Wanderson Luiz da Silva

AGRUPAMENTOS DE SÉRIES TEMPORAIS DE IMAGENS DE SATÉLITE POR *VNS* BÁSICO COM BUSCA LOCAL E RESTRIÇÕES

Tese de Doutorado apresentada ao Instituto de Matemática, Estatística e Computação Científica da Unicamp como requisito parcial para obtenção do título de Doutor em Matemática Aplicada e Computacional. Área de concentração: Matemática Aplicada.

Orientador: Prof. Dr. Francisco A. M. Neto
Co-orientador: Prof. Dr. Jurandir Zullo Junior
Co-orientador: Prof. Dr. Stanley R. M. Oliveira

Campinas, SP
2013

Sumário

Lista de Figuras

List of Algorithms

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 $(e_j)_i = 0 sei \neq j$ 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
SR	-	sensoriamento remoto

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

Nessa tese investigamos como resolver o 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. 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 transdisciplinar 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 2 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 a evolução humana. 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 relativamente recente chamada mineração de dados[2].

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 e matemática aplicada. 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.

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

[2] A área de mineração de dados se formou em meados da década de 90, a partir da contribuição de diversas áreas como estatística, aprendizado de máquina e banco de dados.

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***".

1.2 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[3]. 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 [58]. Em especial, eles podem gerar estimativas de produção para culturas como cana-de-açúcar, que é usada para a produção de biocombustível e tem importância estratégica para a economia do país[4]. 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 uma ferramenta adequada para transformar esse labirinto de imagens em auxílios para a tomada de decisão em demandas do agronegócio[74].

A classificação de imagens de satélite auxilia na estimativa de áreas cultivadas, que, por sua vez, favorecem um reescalonamento no plantio das culturas, para que não haja 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 é o

[3] 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.

[4] 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

caso do 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 [2].

A classificação de imagens multitemporais de satélite tem a função de obter 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 [58]. 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 [3] e pode permitir a obtenção automática ou semi-automática de informações a partir de extensos bancos de imagens e dados de sensoriamento remoto.

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, onde a partição mais adequada maximiza ou minimiza uma ou múltiplas funções objetivos [91, 32, 50, 47]. 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 [28]. Por isso, existe a necessidade de métodos que previnam essas falhas. O crescimento de investidas para sanar estes inconvenientes é facilmente atestada pelo crescimento no número de trabalhos que usam meta-heurísticas para geração de agrupamentos.

Disso veio a proposta de criação de um método eficiente para resolver o problema combinatorial de agrupamento k (com restrições) e 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 classicamente sugeridas na literatura.

Os objetivos por trás dessa hipótese são dois:

i) verificar o ganho que se estabelece em um agrupamento obtido pelo 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 comparadas a qualidade dos agrupamentos obtidos com a incorporação da *VNS*.

Tanto a escolha da heurística k-médias quanto da meta-heurística *VNS* se justificam pelo poder de cada uma e das várias características que serão expostas e detalhadas nos capítulos mais a frente. Além disso o k-médias pode ser facilmente implementado e compreendido, com baixo custo computacional, enquanto a *VNS* explora conceitos simples 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 no 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 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*). Também existem trabalhos 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 da métrica usual (L_2) sobre um conjunto de características baseadas nos coeficientes do polinômio trigonométrico de ajuste da séries temporal. Em outras palavras, tem-se a hipótese:

Hipótese 2: Seja $\mathscr{F}$ uma extração de características baseadas no ajuste pela base de Fourier (detalhes em 4.5) de séries temporais do índice sumarizante *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, i.e. 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.

Isso 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 no lugar dos dados originais. Assim, a partir da representação contínua dos dados por regressões por polinômios trigonométricos, damos mais pesos às características de oscilação da série do que de seus valores propriamente ditos.

1.4 Contribuições

A principal contribuição dessa tese é o algoritmo de agrupamento que se baseia na proposta inédita de uso da *VNS* acoplada a variantes que usam restrições (semissuperviosionadas) do *kmeans*. Existem trabalhos que fazem uso da *VNS* com buscas locais baseadas em heurísticas simples como *k*-médias ou *k-harmonic means* [4, 48, 16] mas nenhuma delas faz uso da incorporação de restrições e nenhuma foi aplicada sobre um conjunto de 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 [89, 7] das séries temporais (*Harmonic ANalysis of Time Series*) de *NDVI* [63, 62, 61, 96, 9]. Mas essa proposta não contempla toda a diversidade de possibilidades, como o uso de operadores diferenciais [94], 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[5] com ganho de qualidade sobre estratégias tradicionais como o k-médias com múltiplos recomeços (*k-means multistart*). O método proposto corrige déficits de sensibilidade a pontos iniciais e inconvenientes de convergências a mínimos locais.

Além da contribuição óbvia decorrente do cumprimentos 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 favorecer áreas que tenham sua utilidade pública mais claramente justificadas.

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, alternado-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. Atendando para o fato de que a forma aportuguesada, clusterização, foi usada como sinônimo de agrupamento.

Este texto está dividido em 6 capítulos e 3 apêndices. Os apêndices incluíram textos que tornam a tese a mais auto contida possível. A ideia é que, pelo caráter multidisciplinar do tema, conceitos relativamente introdutórios para alguns segmentos, como a conceituação da *KDD*, fossem apresentados de forma breve a especialistas de áreas alheias a ciência da computação, sem a necessidade de consulta externa. Os capítulos centrais da tese e seus conteúdos 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 aquilo que é

[5] 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*).

necessário para compreender a natureza dos dados de sensores orbitais, principalmente os do tipo AVHRR/NOOA, 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 sobre o ponto de vista das teorias que foram efetivamente usadas 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 de *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 da *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-médiaspadrão 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 trabalhos que fazem uso dela. Para conclusão, aplicou-se o algoritmo nos dados de refletância da cidade de Jaboticabal em São Paulo obtidos pelo sensor *AVHRR* a bordo do satélite *NOAA*-15 e nos dados de reflectância do estado do Mato Grosso obtidas pelo sensor *MODIS* a bordo do satélite *TERRA*. Neste ponto, os testes se diferenciam dos acima pela natureza tempo-dependente das séries temporais de imagens de satélite. Portanto é para estes dados que se faz 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 interdisciplinares necessários à compreensão do problema e da proposta apresentadas. Em sensoriamento remoto (SR) apresentamos aquilo que é necessário para compreender a natureza dos dados de sensores orbitais (principalmente os do tipo *AVHRR/NOOA*) e a transformação de sumarização destes dados, conhecida como índices vegetativos (em especial o *NDVI*). Sobre agrupamentos, apresentamos, 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 SR 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) [27]. 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 com sistemas de estimativas agrícolas [60].

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 [38].*

O sensoriamento remoto orbital se ocupa de medir as propriedades da superfície da terra, sem possuir um vínculo físico com ela, através da análise e processamento das iterações 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, são os satélites que se tornaram o instrumento de captação mais comum para análises e pesquisas na área. O diagrama 2.1 exemplifica o funcionamento de um sistema de imageamento de SR passivo, ou seja, imageamento de áreas contanto com uma fonte de iluminação/radiação natural.

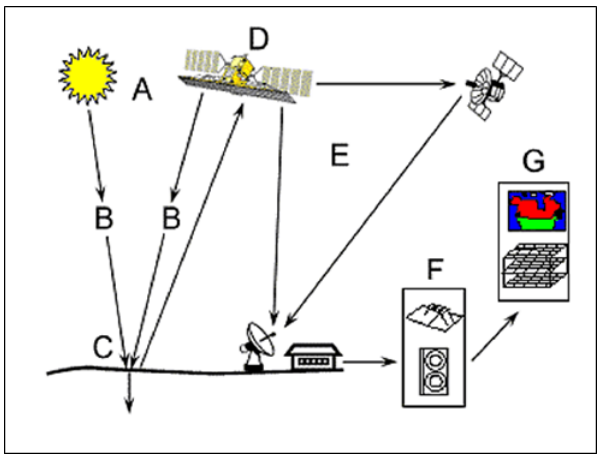


Fig. 2.1: Elementos de um sistema orbital passivo de sensoriamento remoto.
Fonte: *Canadian Centre of Remote Sensing* (*CCRS*).

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 uma da principais fontes 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 espectais 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 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 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 fim de um sistema de SR sempre é usar as informações obtidas a partir do acervo de imagens do alvo 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, trata-se do registro das refletâncias das áreas visadas pelo satélite, 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 do solo (*land cover*) ou o acompanhamento de mudanças de superfície. A curva de refletância espectral (Fig. 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. O desenho de sua curva, a assinatura espectral da vegetação, pode ser explicado por

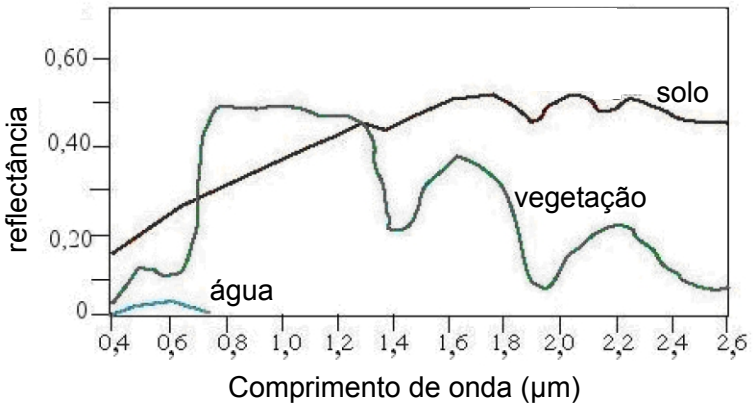


Fig. 2.2: Curva espectral do solo, água e vegetação
Fonte: Adaptado de Introduction to Remote Sensing, James B. Campbell, 2006.

sua fisiologia [53, 45].

2.1.2 Satélites e sensores

O satélite é um veículo posto em órbita, baixa ou alta, e é comumente composto por 3 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 REM atravessa um sistema óptico e é focalizada sobre detectores. Estes transformam a radiação em sinais elétricos que são gravados em fita magnética. 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 1,6km terá pixels de $1{,}6 \times 1{,}6$ km^2 de área. Em outras palavras, o registro da a intensidade luminosa captada é associada a uma região definida pela área do pixel.

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 diferenciar. 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 a quais faixas do espectro de luz o sensor do satélite pode coletar. Existem sensores que captam aspectos termais do alvo, relacionados ao infravermelho longo (*far infrared*), 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/NOOA 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 imprópria 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 desenvolver 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) [67], os índices de vegetação cresceram em variantes, especialmente na região do visível e 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 vegetações usam o índice de vegetação pela diferença normalizada (*NDVI*, do inglês *Normalized Difference Vegetation Index*) [68].

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 e produtividade [55]. Seus valores variam, em teoria, entre -1 e 1, de forma que valores próximos de 1 indicam

áreas altamente vegetadas, enquanto valores próximos de zero, ou negativos, indicam ausência de vegetação.

2.1.4 Imagens digitais

As imagens SR são normalmente compostas por dois arquivos: um cabeçalho (*header*) 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), onde cada qual correspondente 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 o brilho ou radiância da área correspondente ao pixel, no terreno. Tanto x e y quanto f só assumem 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 *NOAAA*, por exemplo, possui abrangência espacial, longevidade e baixo custo para aquisição de imagens, o que permite coberturas diárias das culturas de interesse.

Desse caráter multitemporal também decorre a análise de produtos como perfis temporais de *NDVI*, que podem gerar parâmetros quantitativos a partir dos quais se obtêm informações sobre a biomassa ao longo dos seus estágios fisiológicos [63, 66, 51]. 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 através da construção de imagens por composição dos valores máximos (*mvc*, do inglês *maximum value composition*) [54].

2.1.5 Classificação de imagens de satélite

A classificação de imagens de satélite vem concentrando esforços desde de que o uso de imagens de satélite se tornaram de uso comum e fácil acesso. As imagens AVHRR/NOAA vêm sendo disponibilizadas pela *NASA* desde 1995 para uso do Cepgri/Unicamp e o INPE vem disponibilizando imagens Landasat e Spot em seu site desde 2001. Existem trabalhos que atuam no problema de classificação fazendo uso destes acervos, aplicando técnicas de *DM* [97, 89, 7], 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 áreas cultivadas com milho, soja ou café. Ao se classificar uma imagem, assume-se que os alvos diferentes apresentam propriedades espectrais específica e que cada ponto pertence a uma única classe (abordagem *crisp*), ou a múltipla classes com intensidades distintas (abordagem *fuzzy*).

Pontos representativos de uma certa classe devem possuir padrões próximos de tonalidade, de cor e de textura. A classificação pode ser por inspeção visual, onde o pesquisador interpreta visualmente os elementos da imagem a fim de identificar o cenário ali registrado. Essa interpretação é fortemente dependente da qualidade discriminativa do observador e de outros aspectos subjetivos atrelados a 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 sobre algum critério. Nessa abordagem, os algoritmos usualmente adotam estratégias baseadas em crescimento de regiões ou detecção de bordas [39]. 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 utilizadas 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 alvos e de sobrepor as classes. Além disso, é fato que as classes, na realidade, não se enquadram em padrões geométricos perfeitos.

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 $p(a < z < b) > l_z$, onde l_z é um limiar mínimo. Os algoritmos mais utilizados neste tipo de classificação são o *k-means* e o método da máxima verossimilhança (*EM*, do inglês *Expectation Maximization*).

2.2 Análise de grupos

Indiscutivelmente, estamos cercados por um oceano 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 [3]. 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 [49].

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 se constituindo numa área de pesquisa efervescente 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 gerado na última meia década, como podemos verificar na Figura 2.3.

