Análise comparativa

As Tabelas 5.7 e 5.8 apresentam, respectivamente, o desempenho dos algoritmos quando aplicados aos dados em sua forma bruta e transformada, enquanto as Tabelas 5.9 e 5.10 apresentam o ranqueamento dos algoritmos com base nesse desempenho.

Tab. 5.7: Valores de média e desvio padrão dos algoritmos aplicados a base de Jaboticabal no formato bruto.

	métodos	B3P ($\times 10^{-2}$)		B3R ($\times 10^{-2}$)		B3F ($\times 10^{-2}$)		SSE ($\times 10^{6}$)	
		μ	σ	μ	σ	μ	σ	μ	σ
1	KM	82,25	0,00	99,87	0,04	90,21	0,02	3,30	0,59
2	PKM-SS	82,25	0,00	99,88	0,04	90,21	0,02	3,36	0,60
3	PKM	82,51	0,32	83,93	20,09	81,97	10,38	3,85	0,14
4	CKM	82,25	0,00	99,86	0,04	90,21	0,02	3,18	0,55
5	VNS - KM - N1	82,25	0,00	99,84	0,00	90,20	0,00	2,83	0,00
6	VNS - KM - N2	82,25	0,00	99,84	0,00	90,20	0,00	2,83	0,00
7	VNS - CKM - N1	82,25	0,00	99,84	0,00	90,20	0,00	2,83	0,00
8	VNS - CKM - N2	82,25	0,00	99,84	0,00	90,20	0,00	2,83	0,00
9	VNS - PKM - N1	82,37	0,15	90,93	11,46	86,09	5,30	3,09	0,91
10	VNS - PKM - N2	82,38	0,16	90,38	11,89	85,82	5,51	3,38	0,57

Vale observar que a mudança de ênfase nas Tabelas 5.7 e 5.8 da métrica B3F para a B3P se deu pelas idiossincrasias da base de Jaboticabal, que é uma base de dados binários, com desequilíbrio severo (dominância de uma das classes). Para essa base de dados, considera-se a precisão (B3P)

Tab. 5.8: Valores de média e desvio padrão dos algoritmos aplicados a base de dados Jaboticabal com transformação e redução de atributos para 3 dimensões.

	métodos	B3P ($\times 10^{-2}$)		B3R ($\times 10^{-2}$)		B3F ($\times 10^{-2}$)		SSE ($\times 10^{3}$)	
		μ	σ	μ	σ	μ	σ	μ	σ
1	KM	82,25	0,00	99,84	0,00	90,20	0,00	1,49	0,00
2	PKM-SS	82,25	0,00	99,84	0,00	90,20	0,00	1,49	0,00
3	PKM	82,54	0,36	82,58	21,70	81,06	11,49	693,97	46,41
4	CKM	82,25	0,00	99,84	0,00	90,20	0,00	1,49	0,00
5	VNS - KM - N1	82,25	0,00	99,84	0,00	90,20	0,00	1,49	0,00
6	VNS - KM - N2	82,25	0,00	99,84	0,00	90,20	0,00	1,49	0,00
7	VNS - CKM - N1	82,25	0,00	99,84	0,00	90,20	0,00	1,49	0,00
8	VNS - CKM - N2	82,25	0,00	99,84	0,00	90,20	0,00	1,49	0,00
9	VNS - PKM - N1	82,54	0,36	81,15	23,49	80,03	12,78	475,92	226,68
10	VNS - PKM - N2	82,54	0,36	81,15	23,49	80,03	12,78	693,96	46,40

Tab. 5.9: Lista em ordem decrescente de desempenho dos algoritmos para a base de Jaboticabal no formato bruto.

	B3F	B3P	B3R	SSE
1	PKM-SS	PKM	PKM-SS	VNS-KM-N1
2	KM	VNS-PKM-N2	KM	VNS-KM-N2
3	CKM	VNS-PKM-N1	CKM	VNS-CKM-N1
4	VNS-KM-N1	VNS-KM-N1	VNS-KM-N1	VNS-CKM-N2
5	VNS-KM-N2	VNS-KM-N2	VNS-KM-N2	VNS-PKM-N1
6	VNS-CKM-N1	VNS-CKM-N1	VNS-CKM-N1	CKM
7	VNS-CKM-N2	VNS-CKM-N2	VNS-CKM-N2	KM
8	VNS-PKM-N1	CKM	VNS-PKM-N1	PKM-SS
9	VNS-PKM-N2	KM	VNS-PKM-N2	VNS-PKM-N2
10	PKM	PKM-SS	PKM	PKM

mais indicada para validar a performance comparativa dos algoritmos, pois a métrica precisão (B3P) mede o desbalanceamento entre as classes, fazendo uma penalização por se eliminar grupos pequenos, enquanto que a lembrança (B3R), por medir a completeza, não é sensível à degradação de grupos pequenos. Uma forma de evidenciar a inabilidade da métrica B3R de lidar com grupos pequenos é classificar todos os elementos como sendo do mesmo grupo e medir seu valor. Para este caso, apesar da degeneração das classes, a lembrança assume valor máximo. Vale indicar que, nesse tipo de dado, a lembrança e o F de Risemberg se tornam indicadores secundários em relação à precisão (B3P).

Os intervalos de confiança representados nos gráficos A e D da Figura 5.12 sugerem que a **transformação dos dados** só afeta o desempenho dos algoritmos VNS-PKM. Observa-se que, apesar da significativa diferença dimensional entre as bases bruta e transformada (redução de 12 para 3), a transformação manteve a qualidade de agrupamento para maioria dos algoritmos

Tab. 5.10: Lista em ordem decrescente de desempenho dos algoritmos para a base tratada de Jaboticabal ($d = 3$).

	B3F	B3P	B3R	SSE
1	KM	PKM	KM	KM
2	PKM-SS	VNS-PKM-N1	PKM-SS	PKM-SS
3	CKM	VNS-PKM-N2	CKM	CKM
4	VNS-KM-N1	KM	VNS-KM-N1	VNS-KM-N1
5	VNS-KM-N2	PKM-SS	VNS-KM-N2	VNS-KM-N2
6	VNS-CKM-N1	CKM	VNS-CKM-N1	VNS-CKM-N1
7	VNS-CKM-N2	VNS-KM-N1	VNS-CKM-N2	VNS-CKM-N2
8	PKM	VNS-KM-N2	PKM	VNS-PKM-N1
9	VNS-PKM-N1	VNS-CKM-N1	VNS-PKM-N1	VNS-PKM-N2
10	VNS-PKM-N2	VNS-CKM-N2	VNS-PKM-N2	PKM

e ampliou o indicador (B3P) para as variantes VNS-PKM. Dessa forma, podemos considerar a transformação um método de redução da dimensionalidade que não afeta a capacidade discriminatória dos dados, podendo em alguns casos ampliá-la.

Avaliando-se em pares os gráficos A e B e os gráficos C e D da Figura 5.12, percebe-se que não existe **coerência** entre o B3P e o SSE nem para os dados brutos, nem para os transformados. Nos dados transformados observa-se que os valores piores de SSE referem-se aos algoritmos que tiveram melhor desempenho na precisão (B3P).

Observando o **efeito do uso da VNS**, têm-se que as variantes VNS dos algoritmos KM, CKM melhoraram a variação intracluster (SSE) em 30% e 40%, respectivamente, quando comparados às suas versões sem VNS, sem alterar as métricas supervisionadas (B3P, B3R e B3F). Por sua vez, as variantes VNS do PKM melhoraram a SSE em 90% das vezes, quando comparadas ao PKM, além de aumentarem o valor da lembrança (B3R) e o F de Risemberg (B3F) em 40% das vezes.

O desempenho dos algoritmos PKM e suas variantes VNS nas métricas B3R e B3F, como observado nas Tabelas 5.7 e 5.8, pode sugerir que, para dados de duas classes e altamente desbalanceados, o atual método de criação de barreiras talvez seja muito restritivo, pois compromete a qualidade do agrupamento sobre o ponto de vista da lembrança (B3R). Entretanto, ratifica-se que para dados bidimensionais e desbalanceados, a lembrança (B3R) é uma medida menos adequada que a precisão (B3P).

Observando as Tabelas 5.9 e 5.10, vê-se que o método PKM teve um desempenho contraditório, sendo o melhor em B3F e o pior em SSE. Suas variantes VNS conseguiram estabelecer um equilíbrio entre as métricas, mantendo altos valores de lembrança (B3R) com valores relativamente reduzidos de SSE. Em uma perspectiva intervalar (ver Figura 5.12), intui-se que o algoritmo VNS-PKM-N1 é a melhor opção para gerar agrupamentos em dados que se assemelhem à base de Jaboticabal, dado que ele estabeleceu uma coerência entre bons valores de B3P com uma boa coesividade de grupo (baixos valores de SSE).

Os resultados mostram que, para dados deste perfil, a transformação não altera a hierarquia de desempenho dos algoritmos. Os gráficos das Figuras 5.13 e 5.14 indicam que o algoritmo PKM

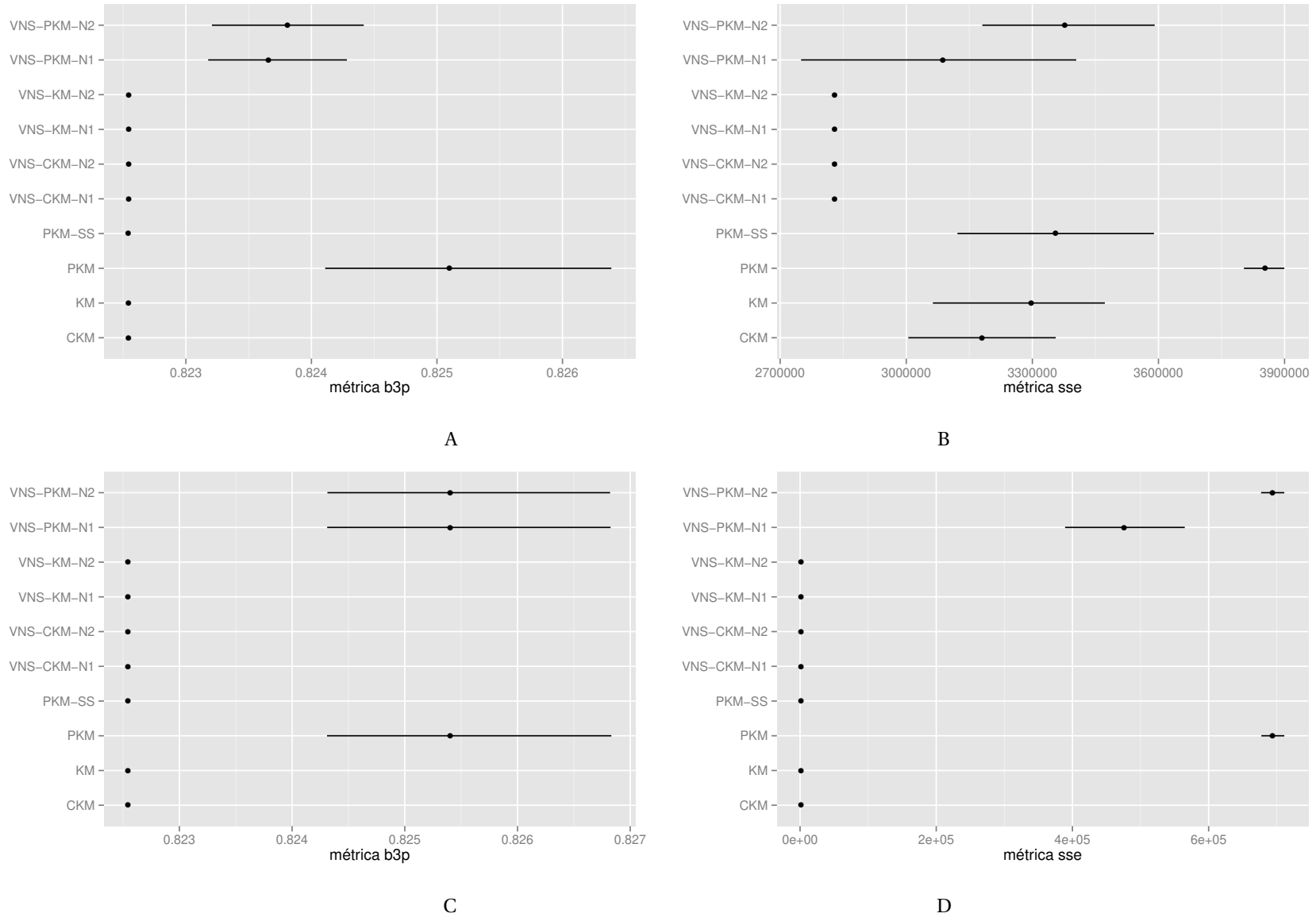


Fig. 5.12: Gráficos referente aos dados brutos de Jaboticabal (A e B) e aos dados transformados (C e D).

alcança o melhor resultado da rodada em pelo menos 90% das vezes. E para uma tolerância de menos de 0,2% do melhor resultado alcançado, nos dados transformados, as variantes VNS-PKM são indistinguíveis do algoritmo PKM. De forma objetiva, o PKM e suas variantes VNS dominam absolutamente todos os concorrentes.

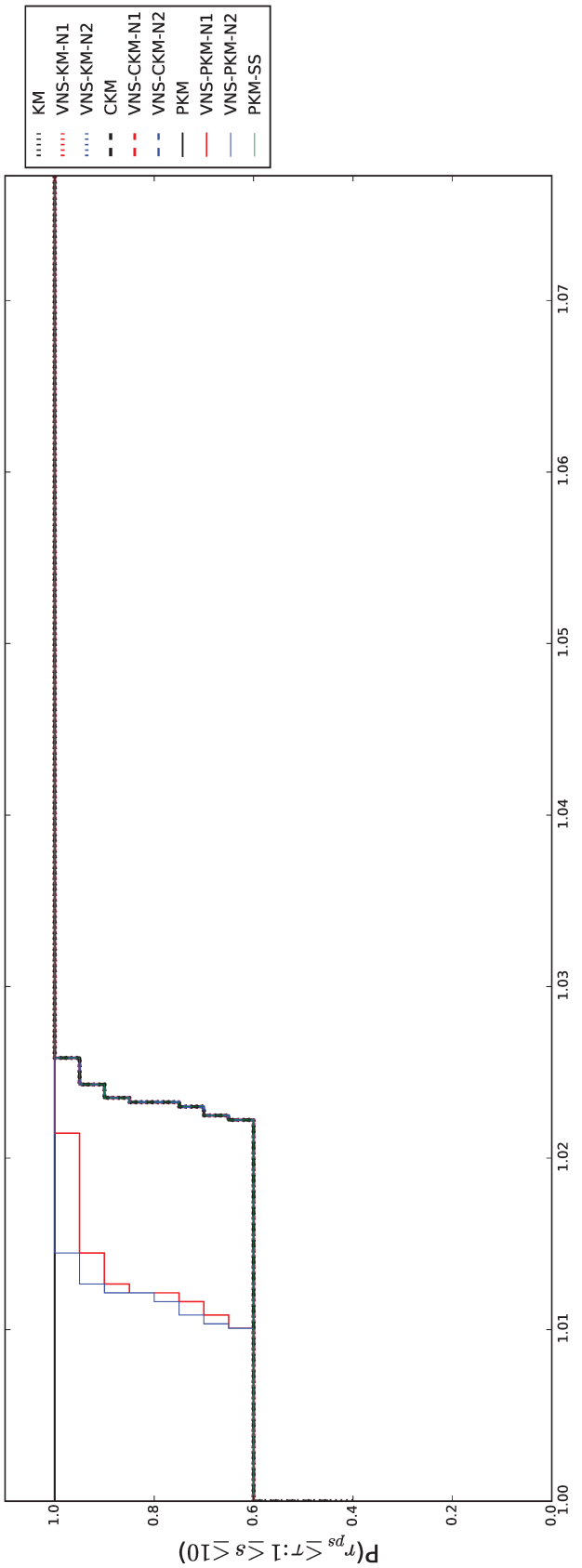


Fig. 5.13: Perfil de desempenho na base de dados brutos de Jaboticabal para uma variedade de 20 condições iniciais.

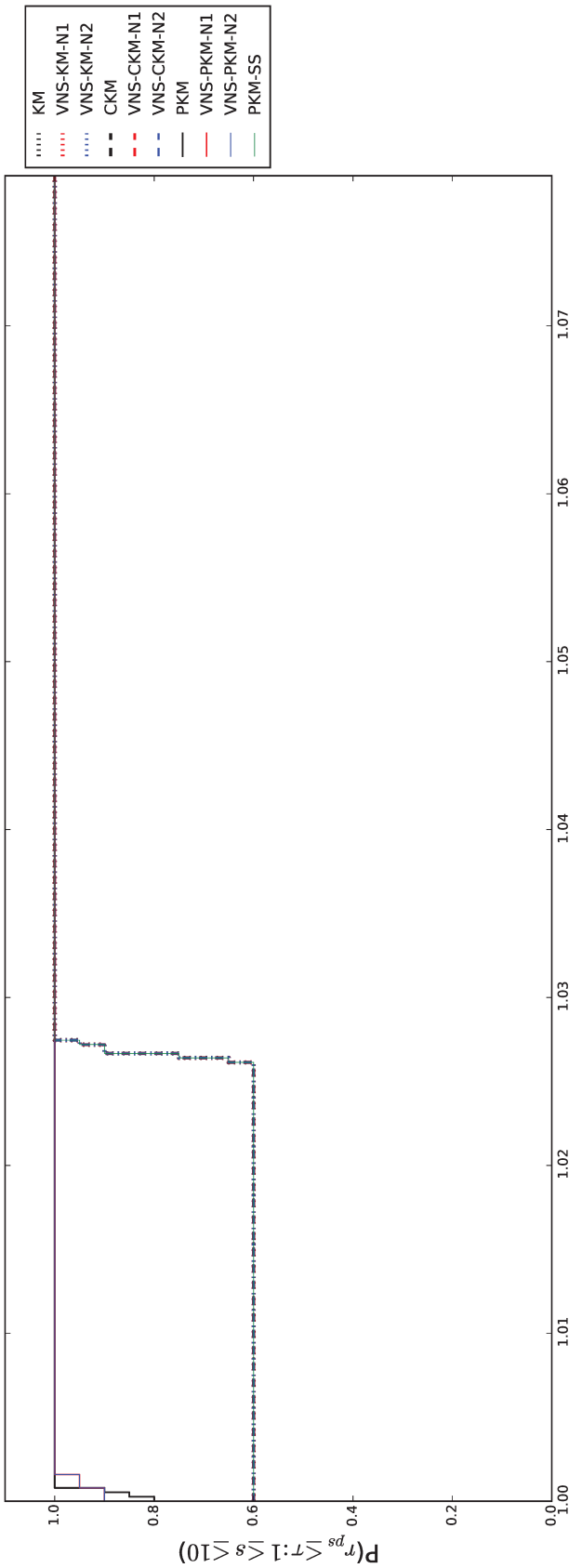


Fig. 5.14: Perfil de desempenho dos dados transformados de Jaboticabal para uma variedade de 20 condições iniciais.

5.3.4 TERRA/MODIS - Mato Grosso 2008/2009

Os dados dessa base constituem-se de séries temporais de NDVI para a região do Mato Grosso no ano safra 2008/2009, com produtos gerados a partir de séries de imagens de 16 dias do sensor TERRA/MODIS, com verdade terrestre obtida a partir de censo agrícola. A amostra analisada contém 1.991 instâncias de dimensão 23 e não possui valores faltantes. Os dados estão distribuídos em 8 classes: soja (85,53%), pasto (5,32%), algodão (3,42%), arroz (2,61%), mata (0,95%), pousio (0,80%), pasto degradado (0,75%) e milho (0,60%).

As imagens usadas foram obtidas a partir do sensor *MODIS* presente no satélite *Terra*. Sua resolução espacial varia entre 1.000m a 250m, dependendo do produto trabalhado. Para os produtos NDVI, as imagens usadas tinham resolução espacial de 250m. As imagens *MODIS* são disponibilizadas pelo LP-DAAC/NASA (*Land Processes Distributed Active Archive Center* - centro de processamento e distribuição ligado à NASA).

Pixels de controle

Os pixels de controle foram obtidos pela visitação dos locais classificados durante o ano safra, entre abril/2008 e março/2009. As visitas de campo foram feitas nas regiões agrícolas tradicionais do Mato Grosso, nos municípios de Campo Novo do Parecis, Campos de Júlio, Nova Mutum, Sapezal, Lucas do Rio Verde, Sinop e Sorriso (região central e oeste do estado).