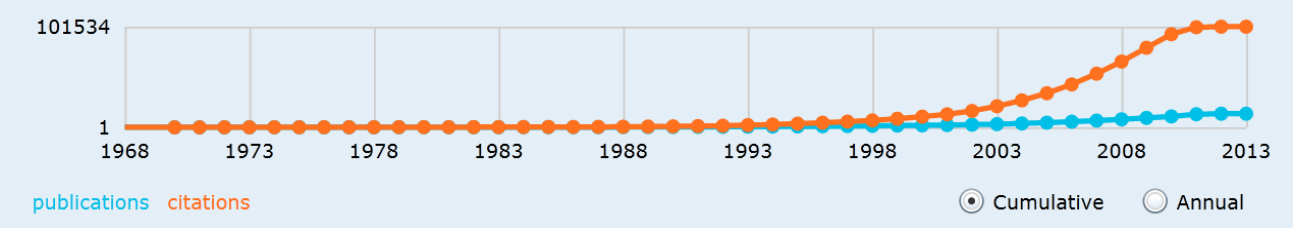


Fig. 2.3: Gráfico do *Microsoft Academic Research* em setembro de 2013, mostrando o crescimento no número de citações e publicações envolvendo o problema de clusterização.

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 agrupamentos orientados por funções objetivo, ou baseada em protótipos, como o *k-means* e 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 uma forma geral, uma divisão bem aceita e recorrente na literatura é a seguinte:

Métodos partitivos são uma coleção de técnicas que, a partir de m particionamento inicial,

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

de forma iterativa, mudam os elementos de grupos buscando a melhoria de algum critério 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 as em que os envoltórios convexos de vários grupos são sobrepostoss. Os algoritmos partitivos usualmente apelam para o conceito de protótipo, onde 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 *k*-médias [78], 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.

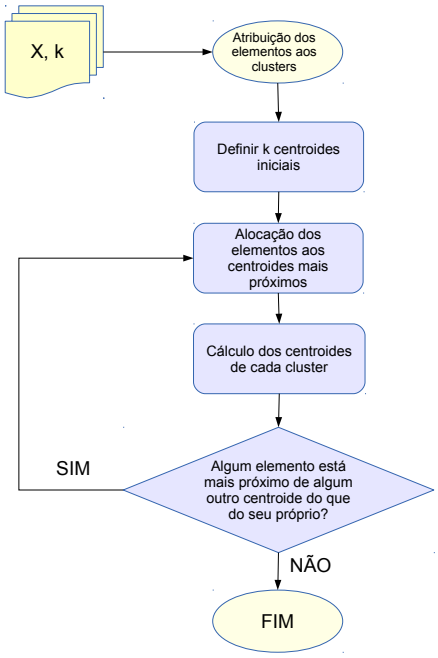


Fig. 2.4: Fluxogramas simplificado do funcionamento do algoritmo *k*-médias.

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

[2]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 clusters.

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.

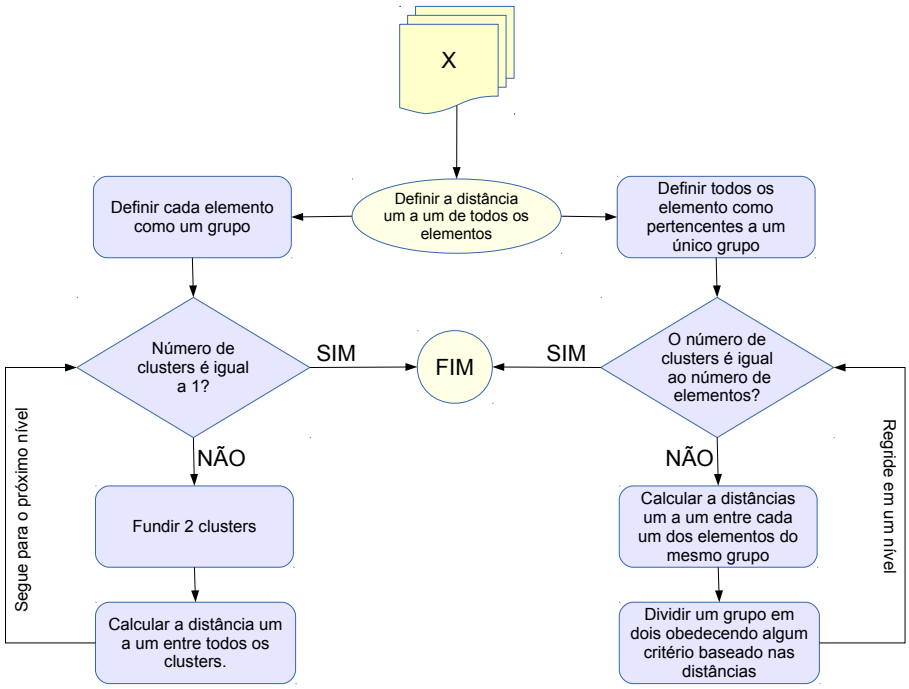


Fig. 2.5: Fluxogramas simplificado do funcionamento dos algoritmos aglomerativos (a esquerda) e divisivos (a direita).

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 pelo 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*) [113].

Uma linha de uso de métodos hierárquicos que atenuam suas deficiências e melhora a qualidade dos agrupamentos obtidos são alternativas híbridas que resultam em algoritmos multi-fases como o BIRCH (*Balanced Iterative Reducing and Clustering Using Hierarchies*) [117] e *Chameleon* [70].

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

O *DBSCAN* [82], representado na Figura 2.6, classifica como pontos centrais aqueles que têm em sua vizinhança ϵ uma quantidade miníma de pontos ($MinPts$), enquanto os de limite são queles 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.

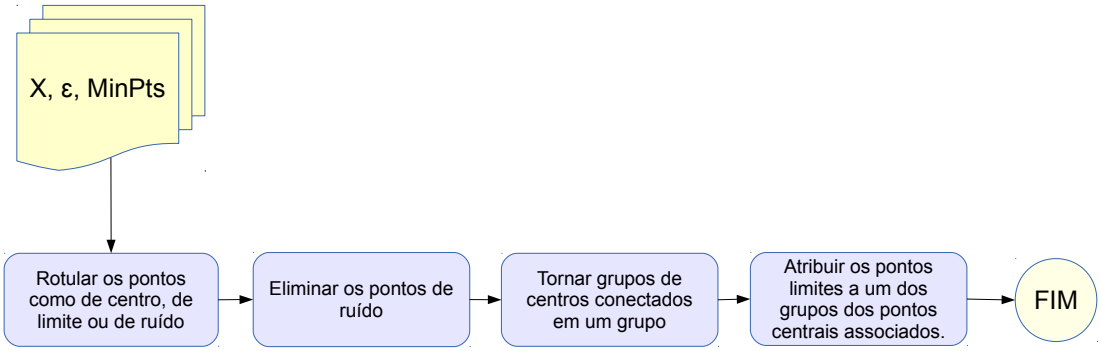


Fig. 2.6: Fluxograma simplificado do funcionamento do *DBSCAN*.

Os algoritmos *DBSCAN* (*Density-Based Saptial Clustering of Aplications with Noise*), *OPTICS* (*Ordering PoinTs to Identify the Clustering Structure*) [6] e *DENCLUE* (*DENsity-based CLUstering*) [52] são alguns dos representantes mais conhecidos de métodos baseados em densidade.

[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.

2.2.2 Clusterização de séries temporais

As técnicas de agrupamento podem ser divididas em agrupamentos de dados estáticos e 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 [86, 75, 69].

Metodologias mais recentes se baseiam na transformação dos dados brutos em um vetor de características de dimensão reduzida [114, 102, 44, 110] ou um conjunto de parâmetros [92, 79]. 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.

O espírito por trás da extração de características é fazer uma transformação dos dados a fim de gerar um *gap* semântico, ou seja, mudar a representação 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 descritas a depender da escolha da métrica e das transformações a que se sujeita os dados. 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:

$< \mathbb{I}, d >$, baseados 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*).

$< E, d >$, 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.

A ideia por trás do descritor $< E, d >$ é que E seja uma transformação que gere vetores cujas coordenadas sejam quantizações 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* [65]. Resultados recentes mostram que na maioria das vezes, o desempenho do classificador supervisionado pode ser melhorado baseado na 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:

- máxima expectativa semissupervisionado (*Semi-supervised EM*) [90],
- treinamento conjunto (*Co-training*) [18],

- máquinas de vetor de suporte transdutiva (*Transdutive SVM's*) [40],
- 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 clusters, 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 [77]. Entretanto, como existem múltiplas formas de agrupamento, o algoritmo pode gerar clusters 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* [22].
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* [15].

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 re-rotulados.

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 *clusters* diferentes. Durante o processo de re-rotulação, o registro é atribuído ao *cluster* mais próximo, desde que 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, enquanto 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 *clusters*

Quando um algoritmo produz um agrupamento, devemos nos perguntar se existe alguma forma de avaliar a qualidade do resultado obtido. Agrupamentos distintos são perfeitamente comuns de serem obtidos a partir de algoritmos distintos, e deve-se ser capaz de escolher a melhor resposta. Entretanto como não existe uma forma inequívoca de dizer qual o melhor algoritmo a partir de um índice universal da qualidade, 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 perceba 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 [59] é uma estratégia que vem ganhando popularidade, na qual 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. A ideia é que cada grupo tenha um desenho característico que permita distingui-los. 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.

Apesar de existirem esforços no sentido de se aprimorar a técnica [87], ocorre uma 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]. De uma forma geral, o coeficiente de silhueta de um *cluster* avalia o quão coesos os elementos

[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.

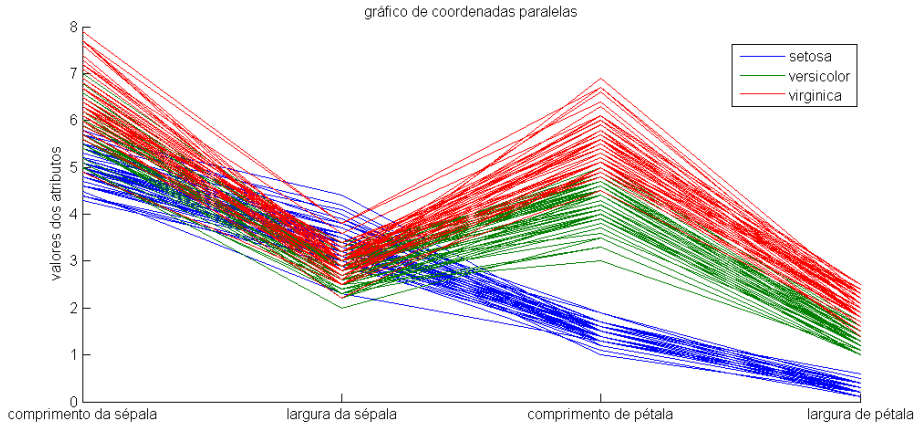


Fig. 2.7: Exemplo de gráfico de coordenadas paralelas para a base de dados Íris de Fisher.

do mesmo grupo são e o quão separados os elementos de grupos distintos estão. Para calcular o coeficiente de silhueta $S(x_i)$, com $x_i \in C$, usa-se

$$a(x_i) = \frac{\sum_{x_j \in C - \{x_i\}} ||x_i - x_j||}{|C| - 1}, \tag{2.2}$$

$$d(x_i) = \frac{\sum_{x_j \in C_l} ||x_i - x_j||}{|C_l|}, \text{ onde } C_l \neq C, \tag{2.3}$$

$$b(x_i) = \min_{l \in \{1,\ldots,k\} - \{i\}} d(x_i), \tag{2.4}$$

e define-se

$$S(x_i) = \frac{b(x_i) - a(x_i)}{\max\{b(x_i), a(x_i)\}}. \tag{2.5}$$

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 media 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 *clusters* 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.

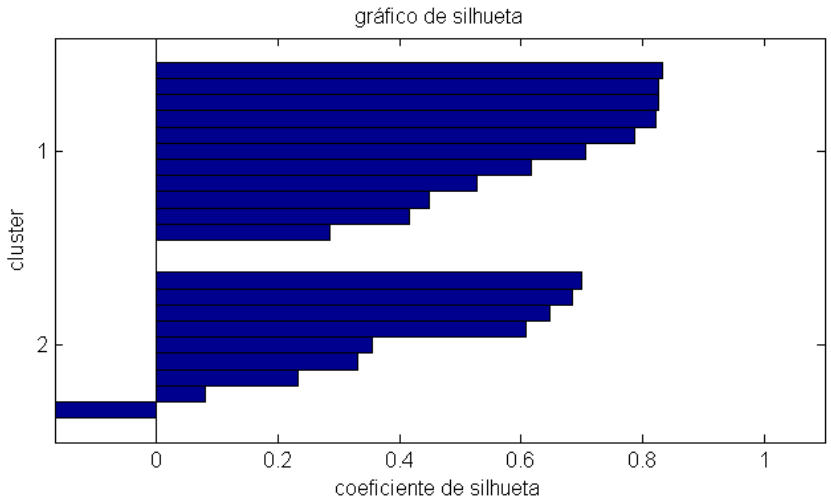


Fig. 2.8: Exemplo de gráfico de silhueta para 20 pontos gerados aleatoriamente em torno dos pontos $(1,1)$ e $(-1,-1)$ e agrupados pelo algoritmo *kmeans*.

Existem várias outras técnicas de representação de dados [65], 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 bizarras, 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 *clusters* 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 a 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 do *SSE* (*sum of squared errors*), além de estar implícito no gráfico de silhueta. Daí, um contraponto é o uso do método extrínseco, onde se pressupõe conhecimento da verdade terrestre.

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 qualidade 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.