5.3.5 Análise comparativa

As Tabelas 5.11 e 5.12 apesentam, respectivamente, os desempenhos dos algoritmos quando aplicados aos dados em sua forma bruta e transformada, enquanto as Tabelas 5.13 e 5.14 apresentam o ranqueamento dos algoritmos com base nesses desempenhos.

Tab. 5.11: Valores de média e desvio padrão dos algoritmos aplicados a base de dados brutos.

	métodos	B3P ($\times 10^{-2}$)		B3R ($\times 10^{-2}$)		B3F ($\times 10^{-2}$)		SSE ($\times 10^{2}$)	
		μ	σ	μ	σ	μ	σ	μ	σ
1	KM	77,69	0,22	21,04	0,63	33,11	0,80	9,99	0,04
2	PKM-SS	77,86	0,33	21,34	0,50	33,49	0,64	10,01	0,04
3	PKM	77,75	0,44	22,24	1,62	34,55	1,93	10,20	0,12
4	CKM	77,69	0,22	20,80	1,00	32,80	1,26	9,99	0,07
5	VNS - KM - N1	77,72	0,33	20,91	0,99	32,94	1,25	9,99	0,04
6	VNS - KM - N2	77,70	0,27	21,15	0,90	33,24	1,14	9,98	0,03
7	VNS - CKM - N1	77,68	0,22	21,23	0,75	33,34	0,93	9,98	0,07
8	VNS - CKM - N2	77,67	0,23	21,04	0,96	33,10	1,20	9,99	0,07
9	VNS - PKM - N1	77,78	0,41	21,69	1,54	33,89	1,86	10,13	0,10
10	VNS - PKM - N2	77,69	0,41	21,94	1,71	34,19	2,05	10,14	0,12

Observando as Tabelas 5.11 e 5.12, constatamos que os valores de precisão (B3P) são muito próximos entre os diversos algoritmos, indicando que, sob a perspectiva de homogeneidade

Tab. 5.12: Valores de média e desvio padrão dos algoritmos aplicados a base de dados do MT com transformação e redução de atributos para 5 dimensões.

	métodos	B3P ($\times 10^{-2}$)		B3R ($\times 10^{-2}$)		B3F ($\times 10^{-2}$)		SSE ($\times 10^{2}$)	
		μ	σ	μ	σ	μ	σ	μ	σ
1	KM	75,64	0,08	28,14	1,24	41,01	1,38	16,21	0,28
2	PKM-SS	75,59	0,13	27,90	1,68	40,73	1,88	16,18	0,18
3	PKM	75,73	0,40	49,21	11,29	58,94	8,46	29,39	3,89
4	CKM	75,63	0,07	28,10	1,21	40,96	1,36	16,15	0,14
5	VNS - KM - N1	76,12	0,01	35,30	0,01	48,24	0,01	16,04	0,00
6	VNS - KM - N2	75,63	0,06	28,60	1,64	41,48	1,63	16,15	0,17
7	VNS - CKM - N1	76,12	0,01	35,31	0,01	48,24	0,01	16,04	0,00
8	VNS - CKM - N2	75,60	0,06	27,96	1,19	40,81	1,33	16,14	0,14
9	VNS - PKM - N1	75,71	0,44	39,79	11,74	51,35	8,16	17,12	0,87
10	VNS - PKM - N2	75,70	0,34	37,26	9,54	49,27	8,20	19,96	4,06

Tab. 5.13: Lista em ordem decrescente de desempenho dos algoritmos para a base de dados de MT no formato bruto

	B3F	B3P	B3R	SSE
1	PKM	PKM-SS	PKM	VNS-KM-N2
2	VNS-PKM-N2	VNS-PKM-N1	VNS-PKM-N2	VNS-CKM-N1
3	VNS-PKM-N1	PKM	VNS-PKM-N1	VNS-CKM-N2
4	PKM-SS	VNS-KM-N1	PKM-SS	KM
5	VNS-CKM-N1	VNS-KM-N2	VNS-CKM-N1	CKM
6	VNS-KM-N2	VNS-PKM-N2	VNS-KM-N2	VNS-KM-N1
7	KM	KM	KM	PKM-SS
8	VNS-CKM-N2	CKM	VNS-CKM-N2	VNS-PKM-N1
9	VNS-KM-N1	VNS-CKM-N1	VNS-KM-N1	VNS-PKM-N2
10	CKM	VNS-CKM-N2	CKM	PKM

e preservação de grupos pequenos, os algoritmos têm resultados muito similares, tanto em dados brutos quanto transformados. Apesar disso, a **transformação** foi capaz de melhorar a lembrança (B3R), em detrimento de uma pequena perda da precisão (B3P). Isso indica que, além da significativa redução de dimensionalidade (23 para 5), em uma comparação ponderada (F de Rijsbergen), a transformação gera sensível ganho de desempenho.

Observando a Figura 5.15, constata-se que o algoritmo PKM e suas variantes VNS possuem intervalos de confiança que, mesmo em seus piores resultados, são competitivos. Entretanto, vale atentar que o algoritmo PKM e sua variantes VNS, apesar de terem os melhores resultados semissupervisionados, tiveram os piores desempenhos na redução da variação intracluster (ver gráficos na Figura 5.15). A VNS reduziu a SSE e manteve uma boa qualidade da B3F, mas, comparada aos demais algoritmos, a variação intracluster foi alta. Talvez isso sirva para indicar que, a depender da estrutura dos dados, ocorrem discrepâncias entre a função objetivo (SSE) e a

Tab. 5.14: Lista em ordem decrescente de desempenho dos algoritmos para a base de dados transformados de MT

	B3F	B3P	B3R	SSE
1	PKM	VNS-KM-N1	PKM	VNS-CKM-N1
2	VNS-PKM-N1	VNS-CKM-N1	VNS-PKM-N1	VNS-KM-N1
3	VNS-PKM-N2	PKM	VNS-PKM-N2	VNS-CKM-N2
4	VNS-CKM-N1	VNS-PKM-N1	VNS-CKM-N1	CKM
5	VNS-KM-N1	VNS-PKM-N2	VNS-KM-N1	VNS-KM-N2
6	VNS-KM-N2	KM	VNS-KM-N2	PKM-SS
7	KM	CKM	KM	KM
8	CKM	VNS-KM-N2	CKM	VNS-PKM-N1
9	VNS-CKM-N2	VNS-CKM-N2	VNS-CKM-N2	VNS-PKM-N2
10	PKM-SS	PKM-SS	PKM-SS	PKM

métrica de qualidade usada para avaliar o agrupamento. Parece adequado dizer que os métodos sem nenhuma restrição de centroide, ou sem algum critério semissupervisionado, podem buscar quedas de SSE que na verdade pioram a qualidade dos grupos.

Avaliando a performance pela Figura 5.16, notamos que o algoritmo PKM alcança o melhor resultado da rodada em 30% das vezes. Por outro lado, depois da transformação (ver Figura 5.17) esse valor aumenta para 80%. Embora o método PKM e sua variantes VNS sejam claramente superiores nas bases brutas e transformadas, a diferença entre eles e os demais algoritmos na base bruta é inexistente para tolerâncias acima de 10%. Entretanto, a superioridade na base transformada é perceptível até para tolerâncias acima de 100%.

Como o algoritmo PKM-SS tem uma performance inferior, admite-se que o desempenho superior do algoritmo PKM e suas variantes VNS está associado às restrições. Entretanto, vale observar que, mesmo com restrições, os algoritmos CKM e suas variantes VNS não foram capazes de se igualar ao algoritmo proposto pela tese.

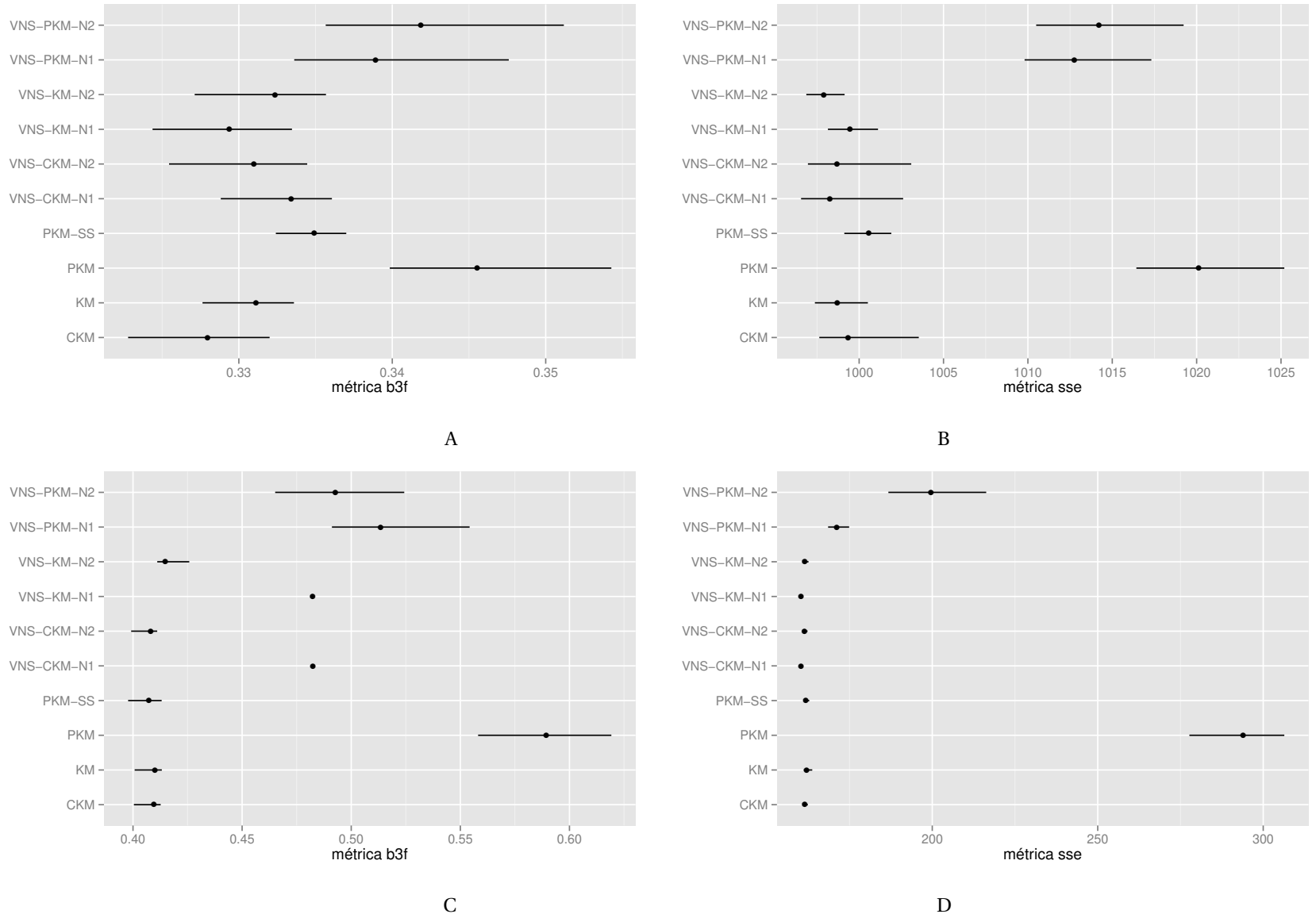


Fig. 5.15: Gráficos A e B referente aos dados brutos de MT e gráficos C e D referentes aos dados transformados.

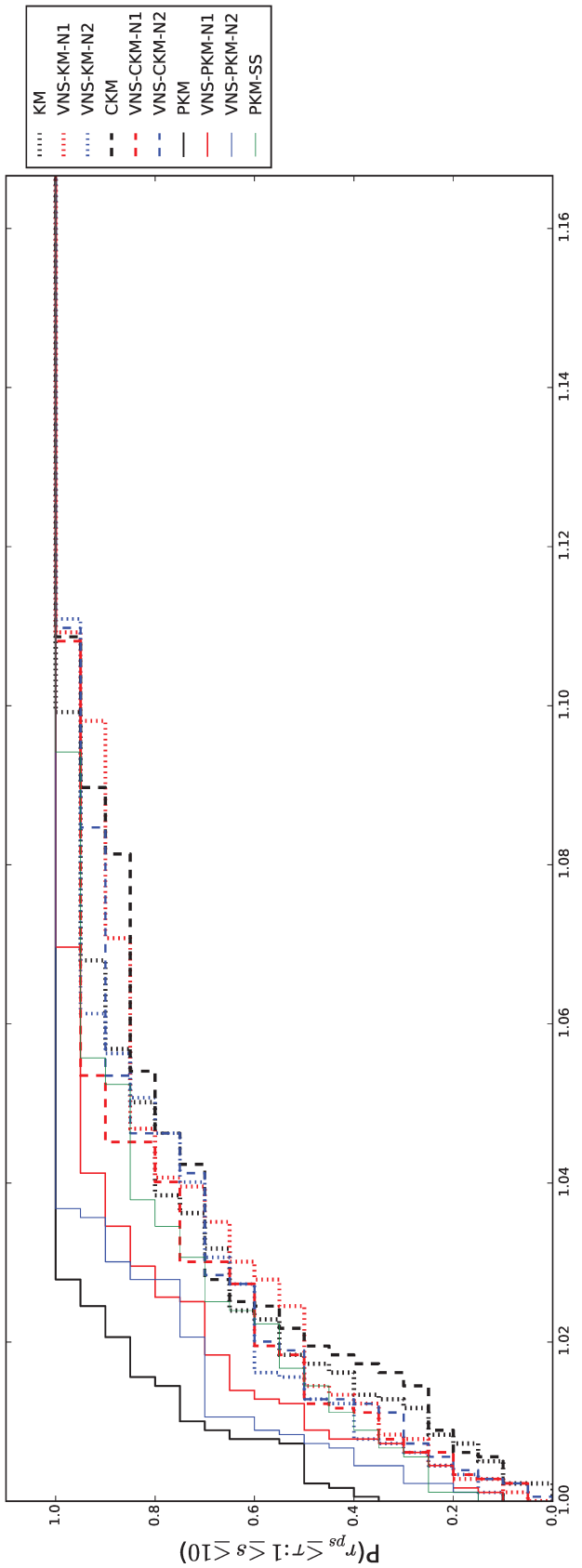


Fig. 5.16: Perfil de desempenho por *b3f* na base de dados brutos de MT em 20 condições iniciais distintas.

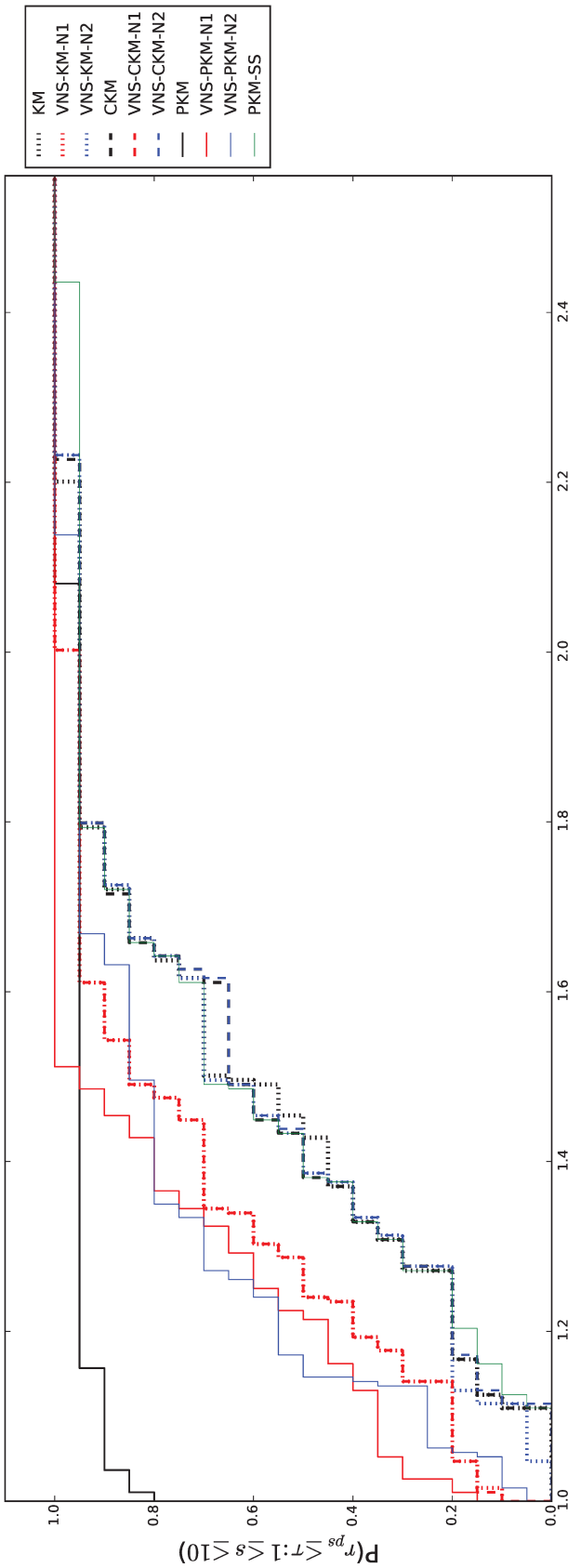


Fig. 5.17: Perfil de desempenho por *b3f* na base de dados transformados de MT em 20 condições iniciais distintas.

5.4 Análise e discussão dos resultados

Os métodos de clusterização e classificação usam, implicita ou explicitamente, alguma função critério em seus algoritmos. A SSE é uma escolha comum dentre alguns dos métodos mais clássicos da literatura, como o k-médias. E estratégias que permitam uma redução do SSE são interessantes por oferecerem uma melhora em um critério amplamente adotado. Nos experimentos realizados, às variantes VNS, quase sempre, reduziram sensivelmente a SSE, e mesmo quando essa queda não ocorreu, o valor obtido era idêntico às variantes que não fizeram uso da VNS. Isso era esperado, dada as especificidades da VNS, mas permite que se afirme que grupos obtidos por algoritmos que fazem uso da VNS serão, na pior das hipóteses, tão coesos quanto seus concorrentes e, na maioria das vezes, melhores.

Observa-se que existe uma clara superioridade dos algoritmos PKM e VNS-PKM, sobre os concorrentes em quase todas as avaliações. Isso só não ocorreu na avaliação de lembrança dos dados de Jaboticabal, mas mesmo nessa avaliação específica, a versão PKM-SS se equiparou aos concorrentes. Talvez isso sugira que, em estruturas de classes binárias, o atual método de criação de barreiras se mostra muito restritivo pois compromete a qualidade do agrupamento sobre o ponto de vista da métrica lembrança. Entretanto, vale observar que nos dados Jaboticabal sobre a perspectiva da precisão (B3P), o método PKM e suas variantes PKM-VNS foram superiores a todos os demais métodos, inclusive ao próprio PKM-SS. E considerando as idiossincrasias dos dados de Jaboticabal, a precisão é a métrica mais adequada para comparar os desempenhos dos algoritmos.

O uso de restrições é muitas vezes uma forma indireta de questionar a ideia de que quanto menor a SSE, melhor é o agrupamento. Existem métodos que implicitamente têm a SSE como função objetivo, mas não avaliam se as quedas da variação intracluster trazem verdadeira vantagem. Por conta disso, usou-se as métricas B3 para validar se, de fato, os métodos que mais reduzem a SSE são melhores e a conclusão foi que a redução da SSE nem sempre traz qualidade ao agrupamento gerado. Apesar de existir alguma correlação, houveram agrupamentos com métricas B3 semelhantes e SSE distintos, e agrupamentos com SSE muito próximos e valores de métricas B3 significamente diferentes. Em alguns testes, o método que menos reduziu a SSE por vezes gerou os agrupamentos com maior escore supervisionado. Daí, métodos sem critério de apoio, como restrições, podem muitas vezes dar a aparência de melhores resultados, sem na verdade alcançá-los e, portanto, é interessante incluir funções objetivo que incorporem restrições ou que penalize violações.