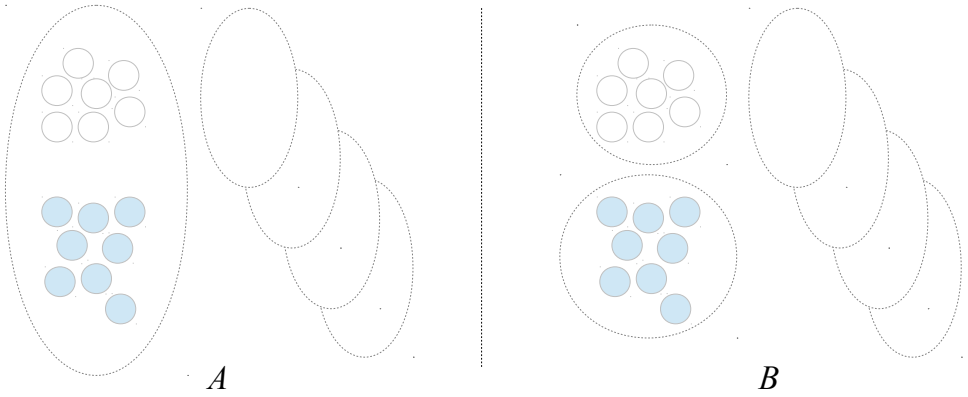


Fig. 2.9: O agrupamento $\mathscr{A}$ tem elementos de dois grupos disjuntos em um mesmo *cluster*, ao contrário do agrupamento $\mathscr{B}$. Uma métrica Q que obedece o critério de homogeneidade terá $Q(\mathscr{A}) < Q(\mathscr{B})$.

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 completeza 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.

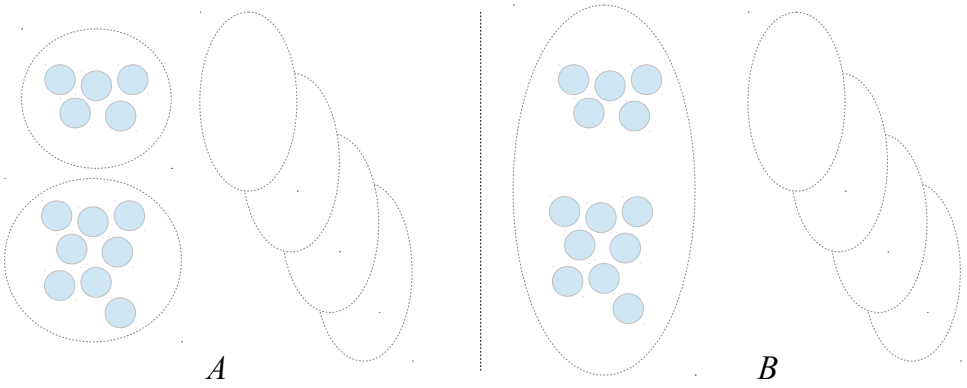


Fig. 2.10: O agrupamento $\mathscr{A}$ separa elementos de um mesmo grupo, ao contrário do agrupamento $\mathscr{B}$. Uma métrica Q que obedece o critério de completeza terá $Q(\mathscr{A}) < Q(\mathscr{B})$.

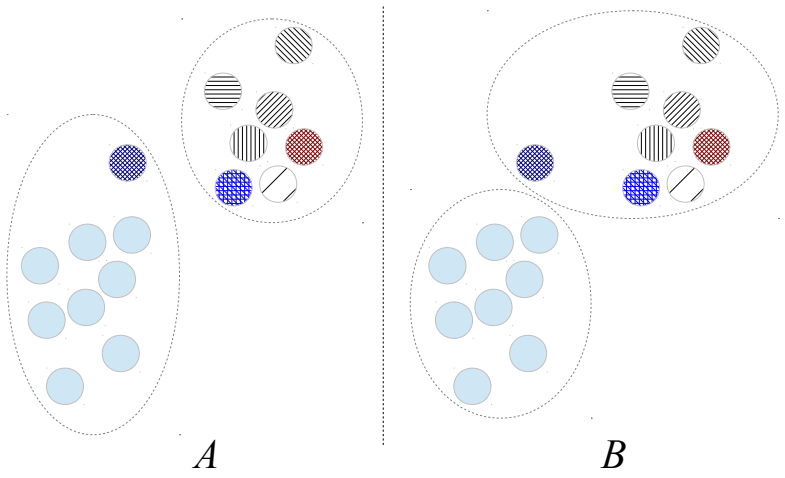


Fig. 2.11: O agrupamento $\mathscr{A}$ incorpora um elemento heterogêneo a um grupo homogêneo. O agrupamento $\mathscr{B}$ cria um grupo exclusivo de termos heterogêneos. Uma métrica Q que cria grupo para elementos não dominantes terá $Q(\mathscr{A}) < Q(\mathscr{B})$.

Preservação de *clusters* pequenos. Grupos pequenos, quando desmembrados, geram uma penalização maior que quando se desmembram 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 as únicas, existindo outras como as restrições de Dom [30] 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, como:

- **métricas baseadas em correspondência de conjuntos**, como pureza e pureza invertida;

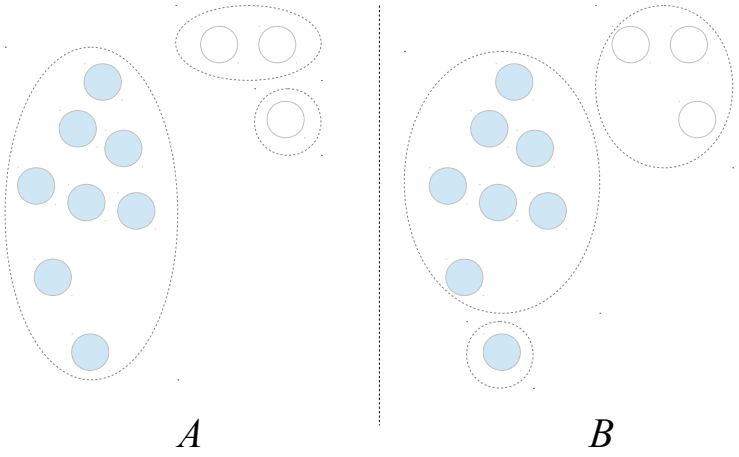


Fig. 2.12: O agrupamento $\mathscr{A}$ quebra o grupo menor em dois, enquanto o agrupamento $\mathscr{B}$ quebra o grupo maior em dois e preserva o grupo menor. Uma métrica Q que obedece o critério de preservação de grupos pequenos terá $Q(\mathscr{A}) < Q(\mathscr{B})$.

- **métricas baseadas em contagem de pares**, como estatística aleatória e coeficiente de *Jaccard*;

- **métricas baseadas em entropias**.

Entretanto, as métricas mais satisfatórias são as *B-cubed* [10], por satisfazerem todos os quatro critérios, ao contrário de todas as outras família de métricas [5, 65].

As métricas lembrança (*B-cubed recall*) e precisão (*B-cubed precision*) 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 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 um outro elemento x' tal que $x \neq x'$, então

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

Daí, temos:

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

onde $P(x_i) = \dfrac{\sum_{x_j \in C_i} exatidão(x_i, x_j)}{||C_i|| - 1}$, sendo $C_i = x_i\{x \in \mathbf{X} | c(x) = c(x_i)\}$ o conjunto dos elementos do mesmo grupo de x_i.

$$R(\mathscr{C}) = \frac{1}{n}\sum_{i=1}^{n} R(x_i) \tag{2.7}$$

onde $R(x_i) = \dfrac{\sum_{x_j \in G_i} exatidão(x_i, x_j)}{||G_i|| - 1}$, 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.

Uma forma comum de combinar índices consiste em usar o coeficiente de Van Rijsbergen (*Van Rijsbergen's F*), que é obtido da seguinte forma

$$F(R, P) = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}}, \tag{2.8}$$

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

Apesar das métricas baseadas em entropia não atenderem todas as restrições declaradas, elas têm sido amplamente adotadas em validação de classificadores. A entropia total de um agrupamento [105] é a média ponderada das entropias de cada grupo, ou seja,

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

onde $\#G_j$ é o número de elementos do grupo j, e E_j é a entropia do grupo j, ou seja é uma medida da confusão de um grupo, avaliando a quantidade relativa de elementos das diversas classes que compõem o *cluster*.

$$E_j = \sum_{i=1}^{k} \frac{\eta_{ij}}{\#G_j} \log_2 \left(\frac{\eta_i j}{\#G_i} \right), \tag{2.10}$$

sendo η_{ij} sendo o número de elementos da classe i no grupo j.

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*)[11], variação de informação (*variation of information*)[116] ou medida$-V$ [98].

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 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:

[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 a área de clusterização.

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 $\mathbf{P} = \{C_1, C_2, \cdots, C_k\}$ 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 necessáriamente obedecer as 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.

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

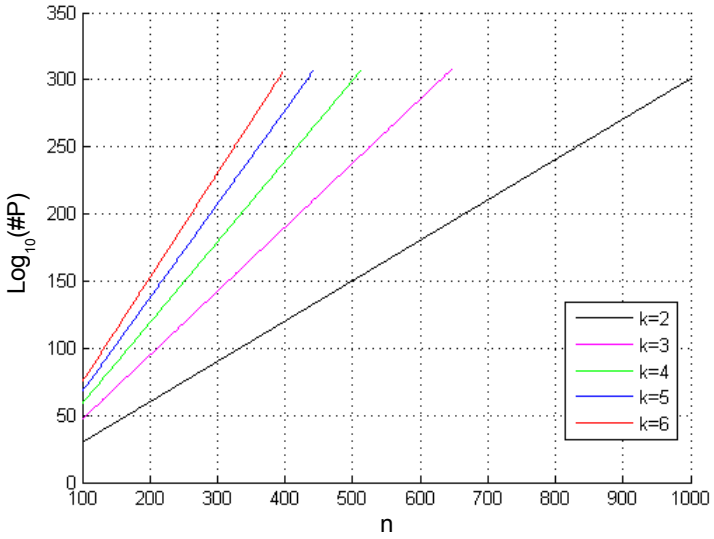


Fig. 3.1: Variação da quantidade de partições ($\#\mathscr{P}$) com k entre 2 e 6.

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 centro $\mathbf{c}_j \in \mathbb{R}^d$, sendo cada instância associada ao centro mais próximo. Daí, o agrupamento é definido pela matriz $C = [\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_k]^t$ de ordem $k \times d$.

Os centroides ocupam um lugar de destaque dentro de vários métodos de agrupamento e podem, de forma geral, ser definidos pela expressão seguinte:

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

onde a função $m(\mathbf{c}_j|\mathbf{x}_i)$ é uma medida de pertinência de $\mathbf{x}_j$ 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 em $[0,1]$, ou *crisp*, caso em que $m(\mathbf{c}_j|\mathbf{x}_i)$ assume necessariamente o valor 0 ou 1. A função $w(\mathbf{x}_i)$ é uma medida do impacto que $\mathbf{x}_i$ tem no cálculo de $\mathbf{c}_i$. É uma função que 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 derivada da média harmônica. Essa mudança na forma de calcular o centroide torna o *k-harmonic means* menos sensível aos pontos de inicialização e a *outliers*.

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 $l = \underset{j \in [k]}{\operatorname{argmin}} ||\mathbf{x}_i - \mathbf{c}_j||$. Para o algoritmo *EM* [29], 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 $\mathbf{c}_j$ [46]. E, para o *k-harmonic means*, temos que

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

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

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*. Mesmo independentemente das características 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. Sendo assim, 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 calcular $\nabla SSE(\mathbf{c}) = \mathbf{0}$ para assegurar o resultado acima. Mesmo para normas arbitrárias, a função $SSE(\mathbf{c})$ é convexa e, com algumas exceções, terá um único minimizador. □

Por todos estes aspectos, o centroide é um 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*

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

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 do 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 possuem 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*. Em vários destes casos, resultados obtidos nestes 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 (Propriedades 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 por um conjunto de centroides e $\mathscr{P}$ o conjunto de todas as partições possíveis, não é difícil perceber que $\#\mathscr{P}' \leq \#\mathscr{P}$. 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 $(\mathbf{w}^t\mathbf{x}) \cdot (\mathbf{w}^t\mathbf{y}) < 0$.*

Disso temos 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.

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

[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.

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 [19].

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.

Outros trabalhos também configuram grandes métodos, como a otimização por enxames de partículas (*particle swarm optimization*) [33, 103] e o algoritmo de evolução diferencial [106], mas os métodos citados na linha do tempo têm figurado como as linhas mais comuns de pesquisa em meta-heurística dos últimos 50 anos.