Comparar os métodos PKM e PKM-SS serve para intuirmos as prováveis diferenças entre usar caixas de factibilidade restritas e caixas de factibilidade flexíveis. Dessa forma, refletimos sobre as vantagens e desvantagens de se incorporar uma barreira rígida, como a adotada, e as possibilidades existentes em versões mais flexíveis, que possam, por exemplo, permitir centroides fora da região de factibilidade com alguma penalização. De certo, ainda cabe muito debate sobre como construir uma região de factibilidade para centroides. Mesmo no modelo atual, cabe debater como mudar a forma e tamanho da região de factibilidade para testar hipóteses que levem em conta a qualidade e a quantidade do conjunto de exemplos. Entretanto, pelos resultados apresentados, acreditamos que a forma e o tamanho usados se mostraram adequados.

A meta-heurística VNS, além da redução da variação intracluster, em especial para a estrutura de vizinhanças do tipo 1, gerou, em vários testes, benefícios nas métricas supervisionadas. Observou-se que as variantes VNS são benéficas, não só como uma forma de redução do SSE,

mas como um método de busca por soluções melhores em relação aos índices extrínsecos.

Acrescido a isso, observou-se que os métodos semissupervisionados nem sempre são melhores que as buscas não supervisionadas, em particular, os métodos CKM e VNS-CKM foram superados em vários testes pelos métodos KM e VNS-KM. Isso contrapõe a ideia intuitiva de que incorporar informação é sempre benéfico à busca. A maneira como se incorpora conhecimento prévio, escolhendo entre relações de paridades ou restrições por caixas, afeta distintamente o algoritmo de agrupamento e a qualidade de seus resultados.

O descritor $< \mathscr{F}, L_2 >$ foi proposto para lidar com as imagens AVHRR/NOAA e TERRA/MODIS, com o propósito de tornar estruturas de valores sequenciais em t-uplas de valores que descrevem a curva de interpolação dos dados em sua forma bruta. A transformação do espaço de busca mostrou-se eficiente para a classificação de diversas coberturas pois a aplicação dos algoritmos sobre os dados transformados criou agrupamentos de qualidades equivalentes, e por vezes melhores, com significativa redução dimensional.

Vale observar que, apesar de apresentarmos uma avaliação comparativa dos algoritmos em todas as bases testadas, a base MT é a mais indicada para o domínio específico de séries temporais agrícolas, por se tratar de uma base multiclasse, com desbalanceamento, onde as séries se referem a culturas misturadas na área investigada, características frequentes em dados de cobertura agrícola. E foi exatamente na base MT que o algoritmo proposto se mostrou significativamente superior a seus concorrentes CKM e KM. Os resultados sugerem que o uso combinado do descritor $< \mathscr{F}, L_2 >$ com o algoritmo PKM é a melhor escolha para gerar agrupamento para dados de imagens agrícolas. Além disso, as variantes VNS do PKM são úteis para estabelecer bons agrupamentos com maior coesividade.

Capítulo 6

Conclusões e trabalhos futuros

"You never fail until you stop trying."
- Albert Einstien

É importante que se diga que as hipóteses inicialmente apresentadas na seção 1.3 foram validadas. O algoritmo proposto é uma variante da heurística k-médias (*k-means*) que faz uso de restrições e pode ser combinado à meta-heurística de busca em vizinhança variável (VNS). Esse método, como descrito na hipótese inicial, permitiu a fuga de mínimos locais e tornou a solução final menos dependente dos pontos iniciais, de maneira que fossem encontradas soluções melhores que as encontradas pelos métodos concorrentes. Isso se deu tanto pela busca em um espaço reduzido (espaço de factibilidade) criado pelas pelas restrições, quanto pela maneira como a VNS faz a exploração desse espaço. Acrescido a isso, a extração de características, que consistiu no uso dos coeficientes da transformada discreta de Fourier da séries temporais, gerou um conjunto de dados transformados que, quando submetido aos principais métodos de agrupamento da literatura e ao método proposto, permitiu que se encontrasse agrupamentos melhores, como era esperado pela segunda hipótese da seção 1.3.

As aplicações interdisciplinares geram visibilidade a teorias formais e conferem credibilidade a áreas consideradas básicas como a matemática. Especificamente na área de agrupamento, o uso de métodos de otimização permite uma visão alternativa de como abordar os problemas e oferece uma gama de métodos, já testados em outros escopos, como a busca de soluções com restrições.

Outros motivos para se interessar em agrupamentos com restrições decorrem de serem técnicas que fazem uso da informação disponível, diferentemente dos métodos usuais de clusterização e classificação. Os métodos de clusterização não fazem uso algum dos dados disponíveis, enquanto métodos de classificação só utilizam um conjunto de dados se ele for qualificado como um conjunto de treinamento. Os métodos propostos, ao contrário da classificação supervisionada, fazem uso de qualquer informação disponível, sem a necessidade de avaliar os dados disponíveis. Como frequentemente é complicado atestar a qualidade de um conjunto de treinamento, estratégias que contornem essa dificuldade, sem perder os benefícios do uso do conhecimento prévio, são desejáveis. Portanto, métodos de agrupamento semissupervisionado são estratégias claramente melhores do que modelos de classificação não supervisionada e podem ser estratégias competitivas, inclusive contra modelos de classificação supervisionada.

A transformação proposta converte dados dinâmicos em estáticos, o que previne a busca por estratégias de agrupamentos especializadas em séries temporais. Ao substituir os valores brutos das séries temporais pelos coeficientes dos harmônicos da curva interpoladora, na verdade estamos trocando os valores brutos por quantidades que caracterizam a curva que se ajusta a estes dados. Essa abordagem, nos dados testados, permitiu grupos de qualidade superior ou equivalente, com uma redução significativa da dimensionalidade. Provavelmente, isso ocorreu pela remoção dos termos referentes a harmônicos de maior frequência, que são altamente relacionados a ruídos. Por fim, nota-se que a transformação é uma alternativa à aplicação direta dos algoritmos sobre os dados brutos, pois atenua o efeito de Hughes. Além disso, nos dados agrícolas, o uso combinado da transformação com métodos de agrupamentos com restrições ofereceu desempenho superior a seus concorrentes e acreditamos que esta estratégia também tenha vantagens sobre classificações baseadas em máscaras, pois é capaz de detectar culturas de expansão.

A meta-heurística VNS básica causou sensíveis melhoras na variação intracluster, mas, como não estabelecemos uma coerência entre a SSE e as métricas B3, é preciso adaptar a VNS para que os agrupamentos por ela gerados tenham um desempenho superior, independentemente da métrica usada. Além disso, é importante enfatizar que a VNS é um método de diversificação que permitiu obter agrupamentos mais coesos do que aqueles obtidos sem o seu uso. Seria interessante investigar o efeito de outras estruturas de vizinhança e outros esquemas de perturbação. É possível, por exemplo, usar estratégias de dados multidimensionais baseadas em densidade para criar estruturas de vizinhança, bem como fazer a pertubação selecionando aleatoriamente um elemento x' e separando algumas características y que o distinguam de x, de modo que a busca local fosse feita apenas no espaço y (VNS decomposto).

Trabalhos futuros

Existem múltiplos aspectos dos algoritmos e da transformação propostos que podem ser revisados para tentar aumentar seu desempenho. Além do mais, uma característica do trabalho multidisciplinar é a possibilidade de se empreender estratégias e ênfases distintas sobre os diversos tópicos que envolvem o problema de pesquisa. Disso, compilou-se uma lista de sugestões dos principais pontos que permitem uma continuação do trabalho baseada em considerações a respeito dos métodos propostos.

Escolha de parâmetros. Existem dois parâmetros implícitos para a calibração dos algoritmos propostos: o tamanho, ou forma, da caixa de restrição e o número máximo de vizinhanças. As escolhas apresentadas para estes parâmetros foram feitas de forma empírica, a partir da experimentação de várias possibilidades. Essa abordagem pode ter gerado valores subótimos e uma análise mais detalhada talvez revele escolhas que melhorem o desempenho do método.

Esquema de restrições. A região de factibilidade da heurística PKM é uma caixa no espaço de atributos, que restringe o deslocamento dos centroides dentro delas. Essa maneira de controlar os centroides é uma abordagem rígida, que poderia ter sua eficiência comparada a uma abordagem que flexibilizasse as barreiras, como permitir qualquer deslocamento do centroide, mas penalizando movimentos de centroides que violem as restrições.

Escolha de d na transformação $\mathscr{F}$. A transformação pode gerar um conjunto de dados dimensionalmente compatível com os dados originais. Entretanto, dimensões maiores não significam necessariamente ganho de informação, havendo uma dimensão ótima de performance, que foi experimentalmente definida como um quarto da dimensão original ($\lceil \frac{d}{4} \rceil$). Esse critério é uma estratégia simplista, que pode ser substituída por métodos que levem em conta as cararacterísticas dos dados transformados.

Comparação entre paradigmas. Os classificadores usam uma intervenção supervisionada para induzir um critério de discriminação e essa abordagem é intuitivamente mais efetiva do que contar apenas com características intrínsecas dos dados para gerar separabilidade que é a abordagem dos clusterizadores. A hipótese implícita dos métodos de agrupamentos é que as características de distribuição são suficientes para definir os grupos. Entretanto, o agrupamento com restrições (semissupervisionado) é uma estratégia híbrida, que faz uso de conjuntos classificados. Isso permite comparar os grupos obtidos pela clusterização semissupervisionada e pela classificação supervisionada e não existe na literatura uma comparação de desempenhos entre esses paradigmas.

Abordagens *fuzzy*. A abordagem *crisp* pode ser substituída por uma abordagem que permita definir pertinências intermediárias do elemento à classe. Nos dados agrícolas, essa forma de encarar a pertinência é um meio de lidar com problemas de mistura espectral.

Transformação por outras bases. A transformação proposta consiste na substituição dos dados brutos pelos coeficientes do polinômio trigonométrico que melhor interpola a série temporal. Entretanto, o uso de uma base trigonométrica não é a única opção. Pode-se testar os efeitos destas transformações em bases como *wavelets e curvelets*.