3.3 Busca em vizinhança variável

"*Simplicity is the ultimate sophistication.* "
- Leonardo da Vinci

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] e tem se mostrado simples, coerente, eficiente e altamente precisa em

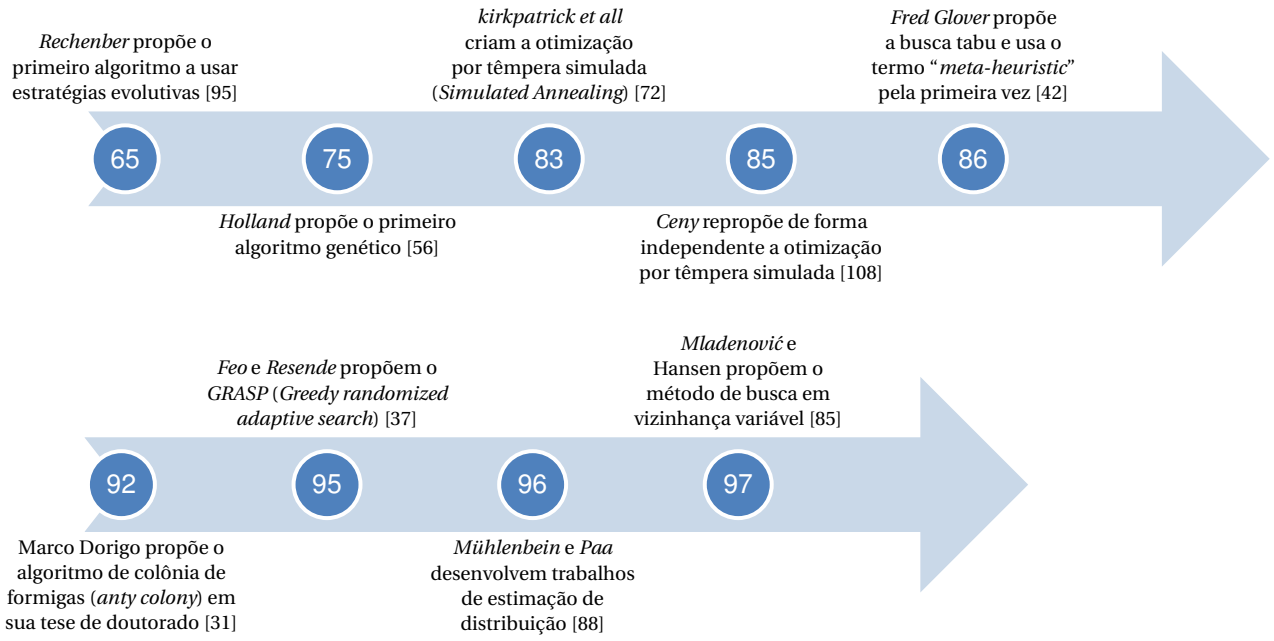


Fig. 3.2: Cronologia das principais meta-heurísticas da atualidade.

vários domínios de aplicação. Ela é frequentemente usada par resolver problemas da forma $\min\{f(x)|x \in \Omega, \Omega \subset S\}$, sendo $f : S \subset \mathbb{R}^d \to \mathbb{R}$. Onde 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. Seus princípios têm o propósito de fornecer mudanças sistemáticas de vizinhanças constituídas de 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 *VNS* é imprescindível que se entenda seus princípios e suas estruturas, e um elemento central dentro do conceito de busca por *VNS* é o conceito de vizinhança. A estrutura de vizinhança de $\mathbf{x}$ é uma coleção de pontos factíveis próximos a $\mathbf{x}$, e pode ser induzida por uma métrica d ou quase-métrica do espaço de soluções. A coleção de pontos $N_i(\mathbf{x}) = \{\bar{\mathbf{x}} \in \Omega | d(\mathbf{x}, \bar{\mathbf{x}}) \leq i\}$ é um exemplo de estrutura de vizinhança aninhada, i.e. $N_i(\mathbf{x}) \subseteq N_{i+1}(\mathbf{x})$. Outros conceitos essenciais são o de mínimo local e o de mínimo global.

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 vinhança.

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

(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. Partindo de uma solução $\mathbf{x} \in \Omega$ a *VNS*investiga $N_i(\mathbf{x})$, conjunto de soluções da vizinhança i de $\mathbf{x}$, até que se encontre uma solução melhor. Em seguida, inicia-se uma busca centrada nesta nova solução (figura 3.3).

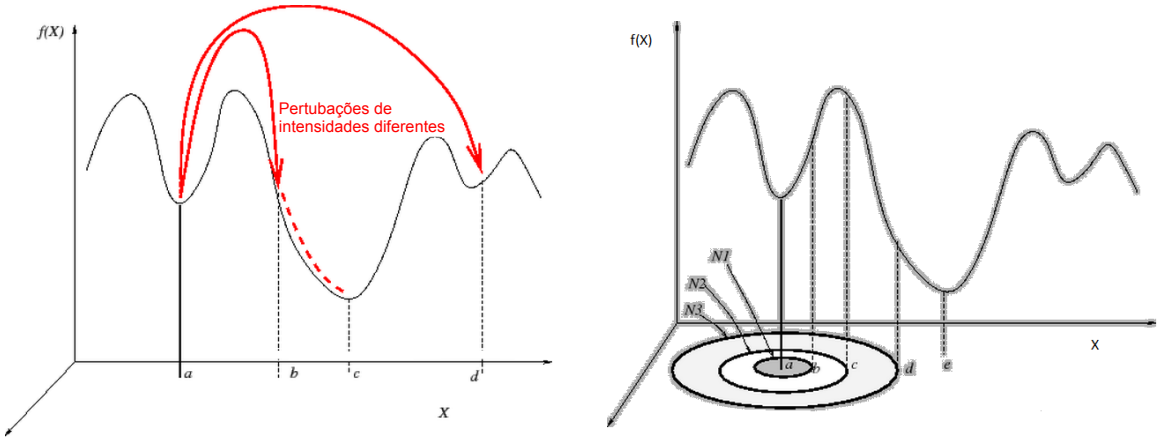


Fig. 3.3: Na *VNS*, as vizinhanças são definidas de forma incremental, a fim de explorar de forma sistemática soluções cada vez mais distantes da solução corrente.
Adaptado de `lion.disi.unitn.it/reactive-search/thebook` em novembro de 2013.

Esquemas da *VNS*

As buscas centradas nas soluções podem ser feitas de forma determinística, estocástica ou híbrida. E a escolha de como fazer a aplicação dos princípios do *VNS* define esquemas distintos.

A aplicação, o tempo e a qualidade exigidas pelo usuário desempenham um papel central na escolha da estrutura de vizinhança e do esquema a se adotar. Se o esquema de *VNS* é altamente desenvolvido e evolutivo, ele pode abranger uma mudança de estrutura de vizinhança em todas as iterações e toda a informação fornecida pode ser utilizada para resolver um problema específico.

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.

Algorithm 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 $N_i(\mathbf{x})$
    $\mathbf{x} \leftarrow$ NChange($\mathbf{x}, \mathbf{x}', i$)
  until $i = i_{max}$;
until *nenhuma melhora é obtida*;

A função NChange, que avalia se $\mathbf{x}'$ é melhor que $\mathbf{x}$, é dada no Algoritmo 2. Como descrito no Algoritmo 1, caso seja $\mathbf{x}'$ melhor que $\mathbf{x}$, reinicia-se o processo centrado em $\mathbf{x}'$, caso contrário, a busca é feita na próxima vizinhança. O propósito dessa ação é comparar a solução corrente à soluções da vizinhança, obtidas por algum tipo de pertubação, com o propósito de fugir de mínimos locais.

Algorithm 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) tem um pressuposto determinístico quando realiza a busca $\mathbf{x}' \leftarrow \underset{\bar{\mathbf{x}} \in N_i(\mathbf{x})}{\operatorname{argmin}} f(\bar{\mathbf{x}})$ e isso pode gerar inconvenientes como a ciclagem. Para domínios que apresentem esse problema, existe o esquema *RVNS*, que usa a função Shake($\mathbf{x}, i$), que escolhe, 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 serve para adicionar uma pertubação aleatória, mas controlada, de $\mathbf{x}$. Por conseguinte, ao contrário do *VND*, o *RVNS* gera diversidade de soluções através de uma estratégia estocástica.

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

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

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

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

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 grande variedade de problemas. Vale citar que a proposta de clusterização por meta-heurística de *VNS* de *Pierre Hansen* e *Nenad Mladenović* é a aplicação do *Basic VNS* com busca local por uma heurística criada por eles, chamada j-means [48].

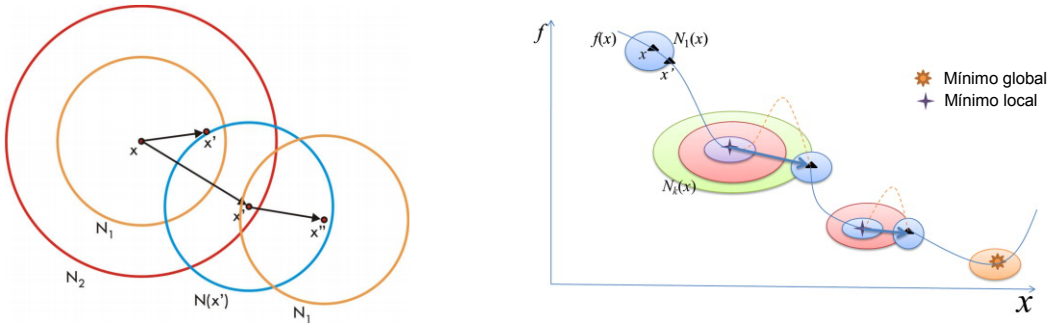


Fig. 3.4: No esquema básico, as vizinhanças são definidas de forma incremental e dentro de cada uma delas se gera uma busca local. A ideia é que as pertubações geradas pela mudança de vizinhança permitam encontrar mínimos locais melhores, dentro de outros vales, ou mesmo o mínimo global.

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

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

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*,

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

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

A atualidade, a arquitetura da *VNS* tem se mostrado eficiente em geração de soluções factíveis para grandes problemas de programação mista e em 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 do primal-dual *VNS*, que tem se mostrado bem sucedido na busca de soluções exatas para problemas de localização de larga escala [41].

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 variedade de soluções a partir de modificações controladas dos centroides. Como exemplo, temos que 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 $i \in [k]$ e $\delta_1 \in \mathbb{R}^d$, 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 [48] 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 barato e simples, quando associado ao *VNS*, tem potencial de encontrar boas soluções com baixo custo computacional. Nesta linha, entre as muitas heurísticas que resolvem o problema de minimizar o *MSSC* (do inglês *minimum sum-of-squares clustering*), a mais conhecida e, provavelmente, a mais utilizada, é a das k-médias [64, 78, 57]. Apesar de não garantir uma solução ótima global, por ficar presa em mínimos locais, suas aplicações mostram que, na prática, ela se sai muito bem em um domínio variado de base de dados.

A função Inicialização($\mathbf{X}, k$) pode assumir de várias formas distintas e pode usar estratégias para promover boas escolhas [23, 112]. As formas mais comuns 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 randômica (*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.

Algorithm 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**

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

  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

Observa-se que, para o algoritmo padrão do 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 randômica 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 [12, 13], algoritmo de Lloyd's [76], *hard c-means* [20] ou *h-means* [48]. 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 sempre o mesmo.

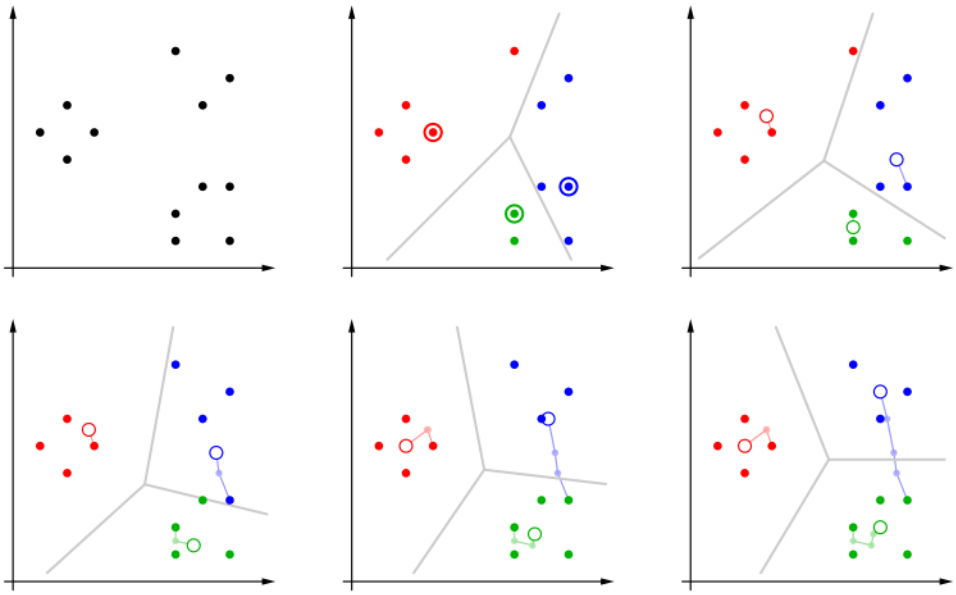


Fig. 3.5: Uso do k-médias em um conjunto simples de dados bidimensionais, onde as linhas em cinza formam um diagrama de *Voronoi*.

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 a 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 [80]. 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. [43, 81, 24, 21, 36, 107]).

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 em 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 Ω, conjunto das soluções factíveis, se $\Omega = \emptyset$, tem-se um problema infactível e, caso $\Omega = S$, 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, ao se estabelecerem trocas de grupos baseadas na quantização de algum critério que seja dependente do arranjo corrente. 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 [35].

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:

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

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

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 $j = \underset{l \in [k]}{\operatorname{argmin}} ||x - c_l||_2^2)$.

(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 $\exists \mathbf{w} \in \mathbb{R}^d$ tal que $(\mathbf{w}^t\mathbf{x}) \cdot (\mathbf{w}^t\mathbf{y}) < 0$.

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 possuí 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 são linearmente separáveis.

Escolhidos a função objetivo e o espaço factível, outros conceitos importantes a se definir são o de mínimo local e o de mínimo global. Um ponto a $C' \in \Omega$ 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'. De forma análoga, C^*, é um 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 (*MSSC*, do ingês *minimum sum of square clustering*) é sabidamente um problema NP-difícil [26], ou seja, é um problema inerentemente complicado, independentemente do algoritmo adotado.

A função $f(C)$ (equação 4.3) é complicada em vários sentidos. 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

[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.

contínua, ou seja, não é possível definir para esta função um gradiente para todo o espaço de solução. Comumente, quando a função objetivo é muito mal comportada, 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 três 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.

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 7.

Definição 7 (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:

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

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

onde L é um limiar mínimo de distância, que pode variar à medida que se muda i.

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*) A distorção de um elemento ao seu centroide deve levar em consideração não só o valor de sua distância, mas de sua distância relativa. 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 *clusters* 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 [73] e em densidade [25].

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:

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

onde

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

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

A ideia é que a realocação leve em conta o quão *outlier* 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:

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

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 *clusters*.