Escolha de índices vegetativos. *O NDVI é uma escolha frequente na literatura. Entretanto, nada nos impede de avaliar a capacidade discriminatória de outros índices de vegetação, como o NDMI (Normalized Difference Moisture Index). O NDMI usa a faixa do infravermelho médio, que é menos afetada pelos aerossóis e vapor de água presentes na atmosfera e pode indicar quando há umidade na vegetação e no solo, o que permite uma maior diferenciação entre os tipos de vegetação. Essas características tornam o NDMI melhor que o NDVI para reconhecer mudanças de biomassa.*

Contexto espacial. *A atual abordagem desconsidera a localização do pixel. Pode-se investigar os benefícios de agrupar os pixels baseando-se na segmentação das imagens. A segmentação é uma tarefa básica no processo de análise de imagens, na qual a imagem é particionada em regiões que devem corresponder às áreas de interesse da aplicação. Entende-se por regiões um conjunto de pixels contíguos que se espalham bidimensionalmente e que apresentam uniformidade em relação a algum atributo.*

Referências Bibliográficas

[1] EM Abdel-Rahman e FB Ahmed, *The application of remote sensing techniques to sugarcane (Saccharum spp. hybrid) production: a review of the literature*, International Journal of Remote Sensing **29** (2008), nº 13, 3753–3767.

[2] SR Ahmed, *Applications of data mining in retail business*, Proceedings of the International Conference on Information Technology: Coding Computing, ITCC **2** (2004), 455–459.

[3] A Alguwaizani, *Variable neighbourhood search based heuristic for K-harmonic means clustering*, Tese de Doutoramento, Brunel University, School of Information Systems, Computing and Mathematics, 2011.

[4] MC Alonso, JA Malpica e AM de Agirre, *Consequences of the Hughes phenomenon on some classification Techniques*, Proceedings of the Annual Conference (ASPRS 2011) (2011), 1–5.

[5] E Amigó, J Gonzalo, J Artiles e F Verdejo, *A comparison of extrinsic clustering evaluation metrics based on formal constraints*, Information retrieval **12** (2009), nº 4, 461–486.

[6] M Ankerst, MM Breunig, HP Kriegel e J Sander, *OPTICS: ordering points to identify the clustering structure*, ACM Sigmod record **28** (1999), nº 2, 49–60.

[7] JFG Antunes, *Aplicação de lógica fuzzy para estimativa de área plantada da cultura de soja utilizando imagens AVHRR-NOAA*, Tese de Mestrado, Universidade Estadual de Campinas, 2005.

[8] JFG Antunes e JCDM Esquerdo, *NAVPRO 3.0: tutorial de instalação e utilização.*, Embrapa Informática Agropecuária-Documentos (INFOTECA-E) (2008).

[9] _____, *Monitoramento agrícola usando análise harmônica de séries temporais de dados NDVI/AVHRR-NOAA.*, Anais XIV Simpósio Brasileiro de Sensoriamento Remoto (2009), 49–55.

[10] JFG Antunes, JCDM Esquerdo e J Zullo Jr, *Desenvolvimento de um sistema automático para a geração de produtos derivados de imagens AVHRR-NOAA*, Anais XII Simpósio Brasileiro de Sensoriamento Remoto (2005).

[11] A Bagga e B Baldwin, *Entity-based cross-document coreferencing using the vector space model*, Proceedings of the 17th international conference on Computational linguistics-Volume 1 (1998), 79–85.

[12] J Bakus, MF Hussin e M Kamel, *A SOM-based document clustering using phrases*, Proceedings of the 9th International Conference on Neural Information (ICONIP'02) **5** (2002), 2212–2216.

[13] GH Ball e DJ Hall, *ISODATA, a novel method of data analysis and pattern classification*, Rel. Téc., Stanford research inst Menlo Park CA, 1965.

[14] _____, *A clustering technique for summarizing multivariate data*, Systems Research and Behavioral Science **12** (1967), nº 2, 153–155.

[15] A Bar-Hillel, T Hertz, N Shental e D Weinshall, *Learning a mahalanobis metric from equivalence constraints*, Journal of Machine Learning Research **6 (jun)** (2005), 937–965.

[16] S Basu, A Banerjee e R Mooney, *Semi-supervised clustering by seeding*, Proceedings of 19th International Conference on Machine Learning (ICML-2002) (2002).

[17] N Belacel, P Hansen e N Mladenovic, *Fuzzy j-Means: a new heuristic for fuzzy clustering*, Pattern Recognition **35** (2002), nº 10, 2193–2200, ISSN 00313203, `http://linkinghub.elsevier.com/retrieve/pii/S0031320301001935`.

[18] M Bilenko, S Basu e RJ Mooney, *Integrating constraints and metric learning in semi-supervised clustering*, Proceedings of the twenty-first international conference on Machine learning (2004), 11.

[19] A Blum e T Mitchell, *Combining labeled and unlabeled data with co-training*, Proceedings of the eleventh annual conference on Computational learning theory (1998), 92–100.

[20] C Blum e A Roli, *Metaheuristics in combinatorial optimization: Overview and conceptual comparison*, ACM computing surveys (CSUR) **35** (2003), nº 3, 268–308.

[21] E Boros, PL Hammer, T Ibaraki, A Kogan, E Mayoraz e I Muchnik, *An implementation of logical analysis of data*, IEEE Transactions on knowledge and Data Engineering **12** (2000), nº 2, 292–306.

[22] A Bowman, *Functional Data Analysis with R and MATLAB*, Journal of Statistical Software **34** (2010), nº 1, 1–2.

[23] PS Bradley, KP Bennett e A Demiriz, *Constrained k-means clustering*, Microsoft Research, Redmond (2000), 1–8.

[24] PS Bradley e UM Fayyad, *Refining Initial Points for K-Means Clustering.*, International Conference on Machine Learning (ICML) **98** (1998), 91–99.

[25] EJ Bredensteiner e KP Bennett, *Multicategory classification by support vector machines*, Pang JS. (eds) Computational Optimization. Springer, Boston, MA (1999), 53–79.

[26] MM Breunig, HP Kriegel, RT Ng e J Sander, *LOF: identifying density-based local outliers*, ACM sigmod record **29** (2000), nº 2, 93–104.

[27] P Brucker, *On the complexity of clustering problems*, Optimization and operations research **157** (1978), 45–54, http://link.springer.com/chapter/10.1007/978-3-642-95322-4_5.

[28] Natural Resourses Canada, *Satellite Imagery and Air Photos*, http://www.nrcan.gc.ca/earth-sciences/geomatics/satellite-imagery-air-photos/satellite-imagery-products/educational-resources/9295.

[29] CEPEA, *PIB do Agronegócio*, 2013.

[30] AP Crosta, *Processamento digital de imagens de sensoriamento remoto*, UNICAMP/Instituto de Geociências, 1999.

[31] S Das, A Abraham e A Konar, *Metaheuristic Clustering*, vol. 1, Springer, março 2009, ISBN 978-35-4092-172-1.

[32] AP Dempster, NM Laird e DB Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the royal statistical society. Series B (methodological) (1977), 1–38.

[33] BE Dom, *An information-theoretic external cluster-validity measure*, Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence (2002), 137–145.

[34] M Dorigo, *Optimization, learning and natural algorithms*, Tese de Doutoramento, Ph. D. Thesis, Politecnico di Milano, Italy, 1992, http://ci.nii.ac.jp/naid/10016599043/.

[35] O du Merle, P Hansen, B Jaumard e N Mladenovic, *An interior point algorithm for minimum sum-of-squares clustering*, SIAM Journal on Scientific Computing **21** (1999), nº 4, 1485–1505.

[36] JCDM Esquerdo, JFG Antunes, DG Baldwin, WJ Emery e JZ Júnior, *An automatic system for AVHRR land surface product generation*, International Journal of Remote Sensing **27** (2006), nº 18, 3925–3942, http://www.tandfonline.com/doi/abs/10.1080/01431160600763956.

[37] M Ester, H P Kriegel, J Sander e X Xu, *A density-based algorithm for discovering clusters in large spatial databases with noise.*, Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96) **34** (1996), 226–231.

[38] V Estivill-Castro, *Why so many clustering algorithms: a position paper*, ACM SIGKDD Explorations Newsletter **4** (2002), 65–75, http://dl.acm.org/citation.cfm?id=568575.

[39] G Felici e K Truemper, *A minsat approach for learning in logic domains*, INFORMS Journal on computing **14** (2002), nº 1, 20–36.

[40] TA Feo e MGC Resende, *Greedy randomized adaptive search procedures*, Journal of global optimization **42** (1995), 32–37, http://link.springer.com/article/10.1007/BF01096763.

[41] LMG Fonseca, *Processamento digital de imagens*, Instituto Nacional de Pesquisas Espaciais (INPE) (2000).

[42] A Gammerman, V Vovk e V Vapnik, *Learning by transduction*, Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence (1998), 148–155.

[43] M Gendreau e J Y Potvin, *Handbook of metaheuristics*, vol. 2, Springer, 2010.

[44] S Ghosh e SK Dubey, *Comparative analysis of k-means and fuzzy c-means algorithms*, International Journal of Advanced Computer Science and Applications **4** (2013), nº 4, 35–39.

[45] F Glover, *Future paths for integer programming and links to artificial intelligence*, Computers & Operations Research **13** (1986), nº 5, 533–549, http://www.sciencedirect.com/science/article/pii/0305054886900481.

[46] _____, *Improved Linear Programming Models for Discriminant Analysis**, Decision Sciences **21** (1990), 771–785, http://onlinelibrary.wiley.com/doi/10.1111/j.1540-5915.1990.tb01249.x/full.

[47] C Goutte, P Toft, E Rostrup, F Å Nielsen e LK Hansen, *On clustering fMRI time series*, NeuroImage **9** (1999), nº 3, 298–310.

[48] I Guyon e A Elisseeff, *An introduction to feature extraction*, Feature extraction, Springer, 2006, pp. 1–25.

[49] G Hamerly e C Elkan, *Alternatives to the k-means algorithm that find better clusterings*, Proceedings of the eleventh international conference on Information and knowledge management (2002), 600–607.

[50] J Han, J Pei e M Kamber, *Data mining: concepts and techniques*, Elsevier, 2011.

[51] P Hansen e B Jaumard, *Cluster analysis and mathematical programming*, Mathematical Programming **79** (1997), nº 1-3, 191–215, ISSN 0025-5610, http://link.springer.com/10.1007/BF02614317.

[52] P Hansen e N Mladenović, *J-means: a new local search heuristic for minimum sum of squares clustering*, Pattern recognition **34** (2001), nº 2, 405–413.

[53] S Harms, D Li, J Deogun e T Tadesse, *Efficient rule discovery in a geo-spatial decision support system*, Proceedings of the 2002 annual national conference on digital government research (2002), 1–7.

[54] RJ Hathaway e JC Bezdek, *Optimization of clustering criteria by reformulation*, IEEE transactions on Fuzzy Systems **3** (1995), nº 2, 241–245.

[55] M J Hill e G E Donald, *Estimating spatio-temporal patterns of agricultural productivity in fragmented landscapes using AVHRR NDVI time series*, Remote Sensing of Environment **84** (2003), nº 3, 367–384.

[56] A Hinneburg e DA Keim, *An efficient approach to clustering in large multimedia databases with noise*, Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'98) **98** (1998), 58–65.

[57] BN Holben, *Characteristics of maximum-value composite images from temporal AVHRR data*, International journal of remote sensing **7** (1986), nº 11, 1417–1434.

[58] BN Holben, CJ Tucker e CJ Fan, *Spectral assessment of soybean leaf area and leaf biomass.*, Photogrammetric Engineering and Remote Sensing **46** (1980), nº 5, 651–656.

[59] JH Holland, *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*, MIT press, 1992.

[60] R Howard, *Classifying a population into homogeneous groups*, Operational Research in Social Science, Tavistock Publ., London (1966).

[61] IBGE, *Levantamento Sistemático da Produção Agrícola*, 2013, `http://www.ibge.gov.br/home/estatistica/indicadores/agropecuaria/lspa/`.

[62] A Inselberg, *Multidimensional detective*, , 1997. Proceedings of IEEE Symposium on Information Visualization (1997), 100–107.

[63] GAl Ippoliti-Ramilo, JCN Epiphanio, Y E Shimabukuro e AR Formaggio, *Sensoriamento remoto orbital como meio auxiliar na previsão de safras*, Agricultura em São Paulo **46** (1999), 89–101.

[64] ME Jakubauskas, DR Legates e JH Kastens, *Harmonic analysis of time-series AVHRR NDVI data*, Photogrammetric engineering and remote sensing **67** (2001), nº 4, 461–470.

[65] _____, *Crop identification using harmonic analysis of time-series AVHRR NDVI data*, Computers and Electronics in Agriculture **37** (2002), nº 1-3, 127–139, ISSN 01681699, `http://linkinghub.elsevier.com/retrieve/pii/S0168169902001163`.

[66] RC Jancey, *Multidimensional group analysis*, Australian Journal of Botany **14** (1966), 127–130, `http://www.publish.csiro.au/?act=view_file&file_id=BT9660127.pdf`.

[67] P Jönsson e L Eklund, *TIMESAT - a program for analyzing time-series of satellite sensor data*, Computers & Geosciences **30** (2004), nº 8, 833–845.

[68] CF Jordan, *Derivation of leaf-area index from quality of light on the forest floor*, Ecology **50** (1969), nº 4, 663–666, `http://www.jstor.org/stable/10.2307/1936256`.

[69] Y Kakizawa, RH Shumway e M Taniguchi, *Discrimination and clustering for multivariate time series*, Journal of the American Statistical Association **93** (1998), nº 441, 328–340.

[70] G Karypis, EH Han e V Kumar, *Chameleon: Hierarchical clustering using dynamic modeling*, Computer **32** (1999), nº 8, 68–75.

[71] L Kaufman e P Rousseeuw, *Clustering by means of medoids*, North-Holland, 1987.

[72] S Kirkpatrick, DG Jr. e MP Vecchi, *Optimization by simulated annealing*, science **220** (1983), nº 4598, 671–680.

[73] F Klawonn, R Kruse e R Winkler, *Fuzzy clustering: More than just fuzzification*, Fuzzy sets and systems **281** (2015), 272–279.

[74] EM Knox e RT Ng, *Algorithms for mining distancebased outliers in large datasets*, Proceedings of the International Conference on Very Large Data Bases (1998), 392–403.

[75] FN Koumboulis, MP Tzamtzi e M Pavlovic, *Decision support systems in agribusiness*, Proceedings of 2006 IEEE International Conference on Mechatronics (2006), 457–461.

[76] M Kumar, NR Patel e J Woo, *Clustering seasonality patterns in the presence of errors*, Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (2002), 557–563.

[77] S Lloyd, *Least squares quantization in PCM*, IEEE transactions on information theory **28** (1982), nº 2, 129–137.

[78] UV Luxburg, *A tutorial on spectral clustering*, Statistics and computing **17** (2007), nº 4, 395–416, http://link.springer.com/article/10.1007/s11222-007-9033-z.

[79] J MacQueen, *Some methods for classification and analysis of multivariate observations*, Proceedings of the fifth Berkeley symposium on mathematical statistics and probability **1** (1967), 281–297.

[80] EA Maharaj, *Cluster of time series*, Journal of Classification **17** (2000), nº 2, 297–314.

[81] OL Mangasarian, *Linear and nonlinear separation of patterns by linear programming*, Operations research **13** (1965), 444–452, http://or.journal.informs.org/content/13/3/444.short.

[82] _____, *Misclassification minimization*, Journal of Global Optimization **5** (1994), nº 4, 309–323.

[83] M Meilua, *Comparing clusterings by the variation of information*, Learning theory and kernel machines, Lecture Notes in Computer Science **2777** (2003), 173–187.

[84] PR Meneses e JS Madeira Netto, *Sensoriamento remoto: reflectância dos alvos naturais*, Editora Universidade de Brasilia; Planaltina: Embrapa Cerrados, 2001.

[85] N Mladenovic e P Hansen, *Variable neighborhood search*, Computers & Operations Research **24** (1997), 1097–1100, http://www.sciencedirect.com/science/article/pii/S0305054897000312.

[86] CS Möller-Levet, F Klawonn, K H Cho e O Wolkenhauer, *Fuzzy clustering of short time-series and unevenly distributed sampling points*, Proceedings of the International Symposium on Intelligent Data Analysis (2003), 330–340.

[87] R Moustafa e EJ Wegman, *On some generalizations of parallel coordinate plots*, Seeing a Million–A Data Visualization Workshop **2** (2002), nº 4, 41–48, http://herakles.zcu.cz/seminars/docs/infovis/papers/Moustafa_generalized_parallel_coordinates.pdf.

[88] H Mühlenbein e G Paass, *From Recombination of Genes to the Estimation of Distributions I. Binary Parameters*, Proceedings of the 4th International Conference on Parallel Problem Solving from Nature (1996), 178–187, http://dl.acm.org/citation.cfm?id=645823.670694.

[89] CR Nascimento, *Utilização de Séries Temporais de Imagens AVHRR/NOAA no Apoio à Estimativa Operacional da Produção da Cana-de-Açúcar no Estado de São Paulo*, Tese de Doutoramento, Unicamp - Universidade Estadual de Campinas, 2010.

[90] K Nigam, A McCallum e T Mitchell, *Semi-supervised text classification using EM*, Proceedings of the Parnec Group Seminars (2006), 33–56.

[91] S Olafsson, X Li e S Wu, *Operations research and data mining*, European Journal of Operational Research **187** (2008), nº 3, 1429–1448, ISSN 03772217, http://dx.doi.org/10.1016/j.ejor.2006.09.023http://linkinghub.elsevier.com/retrieve/pii/S037722170600854X.

[92] D Piccolo, *A distance measure for classifying ARIMA models*, Journal of Time Series Analysis (1990), 153–163, http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9892.1990.tb00048.x/abstract.

[93] J Puchinger e GR Raidl, *Combining metaheuristics and exact algorithms in combinatorial optimization: A survey and classification*, Proceedings of the International Work-Conference on the Interplay Between Natural and Artificial Computation (2005), 41–53.

[94] I Rechenberg, *Cybernetic solution path of an experimental problem*, Ministry of Aviation, Royal Aircraft Establishment (1965), http://www.citeulike.org/group/1662/article/1505176.

[95] GJ Roerink, M Menenti e W Verhoef, *Reconstructing cloudfree NDVI composites using Fourier analysis of time series*, International Journal of Remote Sensing **21** (2000), nº 9, 1911–1917, ISSN 0143-1161, http://www.tandfonline.com/doi/abs/10.1080/014311600209814.

[96] LAS Romani, RRV Gonçalves, J Zullo Jr, C Traina e AJM Traina, *New DTW-based method to similarity search in sugar cane regions represented by climate and remote sensing time series*, Proceedings of the 2010 IEEE International Geoscience and Remote Sensing Symposium (IGARSS) (2010), 355–358.

[97] A Rosenberg e J Hirschberg, *V-measure: A conditional entropy-based external cluster evaluation measure*, Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL) **7** (2007), 410–420.

[98] JW Rouse Jr, RH Haas, JA Schell e DW Deering, *Monitoring vegetation systems in the Great Plains with ERTS*, NASA. Goddard Space Flight Center 3d ERTS-1 Symp. **1, Sect. A** (1974), 309–317.

[99] PJ Rousseeuw, *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis*, Journal of computational and applied mathematics **20** (1987), 53–65.

[100] T Sasao e M Fujita (eds.), *Representations of Discrete Functions*, Springer US, 1996, ISBN 978-1-4612-8599-1, http://www.springerlink.com/index/10.1007/978-1-4613-1385-4.

[101] RA Schowengerdt, *Techniques for image processing and classification in remote sensing*, Academic Press, 1983, http://www.slac.stanford.edu/spires/find/books/www?key=137618.

[102] P Senin, *Dynamic time warping algorithm review*, Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA (2008), 1–23.

[103] CT Shaw e GP King, *Using cluster analysis to classify time series*, Physica D: Nonlinear Phenomena **58** (1992), nº 1-4, 288–298.

[104] H Spath, *Cluster analysis algorithms for data reduction and classification of objects*, Horwood, 1980, http://www.slac.stanford.edu/spires/find/books?irn=108227.

[105] Michael Steinbach, George Karypis, Vipin Kumar et al., *A comparison of document clustering techniques*, Proceedings of the KDD workshop on text mining **400** (2000), nº 1, 525–526.

[106] WN Street, *Oblique multicategory decision trees using nonlinear programming*, INFORMS Journal on Computing **17** (2005), nº 1, 25–31.

[107] PN Tan, *Introduction To Data Mining*, Pearson Education, 2006, ISBN 9788131714720.

[108] V Černý, *Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm*, Journal of optimization theory and applications **45** (1985), 41–51, http://link.springer.com/article/10.1007/BF00940812.

[109] HD Vinod, *Integer Programming and the Theory of Grouping*, Journal of the American Statistical Association **64** (1969), nº 326, 506–519, http://www.tandfonline.com/doi/abs/10.1080/01621459.1969.10500990.