As propostas de vizinhança $N^{(1)}$ e $N^{(2)}$ vão gradativamente aumentando o nível de pertubação ao realocar um número maior de elementos. Entretanto mas 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, como:

$$N_i^{(3)}(C_1) = \left\{ C \in \Omega | \Theta(C_i) \cdot \mathbf{e}_j \geq L_i \right\}, \text{com } L_i = 5 - 0,1i. \tag{4.10}$$

A vizinhança $N_i^{(3)}$ é uma vizinhança definida pela realocação de todos os *outliers*, onde o conceito de *outlier* vai variando à medida que se muda de vizinhança. Inicialmente, existe um critério mais rígido, onde o elemento a ser realocado deve estar a mais de 5σ da média das distâncias dos elementos do grupo ao centroide. Em seguida, sucessivamente, relaxa-se essa condição até chegar ao ponto que qualquer instância que diste mais de $(5-0,1i_{max})\sigma$ do centroide é realocada.

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

4.3 Incorporando conhecimento

Incorporar conhecimento, mesmo em pequena quantidade, pode beneficiar a qualidade dos grupos formados [14, 17, 111, 115]. 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 em 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, de forma implícita ou explícita, uma superfície de decisão 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.

Seeds **para inicialização,** que são um conjunto de valores a serem usados como pontos iniciais. 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 a que nos propomos, 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 competitivos 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 de função objetivo. 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 obedecendo às restrições *must-link* e *cannot-link* e, durante o seu processo de rerotulação, esse algoritmo atribui um registro ao grupo mais próximo, desde que isso não viole nenhuma restrição.

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 incorporar as restrições é que ele estes algoritmos 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 8 (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 um*

hiperparalelepípedo d-dimensional:

$$\mathscr{H}_i = \left\{ \mathbf{x} \in \mathbb{R}^d \middle| a_j \leq x_j \leq b_j \right\}, \forall\, j \in \{1, \cdots, d\}, \tag{4.11}$$

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

Usa-se estatísticas extraídas das amostras de classes, ou das relações de paridade obrigatória (*must-link*), para limitar o movimento do centroide ao qual os pontos com vínculos estão associados. A ideia é que essas restrições funcionem como um guia para o algoritmo de agrupamento, na tentativa de encontrar aglomerados melhores ao restringir o movimento dos centroides. O uso de restrições de caixas é uma abordagem inédita e seu objetivo é transformar as restrições de instâncias em restrições espaciais.

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 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*.

Funções Shake **e** NeighbourhoodChange

A função Shake serve para ampliar a diversidade da exploração, ao escolher a solução que será usada para inicializar a busca local (LocalSearch). Com os conceitos de vizinhança $N^{(1)}$, $N^{(2)}$ e $N^{(3)}$ propostos na Seção 4.2, podemos escolher, de forma aleatória, uma solução $\mathbf{C}'$ da i-ésima vizinhança da solução corrente $\mathbf{C}$.

A função NeighbourhoodChange($\mathbf{C}, \mathbf{C}'', i$) serve tão somente para comparar a solução corrente $\mathbf{C}$ com a solução $\mathbf{C}''$ obtida pela busca local. Caso $\mathbf{C}''$ seja pior, o valor atual é mantido e muda-se para o próximo nível de vizinhança. Caso $\mathbf{C}''$ seja melhor, o valor atual é descartado, a nova solução é usada em seu lugar e passa-se a uma nova interação, onde se usa a primeira vizinhança da nova solução.

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. Entretanto, duas regras heurísticas que têm se

Algorithm 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($\mathbf{X}, \mathbf{C}, \Omega_c, k, t_{max}$)
$i_{max} \longleftarrow \frac{n}{10}$
repeat
  $i \leftarrow 1$;
  repeat
    $\mathbf{C}' \leftarrow$ Shake($\mathbf{C}, i$);
    $\mathbf{C}'' \leftarrow$ LocalSearch($\mathbf{X}, k, \mathbf{C}', \Omega_c$);
    $\mathbf{C}, i \leftarrow$ NeighbourhoodChange($\mathbf{C}, \mathbf{C}'', i$)
  until $i = i_{max}$;
  $t \leftarrow$ tempo de processamento
until $t > t_{max}$;

mostrado consistentes são assumir que a pertubação por realocação seja limitada $i_{max} = \frac{n}{2k}$ 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 9 (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 têm-se:*

$$\mathscr{H}_i = \left\{ \mathbf{x} \in \mathbb{R}^d \middle| \mu_j - \frac{3\sigma_j}{\sqrt{n_i}} \leq x_j \leq \mu_j - \frac{3\sigma_j}{\sqrt{n_i}}, \forall\, j \in \{1, \cdots, d\} \right\}. \tag{4.12}$$

Daí, Ω_c é o conjunto das k' restrições, ou seja, $\Omega_c = \{\mathscr{H}_1, \cdots, \mathscr{H}_{k'}\}$.

Uma solução C satisfaz as restrições de caixa Ω_c se, para cada uma das k' restrições, existe ao menos um centroide $c_j \in \mathscr{H}_j$. O princípio por trás dessa forma de usar o conhecimento prévio é que há um conhecimento *a priori* de uma amostra das classes que se quer caracterizar. Portanto, é razoável supor que o centroide da população esteja dentro de uma região de confiança

determinada pela amostra. Disso, propôs-se o uso de uma função de busca local (LocalSearch) que aproveitasse esse conhecimento na forma de restrições por caixas.

Algorithm 7: descrição por passos da busca local com restrições por caixas

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

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): $C' \longleftarrow \text{Projection}(C', C, \Omega_c)$
passo 5 (atualização): trocar os centroides C da solução corrente por C' e voltar para o **passo 2**

A forma geral do método de busca local é apresentada no algoritmo 7, usando-se o conceito de projeção definido pela função $\text{Projection}(C', C, \Omega_c)$. Se o novo conjunto $\mathbf{C}'$ não tem ao menos um centroide em cada uma das caixas, troca-se os centroides infactíveis, $\mathbf{c}_j \notin \Omega_c$, por suas projeções, como descrito pelo algoritmo 8. Para cada centroide infactível, a projeção é um ponto da superfície da caixa que satisfaz a seguinte condição:

Definição 10 (regra de projeção do centroide sobre a caixa)**.** *Seja* $\mathbf{c}_{new} \in C'$ *a atualização infactível de* $\mathbf{c}_{old} \in C$ *e* $\mathscr{H}_i = \left\{\mathbf{x} \in \mathbb{R}^d | a_j \leq x_j \leq b_j, \forall\, j \in \{1, \cdot, d\}\right\}$ *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*

$$\alpha = \min\left\{1, \min_{j|d_j>0}\left\{\frac{b_j - c_j}{d_j}\right\}, \min_{j|d_j<0}\left\{\frac{a_j - c_j}{d_j}\right\}\right\}, \tag{4.13}$$

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 esvaziados, 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].

Algorithm 8: projeção sobre a caixa

$C' \leftarrow \text{Projection}(C', C, \Omega_c)$

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 10)

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

Sendo a solução final da busca local uma solução degenerada, então $\exists k_d > 0$, tal que o número de *clusters* 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 inclusão de uma instrução adicional entre os passos de teste de otimalidade local e de atualização:

passo 3 (teste de otimalidade local)

Se não houve alguma mudança no passo anterior, então

Se a solução corrente não é degenerada, parar aqui,

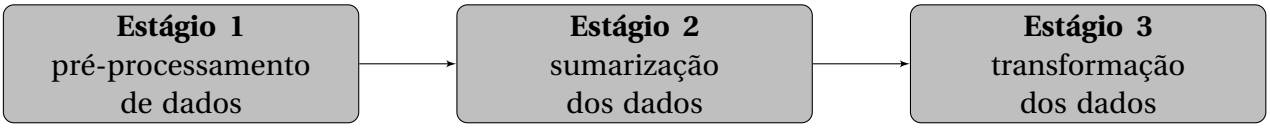
caso contrário, selecionar k_d pontos como centroides e voltar ao **passo 2**

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. E mais que o simples ajustamento dos dados ao método proposto, objetiva-se 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 das culturas a serem identificadas e que mitigue os ruídos decorrentes do processo de captação e das condições climáticas.

Para essas fases, empregou-se uma solução já presente na literatura [1, 34], 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. Este sistema foi capaz de gerar automaticamente imagens com deslocamentos máximos de 1 pixel, valor aceitável para aplicações com dados de baixa resolução espacial.

4.5.1 Transformação dos dados

Os algoritmos de clusterização, em geral, dependem implicitamente da imposição de certas hipóteses a respeito da forma dos *clusters* ou da configuração dos múltiplos *clusters*. Os dados dificilmente estarão 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), a fim de gerar uma salto semântico entre os dados brutos e os dados a serem agrupados. Para isso, lançou-se mão da análise de dados funcionais (*Functional Data Analysis*) 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, em parte, 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, em simultaneamente estimar a velocidade de crescimento ou a aceleração como função do tempo para cada indivíduo. Embora a área de análise funcional de dados seja cativante e de grande interesse da comunidade estatística internacional, ela ainda é incipiente no Brasil, contando com poucos autores nacionais.

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*). Por exemplo, os dados de temperatura na Figura 4.1 foram ajustados por suavização, usando uma série finita de *Fourier*. Essa técnica permite obter informações de alta qualidade sobre as derivadas.

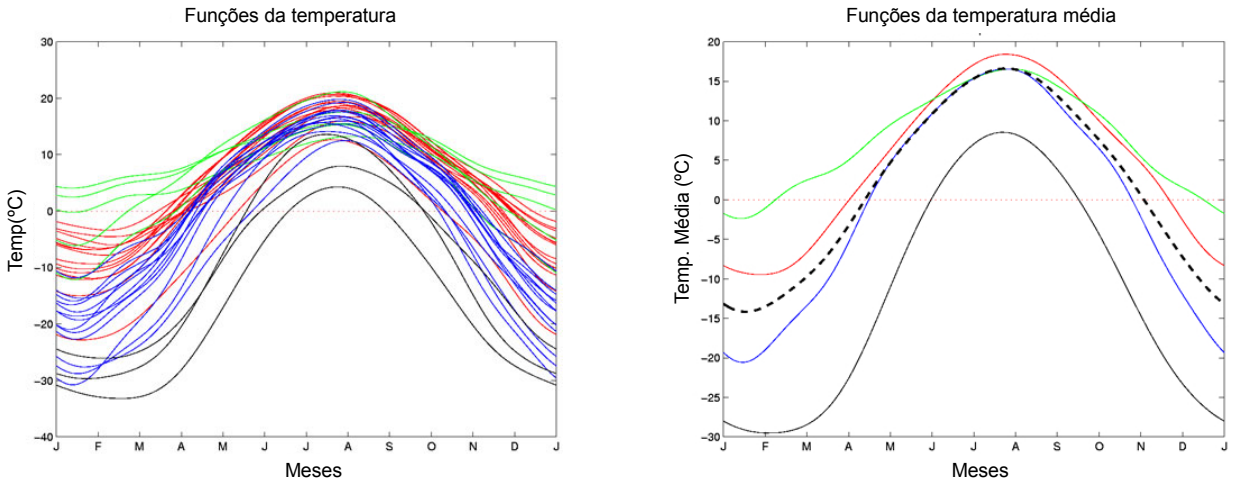


Fig. 4.1: Ajuste de curvas para as temperaturas e médias de temperaturas coletadas mensalmente em quatro estações canadenses de clima: *Montreal*, *Edmonton*, *Pr. Rupert* e *Resolute*.
Fonte: Adaptado de `www.functionaldata.org` em janeiro de 2014.

É razoável pensar em estratégias que transformem os dados dinâmicos em estruturas mais adequadas aos algoritmos, sem perda das características originais. Para os dados temporais, é razoável supor que suas principais informações são derivadas de sua curva. Sendo assim, propôs-se a transformação $\mathscr{F}$, que substitui as séries originais por vetores mais significativos para o desenho da curva dos dados. Mais especificamente, a cada série temporal associou-se

um conjunto de coeficientes derivados do ajuste por uma série de Fourier truncada.

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*) [118].

A transformação $\mathscr{F}$ decompõe um sinal em um número infinito de componentes (harmônicos), onde 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:

$$f(x) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty}\left(a_n \cos\frac{2\pi nx}{L} + b_n \operatorname{sen}\frac{2\pi nx}{L}\right). \tag{4.14}$$

O lado direito da equação 4.14 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:

$$a_n = \frac{2}{L}\int_0^L f(x)\cos\frac{2\pi nx}{L}dx \quad \text{e} \quad b_n = \frac{2}{L}\int_0^L f(x)\operatorname{sen}\frac{2\pi nx}{L}dx \quad \text{com } n \geq 1. \tag{4.15}$$

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. Vamos definir o j-ésimo harmônico como sendo o j-ésimo termo da série de Fourier (para $j \geq 1$), dado por:

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

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

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

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 $[-\frac{\pi}{2}, \frac{\pi}{2}]$, sempre que $a_j < 0$, adotamos a forma modificada $\phi = tg^{-1}\left(\frac{b_j}{a_j}\right) + \pi$. 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 do n-ésimo termo (n-ésimo harmônico) e ϕ_n é o ângulo de fase do n-ésimo termo.

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 onde se substitui as integrais de Riemann, em (4.15), por aproximações trapezoidais. Qualquer série de n pontos pode ser representada exatamente pela expressão:

$$y_t = \bar{y} + \sum_{k=1}^{\lfloor \frac{n}{2} \rfloor} \left[C_k \cos\left(\frac{2\pi kt}{n} - \Phi_k\right)\right] = \tag{4.16}$$

$$= \bar{y} + \sum_{k=1}^{\lfloor \frac{n}{2} \rfloor} \left[A_k \cos\left(\frac{2\pi kt}{n}\right) + B_k \operatorname{sen}\left(\frac{2\pi kt}{n}\right)\right], \tag{4.17}$$

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:

$$A_k = \frac{2}{n}\sum_{t=1}^{n} y_t \cos\left(\frac{2\pi kt}{n}\right) \qquad \text{e} \qquad B_k = \frac{2}{n}\sum_{t=1}^{n} y_t \operatorname{sen}\left(\frac{2\pi kt}{n}\right). \tag{4.18}$$

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 $\Phi_k = tg^{-1}\left(\frac{B_j}{A_j}\right)$.

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)

Em larga escala, é vital que pesquisadores que programem com regularidade reutilizem código ao invés de re-escrevê-lo. A reprodutibilidade é maximizada quando os pesquisadores colocam tudo que criaram manualmente em um sistema de controle de versões, incluindo programas, observações de campo e seus arquivos e artigos. Saídas automáticas e artigos intermediários podem ser gerados a medida que são necessários. Pela crença na importância da reprodutibilidade e continuação do trabalho, disponibilizou-se uma instância virtual da tese, que conta com a relação de todos os programas desenvolvidos em *MatLab* e arquivos de banco de dados. Para tanto usou-se o *git*, um sistema de controle de versão distribuído e gerenciamento de código fonte, em um repositório disponível em `https://github.com/baldez/doctorate`. Todos os códigos fontes e executáveis estão disponíveis sobre a licença *Gnu General Public License* ou *Gnu Lesser General Public Licence*. As licenças aplicadas estão declaradas na documentação do programa.

Os dados de teste estão divididos em dois arquivos: um com a extensão *arf*, contendo os metadados do banco de dados produzido a partir da combinação do cabeçalho*arff* do Weka com as informações disponibilizadas no «<site que tem os bancos de dados»»>., e o arquivo *csv*, contendo os dados propriamente.

5.1 Dados sintéticos

"However, experiments require a lot of work, so the reader may be warned: performing a good experiment is as demanding as proving a new theorem."
- Hans-Paul Schwefel

visualização de agrupamentos

Incluir diagrama de voronoi, envoltório convexo e análise de tragetória central.

avaliação comparativa - *benchmarking*

Tab. 5.1: Performace dos algoritmos na base de dados sintética

parâmetros $t = 10s$; $i_{max} =$				
	algoritmos	B-cubed recall	B-cubed precision	MSSC
1	k-médias	16.128	+8.872	16.128
2	COP k-médias	-2.509	3.442	0.299
3	projected k-médias	-0.363	1.826	0.159
4	*VNS*+ $N^{(1)}$ + k-médias	-0.466	0.641	0.056
5	*VNS*+ $N^{(1)}$ + COP k-médias	0.45	0.039	
6	*VNS*+ $N^{(1)}$ + projected k-médias	-0.597	0.598	0.052
7	*VNS*+ $N^{(2)}$ + k-médias	-0.466	0.641	0.056
8	*VNS*+ $N^{(2)}$ + COP k-médias	0.45	0.039	
9	*VNS*+ $N^{(2)}$ + projected k-médias	-0.597	0.598	0.052
10	*VNS*+ $N^{(3)}$ + k-médias	-0.466	0.641	0.056
11	*VNS*+ $N^{(3)}$ + COP k-médias	0.45	0.039	
12	*VNS*+ $N^{(3)}$ + projected k-médias	-0.597	0.598	0.052

5.2 Base de dados Íris - *Iris Plants*

Esta é uma base de dados bem conhecida que apresenta as medidas em centímetro da largura e comprimento das pétalas e sépalas de três espécies de flor íris (Setosa, Virginica e Versicolor) (Fisher, 1936). Esta base contém 150 amostras e 5 variáveis:

5.3 Aplicações em séries temporais de satélite

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

Descrevemos a seguir o uso do algoritmo proposto sobre a hipótese de que os elementos são pequenas variações de um modelo de comportamento idealizado. Ou seja, a clusterização tem como objetivo classificar séries temporais de *NDVI*a partir de sua similaridade com o que chamamos de **curva característica**, que nada mais é que um **protótipo** de comportamento que o

Tab. 5.2: Performace dos algoritmos na base de dados Íris

parâmetros $t = 10s$; $i_{max} =$				
	algoritmos	B-cubed recall	B-cubed precision	MSSC
1	k-médias	16.128	+8.872	16.128
2	COP k-médias	-2.509	3.442	0.299
3	projected k-médias	-0.363	1.826	0.159
4	*VNS*+ $N^{(1)}$ + k-médias	-0.466	0.641	0.056
5	*VNS*+ $N^{(1)}$ + COP k-médias	0.45	0.039	
6	*VNS*+ $N^{(1)}$ + projected k-médias	-0.597	0.598	0.052
7	*VNS*+ $N^{(2)}$ + k-médias	-0.466	0.641	0.056
8	*VNS*+ $N^{(2)}$ + COP k-médias	0.45	0.039	
9	*VNS*+ $N^{(2)}$ + projected k-médias	-0.597	0.598	0.052
10	*VNS*+ $N^{(3)}$ + k-médias	-0.466	0.641	0.056
11	*VNS*+ $N^{(3)}$ + COP k-médias	0.45	0.039	
12	*VNS*+ $N^{(3)}$ + projected k-médias	-0.597	0.598	0.052

*NDVI*assumiria durante a evolução fenológica da cultura se não ouvesse nenhuma tipo de ruído na captação do sinal.

De forma mais específica, temos que para cada curva desenhada pela interpolação da série temporal no sistema de coordenadas paralelas, temos associado um vetor de características que define os coeficientes desta curva em uma espação de Fourier. Sendo assim temos que,

Definição 11 (curva característica de uma cultura)**.** *É a curva obtida pelo uso da base de Fourier, cujo os coeficientes são determinados pelo vetor centroide dos pontos que foram associados a cultura.*

Vamos admitir que uma imagem de satélite é bem caracterizada pela região R que ela contempla, número de bandas espectrais k e intervalo de tempo de revisita δt. Admitimos também que a região R está ajustada sobre um retângulo de dimensões $i \times j$, que necessariamente leva em conta a resolução espacial do sensor. Sendo assim, propomos a seguinte definição,

Definição 12 (Imagem de SR)**.** *Uma imagem de SR, em um dado instante t, é uma matriz $\mathcal{M}_t$ de dimensões $i \times j$, onde os elementos da matriz são vetores $\mathbf{b} \in \Omega_1$, onde k é número de bandas espectrais contidas no sensor que gerou a imagem. Para cada coordenada, o vetor $\mathbf{b}$ é k-upla contendo as quantificações*[1] *da capacidade de reflexão e de emissão de energia eletromagnética dos alvos contemplados pelo pixel correspondente a coordenada $(\bar{i}, \bar{j})$ de $\mathcal{M}_t$.*

O intervalo de revisita δt é o tempo necessário para o satélite produzir outra imagem contemplando a mesma região de interesse. Dessa forma os objetos de entrada para o que

[1] Os valores das coordenadas de $\mathbf{b}$ são dependentes da resolução radiométrica do sensor.

chamaremos de problema de classificação de dados de SR (PCSR) serão estruturas da forma,

$$\{\mathscr{M}_{t_1}, \mathscr{M}_{t_2}, ..., \mathscr{M}_{t_n}\} = S_T^n \in S \tag{5.1}$$

onde S_T^n é uma sequência temporal de tamanho n de imagens obtidas no períodos $T = \{t_1, ..., t_n$, onde assumimos $t_i = t_1 + i\delta t$, e S é a σ-álgebra do conjunto das imagens. A partir de S_T^n, para cada pixel da imagem, referente a coordenada $(\bar{i}, \bar{j})$ temos a sequência,

$$\{\mathbf{b}_{t_1}, \mathbf{b}_{t_2}, ..., \mathbf{b}_{t_n}\} = P_{\bar{i}\bar{j}} \tag{5.2}$$

A partir dessas definições vamos propor que o PCSR como um problema de dois estágios:

i. construção do extrator,

ii. construção do classificador.

A construção do extrator significa definir uma metodologia ψ que transforma uma sequência $P_{\bar{i}\bar{j}}$ em um *vetor de características* $\mathbf{c}_{\bar{i}\bar{j}}$.

Para os propósitos deste trabalho, o problema tratado pode ser entendido como um problema geral de agrupamento k (onde o número de grupos é conhecido previamente) de d-uplas reais, onde as d-uplas são séries temporais discretas, univariadas ($\gamma(t) \in \Re$) e com espaço de tempo equidistante. Acrescenta-se que o conjunto **X** de todas a séries a serem agrupadas caracterizam uma base mais latitudinal que longitudinal ou seja $||\mathbf{X}|| >> d$. Daí não é necessário mitigar o fenômeno de Hughes (também conhecido como maldição da dimensionalidade - *curse of dimensionality*)

O problema de classificação automática ou semi-automática, pelo uso de séries completas e parciais de imagens de satélite foi abordado usando métodos de pré-processamento e correções (ver figura **??**), encontrados na literatura, que mantivessem o problema de classificação de séries temporais de imagens de satélite como um equivalente do problema mais geral de agrupamento de séries temporais, onde não há preocupação de inovar na metodologias de remoções e correções de distorções por intemperes climáticas.

5.3.1 Pré-processamento

correção e geração de imagens pelo NAVPRO

Georreferenciamento

Correção geométrica

Correção atmosférica

Correção radiométrica

composição de valor máximo (*mvc*)

O *NDVI* de um pixel pode ser distorcido por condições atmosféricas, uma vez que nuvens podem diminuir seu valor ou mesmo torná-lo negativo. Para mitigar este problema, um método amplamente adotado é a composição por máximos valores (*mvc*, do inglês: *maximum value composition*). Esse método consiste em usar o valor máximo de um pixel, de uma sequência de imagens, para construir uma única imagem final 5.1.

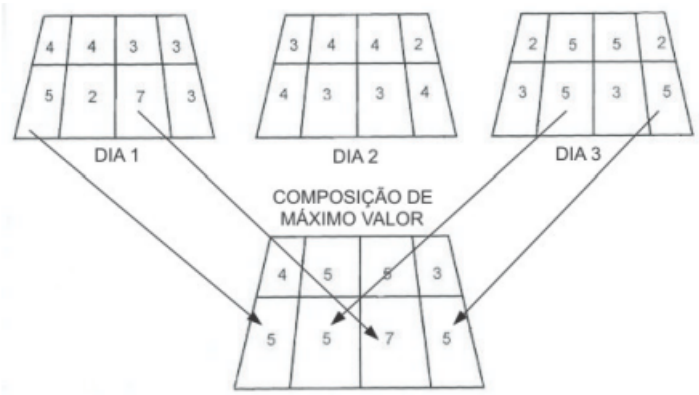


Fig. 5.1: Elaboração de uma composição de máximo valor a partir de dados diários.
Fonte: Embrapa Informática Agropecuária. Comunicado técnico, 107.

O grande volume de imagens, com georreferenciamento preciso, permite construir composições a partir de séries temporais de imagens, assim pixels afetados por estes efeitos de distorção têm menos probabilidade de fazer parte da composição final. Imagens *NDVI*/*mvc* têm redução nos impactos da reflectância direcional, efeitos de sombra e efeitos de partículas no ar [citar HOLBEN; Caracteristics of maximum value composite images from temporal AVHRR data; 1986]. ⟸
Quanto maior a série, melhor será a qualidade da atenuação, entretanto deve-se observar que o uso de períodos muito longos comprometem a análise da dinâmica de cobertura da área investigada.

extração de séries temporais de NDVI

Usou-se a área agrícola referente a cidade de jabotical, no estado de São Paulo, Brasil, entre os períodos de abril/2004 e março/2006. Para cada ano safra, as série de NDVI variaram desde o início do plantio, passando pelo período de maior vigor vegetativo e indo até o final da colheita.

Roughly speaking, temos para cada cada pixel, de uma área a ser classificada, uma série temporal de NDVI é ajustada por um polinômio trigonométrico, que por sua vez é definido ao obtermos os valores de seus coeficientes. Esses coeficientes são comparados com os coeficientes de uma curva característica média da cana-de-açúcar, definida previamente através de pixels de controle.

Para se aproximar da série, pode-se utilizar qualquer base, não necessariamente a de senos e cossenos. Os coeficientes encontrados nessa aproximação serão as *características* dessa série temporal, i.e.

Definition 1 características
Características de uma série temporal relacionada a uma base, são os coeficientes encontrados ao se ter a melhor aproximação possível dessa série por elementos dessa base ou parte dela, quando se tratar de uma representação truncada.

Para classificar um pixel, usou-se a função harmônica da série de NDVI deste pixel. Após extrair suas características, faz-se a comparação destas com as caracteristicas médias de um grupos de controle. Se essa diferença for menor que uma certa tolerância, dizemos que o pixel é pertencente a classe cana-de-açúcar.

Acrescenta-se que o conjunto de entrada $\mathbf{X}$, de todas a séries a serem agrupadas, caracterizam uma base mais latitudinal que longitudinal, ou seja $||\mathbf{X}|| >> d$. Quando temos uma quantidade de instâncias muito maior que o número de dimensões, não é necessário mitigar o fenômeno de Hughes também conhecido como maldição da dimensionalidade (*curse of dimensionality*).

5.3.2 Dados AVHRR/NOAA - Jaboticabal 2004/2005

A classificação binária denotada por y e no caso mais simples assume valores no conjunto binário $Y = \{-1, 1\}$. Neste trabalho, vamos nos restringir apenas a classificação binária.