[110] M Vlachos, J Lin, E Keogh e D Gunopulos, *A wavelet-based anytime algorithm for k-means clustering of time series*, Proceedings of the Workshop on Clustering High Dimensionality Data and Its Applications (2003), 23–30.

[111] K Wagstaff, C Cardie, S Rogers e S Schrödl, *Constrained k-means clustering with background knowledge*, Proceedings of 18th International Conference on Machine Learning (ICML-2001) **1** (2001), 577–584.

[112] J Wang, M Li, J Chen e Y Pan, *A fast hierarchical clustering algorithm for functional modules discovery in protein interaction networks*, IEEE/ACM Transactions on Computational Biology and Bioinformatics **8** (2011), nº 3, 607–620.

[113] SQ Wang e DM Zhu, *Research on selecting initial points for k-means clustering*, Machine Learning and Cybernetics **5** (2008), 2673–2677, http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4620860.

[114] T Warrenliao e T. Warren Liao, *Clustering of time series data - a survey*, Pattern Recognition **38** (2005), nº 11, 1857–1874, ISSN 00313203, http://linkinghub.elsevier.com/retrieve/pii/S0031320305001305.

[115] J Wilpon e L Rabiner, *A modified K-means clustering algorithm for use in isolated work recognition*, IEEE Transactions on Acoustics, Speech, and Signal Processing **33** (1985), nº 3, 587–594.

[116] EP Xing, MI Jordan, SJ Russell e A Y Ng, *Distance metric learning with application to clustering with side-information*, Electronic Proceedings of the Neural Information Processing Systems Conference (2003), 521–528.

[117] W Xu, X Liu e Y Gong, *Document clustering based on non-negative matrix factorization*, Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval (2003), 267–273.

[118] T Zhang, R Ramakrishnan e M Livny, *BIRCH: an efficient data clustering method for very large databases*, ACM Sigmod Record **25** (1996), nº 2, 103–114.

[119] J Zhou, L Jia, G Hu e M Menenti, *Evaluation of Harmonic Analysis of Time Series (HANTS): impact of gaps on time series reconstruction*, Proceedings on the 2012 Second International Workshop on Earth Observation and Remote Sensing Applications (2012), 31–35, http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6261129.

Apêndice A

Complementos Gerais

A.1 Definindo os termos DM e *KDD*

O termo **mineração de dados** (*DM*, do inglês *Data Mining*) pode significar coisas distintas, dependendo do foco do especialista. Essa confusão de terminologia pode ser atribuída, em parte, ao fato desta ter sido influenciada por múltiplas áreas, como aprendizado de máquina, análise exploratória de dados e reconhecimento de padrões. Muitos dos que utilizam suas técnicas a vêem como extensões do que eles já fazem a muitos anos em seus campos.

Os próprios especialistas em DM têm contribuído para essa confusão. Alguns usam o termo como um único passo em uma sequência definida pelo processo de descoberta de conhecimento em banco de dados (*KDD*, do inglês *knowledge-discovery in databases*), enquanto outros autores se referem à mineração de dados como:

> *O processo de extração de informações válidas, compreensíveis e úteis, embora previamente desconhecidas, de grandes bases de conhecimento, para tomada de decisões cruciais em negócios.*

Nesse ponto de vista, o processo de DM é aplicado sobre uma *data warehouse* e é composto por quatro passos básicos: seleção de dados, transformação de dados, mineração de dados e interpretação de resultados. Disso, a DM, e não a *KDD*, é o processo completo de extração de informação útil das bases de dados. Como é difícil entrar em consenso sobre uma definição, para os fins deste texto, não fiz distinção entre os termos DM e *KDD* e muito menos me ocupei em fazer a distinção entre DM e alguma das diversas áreas que a influenciaram, como a estatística.

Particularmente, considero que o significado da DM é governado pelos objetivos do pesquisador, ou seja, pelas razões que o levaram a analisar os dados. Normalmente, o pesquisador que coleta, ou escolhe, os dados, o faz pensando em uma questão específica. Às vezes, a questão é bem definida e é clara a abordagem a ser aplicada, mas, às vezes, a questão tratada, apesar de bem definida, não conta com um procedimento claro para obtenção de sua resposta. Com regularidade, o processo de resolução de uma questão faz emergir outras questões, e os dados podem ser usados para responder questões completamente desvinculadas da questão inicial. Em outras situações, o pesquisador pode não ter uma questão específica, mas estar simplesmente interessado no que os dados contêm. Esse é o caso quando se analisa dados dos quais se têm pouco entendimento.

A.2 Aquisição de imagens de satélite

Texto adaptado do manual *Conceitos Básicos de Sensoriamento Remoto*
diponibilizado pela Conab (Companhia Nacional de Abastecimento).

Uma dúvida comum da comunidade de usuários tem sido como proceder para obter uma imagem de satélite. O primeiro passo consiste em identificar as instituições que comercializam ou distribuem imagens. No Brasil, o Instituto Nacional de Pesquisas Espaciais (INPE) é distribuidor das imagens *LANDSAT*, *SPOT* e CBERS. O INPE possui uma estação de recepção destas imagens em Cuiabá-MT. As instituições proprietárias dos satélites *LANDSAT* e *SPOT* cobram para disponibilizar as imagens nas estações. Algumas empresas privadas também comercializam estas e outras imagens, como, por exemplo, as imagens *Ikonos*.

As imagens *NOAA* têm custo menor, porque a instituição proprietária do satélite não cobra para disponibilizar as imagens nas estações receptoras. Várias instituições públicas e privadas recebem as imagens *NOAA*: o INPE, o INMET, a FUNCEME, a UFRGS e o CEPAGRI (Centro de Pesquisas Meteorológicas e Climáticas Aplicadas à Agricultura vinculado a Unicamp).

Uma vez escolhido o fornecedor de imagem, o passo seguinte é definir a área de interesse, como por exemplo, um município, ou mesmo uma parte do município, caso este seja de grande dimensão territorial. Se possível, deve-se determinar as coordenadas geográficas da área. O *GPS* pode ajudar nesta tarefa definindo uma coordenada central ou um polígono que envolva a região. Dependendo da localização e da dimensão da região, uma imagem pode ser suficiente. Contudo, existem casos, mesmo de pequenas áreas, onde há necessidade de se adquirir várias imagens, como na situação em que a região está localizada nos cantos das imagens. Definida a área, é possível identificar a(s) imagem(ns) a ser(em) adquirida(s). O *LANDSAT* e o *SPOT* têm um sistema de identificação das imagens composto de dois números, sendo o primeiro o número da órbita e o segundo o número da imagem dentro da órbita, também chamado de ponto. A identificação das imagens pode ser obtida no mapa denominado Sistema de Referência Universal, fornecido pelo INPE.

Por exemplo, a imagem *LANDSAT* que cobre o DF é a 221/71. A imagem pode ser adquirida inteira ou parcialmente. No caso do *LANDSAT*, a menor fração da imagem é um sub-quadrante de 45 km ×45 km. Esses sub-quadrantes são identificados pelos números de 1 a 16. Pode-se adquirir também quadrantes de 90 km × 90 km, que são identificados pelas letras A, B, C, D, E, S, W, N e X.

A.3 Softwares usados

Segue abaixo uma breve descrição dos principais *softwares* utilizados nessa tese.

MatLab é uma linguagem de programação apropriada ao desenvolvimento de aplicativos de natureza técnica. Como o próprio nome sugere, o *MATLAB* é bem adequado àqueles que desejam implementar e testar soluções com facilidade e precisão, sem perder tempo com detalhes específicos de linguagem de programação. Para isso, possui facilidades de computação, visualização e programação, dentro de um ambiente amigável e de fácil aprendizado. O nome *MATLAB* vem do inglês *Matrix Laboratory*. Essa ferramenta

foi originalmente desenvolvida para tratamento de vetores e matrizes. Os elementos básicos da linguagem são exatamente os vetores e as matrizes, embora atualmente, o *MATLAB* disponha de uma biblioteca bastante abrangente de funções matemáticas, geração de gráficos e manipulação de dados, que auxiliam muito o trabalho do programador. Além disso, possui uma vasta coleção de bibliotecas, denominadas *toolboxes*, para áreas específicas como: equações diferenciais ordinárias e parciais, estatística, processamento de imagens, processamento de sinais e finanças. A linguagem e o ambiente de programação também permitem que o usuário escreva suas próprias bibliotecas em *MATLAB*.

Weka é uma suíte de mineração de dados muito popular no meio acadêmico, desenvolvida utilizando a linguagem Java. Foi criada nas dependências da Universidade de *Waikato*, Nova Zelândia. Atualmente, é mantida por uma comunidade de entusiastas, por ser um software livre, disponível sobre a licença *GPL*.

Envi **e** *IDL Envi* . A linguagem *IDL* é a base de desenvolvimento do software *ENVI*, que serve para o processamento e para a análise de imagens de satélite. Atualmente, o *ENVI* é reconhecido mundialmente como o software líder na área de sensoriamento remoto. O *IDL* oferece uma grande variedade de rotinas gráficas, controles de interface amigáveis, além de possibilitar a adição de novas rotinas, resultando em um poderoso instrumento no desenvolvimento de visualizações interativas aplicadas ao sensoriamento remoto e ao SIG. Tanto o software *ENVI* quanto o *ENVI IDL* são programas proprietários. Empresas e organizações de investigação como *PORSCHE*, *SIEMENS* e *NASA* usam o *IDL* para desenvolver suas aplicações de visualização e de análise de dados.

*ADaM***:** *Algorithm Development and Mining* é um projeto da *NASA*, em conjunto com a Universidade de Alabama em Huntsville. É um conjunto de ferramentas de mineração de dados científicos e de imagens. Suas funcionalidades incluem reconhecimento de padrões, processamento de imagens, otimização, mineração de regras de associação, dentre outros. O sistema é composto por uma série de componentes individuais, que podem ser utilizados em conjunto para realizar tarefas complexas. O *software* possui módulos implementados em C, C++ e componentes *Python*. Um dos focos do projeto é a implementação eficiente de componentes de desempenho crítico, além do cuidado de manter cada componente do sistema o mais independente possível, visando permitir a utilização de subconjuntos de módulos apropriados para determinadas aplicações, inclusive aproveitando componentes de terceiros.