Embora já existam disponíveis sensores de maior qualidade espectral e espacial, como o MODIS, SPOT/VEGETATION e WFI, as imagens AVHRR-NOAA continuam sendo de grande uso em estudos envolvendoa análise de ecossistemas, em função da disponibilidade de longas séries temporais deimagens pelo grande acervo histórico de dados NOAA. O Centro de Pesquisas Meteorológicas e Climáticas Aplicadas à Agricultura (CEPAGRI) da Universidade Estadual de Campinas (UNICAMP) possui um banco de imagens AVHRRNOAA iniciado em abril de 1995, com aproximadamente dois terabytes dedados. Atualmente são recebidas e armazenadas em média 10 imagens por dia dos satélites NOAA 12, 14, 15, 16 e 17 atualmente em operação. Ainda que

apresentem resolução espacial e espectral menores em comparação a outros sensores, as imagens AVHRR-NOAA possuem características fundamentais no estudo de

Alvos com grande dinâmica espectral, tais como sua alta resolução temporal(coberturadiária), garantia de cobertura global e gratuidade das imagens. O elevado número de imagens diariamente disponíveis pelos satélites hoje em operação torna a geração de produtos derivados do sensor AVHRR uma rotina trabalhosa e com um certo nível de intervenção humana. Uma etapa fundamental para a geração de produtos é a correção geométrica precisa das imagens, o que nem sempre é exeqüível por meio dos softwares comuns atualmente disponíveis. Diante dessas dificuldades e da necessidade de se gerar produtos com qualidade de forma automática, o presente trabalho aborda o desenvolvimento de um sistema para o processamento completo das imagens AVHRRNOAA, sem qualquer intervenção humana.

O primeiro grande projeto de mapeamento de área de cana-de-açúcar no estado de São Paulo ocorreu no ano safra 1979/80 (Mendonça et al.,1981). Este projeto foi realizado através de interpretação visual de imagens Landast-MSS (Multispectral Scanner System) na escala 1:250.000, nas bandas MSS5 (vermelho) e MSS7 (infravermelho próximo). Foram utilizadas 44 imagens de 185x185 km obtendo-se uma estimativa de área plantada de 801.950 ha, para todo Brasil. Posteriormente, foram realizados estudos que visaram estimar a produtividade agrícola da cana-de-açúcar através de imagens do Landsat e de um modelo agrometeorológicodesta (Rudorff e Batista, 1990).

visualização de agrupamentos

A seguir temos a representação por coordenadas paralelas das curvas classificadas antes e depois das transformações.

avaliação comparativa - *benchmarking*

5.3.3 Dados TERRA/MODIS - Mato Grosso 2008/2009

avaliação comparativa - *benchmarking*

Tab. 5.3: Performace dos algoritmos na base de dados brutos

parâmetros $t = 10s$; $i_{max} =$				
	algoritmos	B-cubed recall	B-cubed precision	MSSC
1	k-médias	16.128	+8.872	16.128
2	COP k-médias	-2.509	3.442	0.299
3	projected k-médias	-0.363	1.826	0.159
4	*VNS*+ $N^{(1)}$ + k-médias	-0.466	0.641	0.056
5	*VNS*+ $N^{(1)}$ + COP k-médias	0.45	0.039	
6	*VNS*+ $N^{(1)}$ + projected k-médias	-0.597	0.598	0.052
7	*VNS*+ $N^{(2)}$ + k-médias	-0.466	0.641	0.056
8	*VNS*+ $N^{(2)}$ + COP k-médias	0.45	0.039	
9	*VNS*+ $N^{(2)}$ + projected k-médias	-0.597	0.598	0.052
10	*VNS*+ $N^{(3)}$ + k-médias	-0.466	0.641	0.056
11	*VNS*+ $N^{(3)}$ + COP k-médias	0.45	0.039	
12	*VNS*+ $N^{(3)}$ + projected k-médias	-0.597	0.598	0.052

Tab. 5.4: Performace dos algoritmos na base de dados transformados

parâmetros $t = 10s$; $i_{max} =$				
	algoritmos	B-cubed recall	B-cubed precision	MSSC
1	k-médias	16.128	+8.872	16.128
2	COP k-médias	-2.509	3.442	0.299
3	projected k-médias	-0.363	1.826	0.159
4	*VNS*+ $N^{(1)}$ + k-médias	-0.466	0.641	0.056
5	*VNS*+ $N^{(1)}$ + COP k-médias	0.45	0.039	
6	*VNS*+ $N^{(1)}$ + projected k-médias	-0.597	0.598	0.052
7	*VNS*+ $N^{(2)}$ + k-médias	-0.466	0.641	0.056
8	*VNS*+ $N^{(2)}$ + COP k-médias	0.45	0.039	
9	*VNS*+ $N^{(2)}$ + projected k-médias	-0.597	0.598	0.052
10	*VNS*+ $N^{(3)}$ + k-médias	-0.466	0.641	0.056
11	*VNS*+ $N^{(3)}$ + COP k-médias	0.45	0.039	
12	*VNS*+ $N^{(3)}$ + projected k-médias	-0.597	0.598	0.052

Tab. 5.5: Performace dos algoritmos na base de dados brutos

parâmetros $t = 10s$; $i_{max} =$				
	algoritmos	B-cubed recall	B-cubed precision	MSSC
1	k-médias	16.128	+8.872	16.128
2	COP k-médias	-2.509	3.442	0.299
3	projected k-médias	-0.363	1.826	0.159
4	*VNS*+ $N^{(1)}$ + k-médias	-0.466	0.641	0.056
5	*VNS*+ $N^{(1)}$ + COP k-médias	0.45	0.039	
6	*VNS*+ $N^{(1)}$ + projected k-médias	-0.597	0.598	0.052
7	*VNS*+ $N^{(2)}$ + k-médias	-0.466	0.641	0.056
8	*VNS*+ $N^{(2)}$ + COP k-médias	0.45	0.039	
9	*VNS*+ $N^{(2)}$ + projected k-médias	-0.597	0.598	0.052
10	*VNS*+ $N^{(3)}$ + k-médias	-0.466	0.641	0.056
11	*VNS*+ $N^{(3)}$ + COP k-médias	0.45	0.039	
12	*VNS*+ $N^{(3)}$ + projected k-médias	-0.597	0.598	0.052

Tab. 5.6: Performace dos algoritmos na base de dados transformados

parâmetros $t = 10s$; $i_{max} =$				
	algoritmos	B-cubed recall	B-cubed precision	MSSC
1	k-médias	16.128	+8.872	16.128
2	COP k-médias	-2.509	3.442	0.299
3	projected k-médias	-0.363	1.826	0.159
4	*VNS*+ $N^{(1)}$ + k-médias	-0.466	0.641	0.056
5	*VNS*+ $N^{(1)}$ + COP k-médias	0.45	0.039	
6	*VNS*+ $N^{(1)}$ + projected k-médias	-0.597	0.598	0.052
7	*VNS*+ $N^{(2)}$ + k-médias	-0.466	0.641	0.056
8	*VNS*+ $N^{(2)}$ + COP k-médias	0.45	0.039	
9	*VNS*+ $N^{(2)}$ + projected k-médias	-0.597	0.598	0.052
10	*VNS*+ $N^{(3)}$ + k-médias	-0.466	0.641	0.056
11	*VNS*+ $N^{(3)}$ + COP k-médias	0.45	0.039	
12	*VNS*+ $N^{(3)}$ + projected k-médias	-0.597	0.598	0.052

Capítulo 6

Conclusões e Trabalhos Futuros

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

A metodologia proposta mostrou-se eficiente com alta taxa de êxito para a classificação da cultura de cana-de-açucar em imagens AVHRR/NOAA sobre áreas agrícolas em nível municipal. Com vantagem sobre formas de classificação baseadas em máscaras pois é capaz de detectar cana de expansão.

As aplicações geram visibilidade a uma área. A formulação de categorias básicas dos métodos de *DM*como problemas de otimização evidenciam as oportunidades para contribuições pela comunidade de pesquisadores da área de otimização.

Trabalhos futuros

"Stay Hungry. Stay Foolish."
- The Whole Earth Catalog - Stewart Brand

Apesar de usar a expressão trabalhos futuros, o que se fez na verdade foi listar os diversos pontos em que cabem melhoras e propostas no contexto dos tópicos abordados neste trabalho.

(a) A base trigonométrica não é necessáriamente a única, podendo ser feitos testes com wavelets e Curvelets;

(b) A distância Euclidiana talvez pudesse ser substituída por uma métrica mais coniente que permitisse definir pertinências intermediarias do pixel a classe de cana (Problema de mistura espectral);

(c) Avaliar Modelo de Mistura espectral

(d) Avaliação da fenologia da cultura pelo método da curva característica permitiria trabalhar com previsões de safra;

(e) Testar outros índices de vegetação como o NDMI. WILSON e SADER (2002) aplicaram o pouco conhecido Normalized Difference Moisture Index (NDMI) para estudar a influência hídrica em vegetações e concluíram que o NDMI explicou melhor a dinâmica da vegetação do que o NDVI, durante a análise de uma série multitemporal de imagens. O NDMI utiliza dados da faixa do infravermelho médio que é sensível à umidade por causa das regiões de absorção de água. Essa banda também é menos influenciada pelos efeitos de absorção e espalhamento dos aerossóis e vapor de água presentes na atmosfera devido ao tamanho do comprimento de onda. Em função disso, pode indicar a presença de umidade na vegetação e no solo permitindo obter um maior contraste entre tipos diferentes de vegetação, refletindo melhor as mudanças de biomassa que o NDVI.

(f) Incluir testes sobre vizinhanças geradas a partir de pertubações sobre os centroides ou controlar as pertubações de partições pelo peso do impacto nos centroides.

(g) Segmentação de imagens é uma tarefa básica no processo de análise de imagens: 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 um dado atributo. Atributos das regiões tais como área, forma, parâmetros estatísticos e textura podem ser extraídos e usados posteriormente no processo de análise. (ou seja, incluir contexto espacial nas restrições).

(h) Ampliar o uso de dados com estrutura funcional como uso de operadores diferenciasi. Functional Data Analysis with R and Matlab; Ramsay-Giles Hooker - Spencer Graves http://www.psych.mcgill.ca/misc/fda/index.html

(i) Abordagens que levem em conta outras estruturas de vizinhança como estratégias para dados multidimensionais baseados em distância [citar Knorr, E. M. and Ng, R. T.: 1998, Algorithms for Mining Distance-Based Outliers in Large Datasets. In:Proceedings of the VLDB Conference. New York, USA, pp. 392-403] e densidade [citar Markus Breunig and Hans-Peter Kriegel and Raymond T. Ng and Jörg Sander: 2000, LOF: Identifying Density-Based Local Outliers. In: Proceedings of the ACM SIGMOD Conference. pp. 93-104].

(j) Testar outros esquemas de *VNS*. Depois de selecionarmos x' aleatoriamente, podemos separar algumas características y que distingue x' de x, e fazer uma busca local levando em conta apenas no espaço y (*VNS*decomposto). Geralmente, usa-se a própria VNS para fazer essa

busca. Neste caso, teríamos a *VNS*em dois níveis (*bi-level VNS*). Pensar também a respeito do uso do primal dual, reduced ou outros esquemas.

(k) Avaliar aspectos como ajustamento dos algoritmos para bons desempenhos em bancos de dados grandes (tornar o algoritmo escalável).

(l) Pensar sobre o uso de hiperheurísticas (ver final da página 18 do artigo http://www.ime.usp.br/ igorrs/monografias/metahiper.pdf e página 36 da conlusão).

(m) Pensar sobre a comparação com GA, busca tabu e outros. (ver artigo operation research and data mining - pag 6).

(n) Medir os limites entre os níveis de qualidade obtiveis pela clusterização semissupervisionada e classificação.

(o) Definir métodos que verifiquem a qualidade das restrições como a robustez do método a inconsistência, ou incoerência de restrições por instância (ver Wagstaff2006). Testar também outros tipos de restrições.

Referências Bibliográficas

[1] *Desenvolvimento de um sistema automático para a geração de produtos derivados de imagens AVHRR-NOAA*, 2005.

[2] E. M. ABDEL-RAHMA AND F. B. AHMED, *The application of remote sensing techniques to sugarcane (Saccharum spp. hybrid) production: a review of the literature*, International Journal of Remote Sensin, 29 (2008), pp. 3753–3767.

[3] S. AHMED, *Applications of data mining in retail business*, International Conference on Information Technology: Coding Computing, ITCC, 2 (2004), pp. 455–459.

[4] A. ALGUWAIZANI, *Variable neighbourhood search based heuristic for K-harmonic means clustering*, School of Information Systems, Computing and . . . , (2011).

[5] E. AMIGÓ, J. GONZALO, J. ARTILES, AND F. VERDEJO, *A comparison of extrinsic clustering evaluation metrics based on formal constraints*, Information retrieval, 12 (2009), pp. 461–486.

[6] M. ANKERST, M. BREUNIG, H. KRIEGEL, AND J. SANDER, *OPTICS: ordering points to identify the clustering structure*, ACM SIGMOD Record, (1999).

[7] J. A. F. G. ANTUNES, *Aplicação de lógica fuzzy para estimativa de área plantada da cultura de soja utilizando imagens AVHRR-NOAA.*, PhD thesis, (Mestrado)-Universidade Estadual de Campinas, 2005.

[8] J. A. F. G. ANTUNES AND J. C. D. M. ESQUERDO, *NAVPRO 3.0: Tutorial de Instalação e Utilização*, 2008.

[9] ——, *Monitoramento agrícola usando análise harmônica de séries temporais de dados NDVI/AVHRR-NOAA.*, 2009.

[10] A. BAGGA AND B. BALDWIN, *Entity-based cross-document coreferencing using the vector space model*, in Proceedings of the 17th international conference on Computational linguistics-Volume 1, Association for Computational Linguistics, 1998, pp. 79–85.

[11] J. BAKUS, M. F. HUSSIN, AND M. KAMEL, *A SOM-based document clustering using phrases*, in Neural Information Processing, 2002. ICONIP'02. Proceedings of the 9th International Conference on, vol. 5, IEEE, 2002, pp. 2212–2216.

[12] G. BALL AND D. HALL, *ISODATA, a novel method of data analysis and pattern classification*, (1965).

[13] ——, *A clustering technique for summarizing multivariate data*, Behavioral science, 12 (1967), pp. 153–155.

[14] A. BAR-HILLEL AND T. HERTZ, *Learning a mahalanobis metric from equivalence constraints*, ... of Machine Learning ..., 6 (2005), pp. 937–965.

[15] S. BASU, A. BANERJEE, AND R. MOONEY, *Semi-supervised clustering by seeding*, ICML, (2002), pp. 19–26.

[16] N. BELACEL, P. HANSEN, AND N. MLADENOVIC, *Fuzzy j-Means: a new heuristic for fuzzy clustering*, Pattern Recognition, 35 (2002), pp. 2193–2200.

[17] M. BILENKO, S. BASU, AND R. MOONEY, *Integrating constraints and metric learning in semi-supervised clustering*, Proceedings of the twenty-first century, (2004).

[18] A. BLUM AND T. MITCHELL, *Combining labeled and unlabeled data with co-training*, Proceedings of the eleventh annual conference on ..., (1998).

[19] C. BLUM AND A. ROLI, *Metaheuristics in combinatorial optimization: Overview and conceptual comparison*, ACM Computing Surveys (CSUR), (2003).

[20] C. BORGELT, *Prototype-based classification and clustering*, (2006).

[21] E. BOROS, P. HAMMER, AND T. IBARAKI, *An implementation of logical analysis of data*, Knowledge and Data ..., 12 (2000), pp. 292–306.

[22] P. BRADLEY, *Constrained k-means clustering*, Microsoft ..., (2000).

[23] P. BRADLEY AND U. FAYYAD, *Refining Initial Points for K-Means Clustering.*, ICML, (1998).

[24] E. BREDENSTEINER AND K. BENNETT, *Multicategory classification by support vector machines*, Computational Optimization, 12 (1999), pp. 53–79.

[25] M. BREUNIG, H. KRIEGEL, R. NG, AND J. SANDER, *LOF: identifying density-based local outliers*, ACM Sigmod Record, (2000), pp. 93–104.

[26] P. BRUCKER, *On the complexity of clustering problems*, Optimization and operations research, 157 (1978), pp. 45–54.

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

[28] S. DAS, A. ABRAHAM, AND A. KONAR, *Metaheuristic Clustering*, vol. 1, Springer, Mar. 2009.

[29] A. DEMPSTER, N. LAIRD, AND D. RUBIN, *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the Royal Statistical Society., (1977).

[30] B. E. DOM, *An information-theoretic external cluster-validity measure*, in Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc., 2002, pp. 137–145.

[31] M. DORIGO, *Optimization, learning and natural algorithms*, Ph. D. Thesis, Politecnico di Milano, Italy, (1992).

[32] O. DU MERLE, P. HANSEN, B. JAUMARD, N. MLADENOVIC, AND O. D. MERLE, *An interior point algorithm for minimum sum-of-squares clustering*, SIAM Journal on Scientific ..., 21 (1999), pp. 1485–1505.

[33] R. EBERHART AND J. KENNEDY, *A new optimizer using particle swarm theory*, Proceedings of the Sixth International Symposium on Micromachine and Human Science, (1995), pp. 39–43.

[34] J. C. D. M. E. ESQUERDO, J. F. G. ANTUNES, D. G. BALDWIN, W. J. EMERY, AND Z. J. JÚNIOR, *An automatic system for AVHRR land surface product generation*, International Journal of Remote Sensing, 27 (2006), pp. 3925–3942.

[35] V. ESTIVILL-CASTRO, *Why so many clustering algorithms: a position paper*, ACM SIGKDD Explorations Newsletter, 4 (2002), pp. 65–75.

[36] G. FELICI AND K. TRUEMPER, *A minsat approach for learning in logic domains*, INFORMS Journal on computing, 14 (2002), pp. 20–36.

[37] T. FEO AND M. RESENDE, *Greedy randomized adaptive search procedures*, Journal of global optimization, 42 (1995), pp. 32–37.

[38] D. FIGUEIREDO, *Conceitos Básicos de Sensoriamento Remoto*, São Paulo, (2005).

[39] L. FONSECA, *Processamento digital de imagens*, INPE - Instituto de Pesquisas Espaciais, (2000).

[40] A. GAMMERMAN, V. VOVK, AND V. VAPNIK, *Learning by transduction*, Proceedings of the Fourteenth ..., (1998).

[41] M. GENDREAU AND J. POTVIN, *Handbook of metaheuristics*, (2010), pp. 61–86.

[42] F. GLOVER, *Future paths for integer programming and links to artificial intelligence*, Computers & Operations Research, 13 (1986), pp. 533–549.

[43] ——, *Improved Linear Programming Models for Discriminant Analysis**, Decision Sciences, 21 (1990), pp. 771–785.

[44] C. GOUTTE, P. TOFT, E. ROSTRUP, F. NIELSEN, AND L. HANSEN, *On clustering fMRI time series*, NeuroImage, (1999).

[45] G. GUYOT, *Signatures spectrales des surfaces naturelles*, (1989).

[46] G. HAMERLY AND C. ELKAN, *Alternatives to the k-means algorithm that find better clusterings*, Proceedings of the eleventh international conference on Information and knowledge management, (2002).

[47] P. HANSEN AND B. JAUMARD, *Cluster analysis and mathematical programming*, Mathematical Programming, 79 (1997), pp. 191–215.

[48] P. HANSEN, N. MLADENOVIC, AND E. HAUTES, *j-means : a new local search heuristic for minimum sum of squares clustering*, 34 (2001).

[49] S. HARMS, D. LI, J. DEOGUN, AND T. TADESSE, *Efficient rule discovery in a geo-spatial decision support system*, Proceedings of the 2002 annual ..., (2002), pp. 1–7.

[50] R. J. HATHAWAY AND J. C. BEZDEK, *Optimization of Clustering Criteria by Reformulation*, IEEE, 1995.

[51] M. HILL AND G. DONALD, *Estimating spatio-temporal patterns of agricultural productivity in fragmented landscapes using AVHRR NDVI time series*, Remote Sensing of Environment, (2003).

[52] A. HINNEBURG AND D. KEIM, *An efficient approach to clustering in large multimedia databases with noise*, KDD, (1998).

[53] R. HOFFER, *Biological and physical considerations in applying computer-aided analysis techniques to remote sensor data*, Remote sensing: The quantitative approach, (1978).

[54] B. HOLBEN, *Characteristics of maximum-value composite images from temporal AVHRR data*, International Journal of Remote Sensing, (1986).

[55] B. HOLBEN, C. TUCKER, AND C. FAN, *Spectral assessment of soybean leaf area and leaf biomass.*, Photogrammetric Engineering and . . . , (1980).

[56] J. HOLLAND, *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence.*, (1975).

[57] R. HOWARD, *Classifying a population into homogeneous groups*, Operational Research in the Social Sciences., (1966).

[58] IBGE, *Levantamento Sistemático da Produção Agrícola*, 2013.

[59] A. INSELBERG, *Multidimensional detective*, Information Visualization, 1997. Proceedings., . . . , (1997).

[60] G. IPPOLITI-RAMILO, J. EPIPHANIO, Y. SHIMABUKURO, AND A. FORMAGGIO, *Sensoriamento remoto orbital como meio auxiliar na previsão de safras, . . .* em São Paulo, (1999).

[61] M. E. JAKUBAUSKAS AND D. R. LEGATES, *Harmonic analysis of time-series AVHRR NDVI data for characterizing US Great Plains land use/land cover*, INTERNATIONAL ARCHIVES OF PHOTOGRAMMETRY AND REMOTE SENSING, 33 (2000), pp. 384–389.

[62] M. E. JAKUBAUSKAS, D. R. LEGATES, AND J. H. KASTENS, *Harmonic Analysis of Time-Series AVHRR NDVI Data*, Photogrammetric Engineering Remote Sensing, 67 (2001), pp. 461–470.

[63] ——, *Crop identification using harmonic analysis of time-series AVHRR NDVI data*, Computers and Electronics in Agriculture, 37 (2002), pp. 127–139.

[64] R. JANCEY, *Multidimensional group analysis*, Australian Journal of Botany, 14 (1966), pp. 127–130.

[65] J. P. JIAWEI HAN MICHELINE KAMBER, *Data Mininng: Concepts and Tecniques*, Morgan KaufMann Publishers, 3st, ed., 2011.

[66] P. JÖNSSON AND L. EKLUND, *TIMESAT - A program for analyzing time-series of satellite sensor data*, Computers & Geosciences, (2004).

[67] C. JORDAN, *Derivation of leaf-area index from quality of light on the forest floor*, Ecology, (1969).

[68] J. R. JR AND R. HAAS, *Monitoring vegetation systems in the Great Plains with ERTS*, NASA special ..., 1 (1974), pp. 309–317.

[69] Y. KAKIZAWA, *Discrimination and clustering for multivariate time series*, Journal of the American ..., (1998).

[70] G. KARYPIS, E. HAN, AND V. KUMAR, *Chameleon: Hierarchical clustering using dynamic modeling*, Computer, (1999).

[71] R. P. KAUFMAN L, *Clustering by means of medoids*, In Statistical Analysis Based Upon the L1 Norm - Edited by Dodge Y. Amsterdam, (1987), pp. 405–416.

[72] S. KIRKPATRICK, D. JR., AND M. VECCHI, *Optimization by simmulated annealing*, science, volume 220 (1983), pp. 671–680.

[73] E. KNOX AND R. NG, *Algorithms for mining distance-based outliers in large datasets*, ...of the International Conference on Very Large Data ..., (1998), pp. 392–403.

[74] F. N. KOUMBOULIS, M. P. TZAMTZI, AND M. PAVLOVIC, *Decision support systems in agribusiness*, 2006 IEEE International Conference on Mechatronics, (2006), pp. 457–461.

[75] M. KUMAR, N. PATEL, AND J. WOO, *Clustering seasonality patterns in the presence of errors*, Proceedings of the eighth ACM SIGKDD ..., (2002).

[76] S. LLOYD, *Least squares quantization in PCM*, Information Theory, IEEE Transactions on, 28 (1982), p. 129?137.

[77] U. V. LUXBURG, *A tutorial on spectral clustering*, Statistics and computing, (2007).

[78] J. MACQUEEN, *Some methods for classification and analysis of multivariate observations*, Proceedings of the fifth Berkeley symposium on ..., 2 (1967), pp. 281–297.

[79] E. MAHARAJ, *Cluster of time series*, Journal of Classification, (2000).

[80] O. MANGASARIAN, *Linear and nonlinear separation of patterns by linear programming*, Operations research, 13 (1965), pp. 444–452.

[81] ——, *Misclassification minimization*, Journal of Global Optimization, 5 (1994), pp. 309–323.

[82] X. X. MARTIN ESTER, HANS-PETER KRIEGEL, JÖRG S, *A density-based algorithm for discovering clusters in large spatial databases with noise*, (1996).

[83] M. MEILUA, *Comparing clusterings by the variation of information*, in Learning theory and kernel machines, Springer, 2003, pp. 173–187.

[84] P. MENESES AND J. M. NETO, *Fundamentos de radiometria óptica espectral, ...* Remoto: reflectância de alvos espectrais. ..., (2001).

[85] N. MLADENOVIC AND P. HANSEN, *Variable neighborhood search*, Computers & Operations Research, 24 (1997), pp. 1097–1100.

[86] C. MÖLLER-LEVET AND F. KLAWONN, *Fuzzy clustering of short time-series and unevenly distributed sampling points*, Advances in Intelligent ..., (2003).

[87] R. E. A. R. MOUSTAFA AND E. J. E. WEGMAN, *On some generalizations of parallel coordinate plots*, in Seeing a Million–A Data Visualization Workshop, Rain am Lech, Germany, vol. 2, 2002, pp. 1–18.

[88] H. MÜHLENBEIN AND 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), pp. 178–187.

[89] C. R. 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*, PhD thesis, Unicamp - Universidade Estadual de Campinas, 2010.

[90] K. NIGAM, A. MCCALLUM, AND T. MITCHELL, *Semi-supervised text classification using EM*, Semi-Supervised Learning, (2006).

[91] S. OLAFSSON, X. LI, AND S. WU, *Operations research and data mining*, European Journal of Operational Research, 187 (2008), pp. 1429–1448.

[92] D. PICCOLO, *A distance measure for classifying ARIMA models*, Journal of Time Series Analysis, (1990), pp. 153–163.

[93] J. PUCHINGER AND G. RAIDL, *Combining metaheuristics and exact algorithms in combinatorial optimization: A survey and classification*, Artificial intelligence and knowledge engineering ..., (2005).

[94] J. O. RAMSAY, G. HOOKER, AND S. GRAVES, *Functional Data Analysis with R and MATLAB*, ece.uvic.ca, (2009).

[95] I. RECHENBERG, *Cybernetic solution path of an experimental problem*, (1965).

[96] G. J. ROERINK, M. MENENTI, AND W. VERHOEF, *Reconstructing cloudfree NDVI composites using Fourier analysis of time series*, International Journal of Remote Sensing, 21 (2000), pp. 1911–1917.

[97] L. ROMANI, *Integrating Time Series Mining and Fractals to Discover Patterns and Extreme Events in Climate and Remote Sensing Databases*, (2010).

[98] A. ROSENBERG AND J. HIRSCHBERG, *V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure.*, in EMNLP-CoNLL, vol. 7, 2007, pp. 410–420.

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

[100] T. SASAO AND M. FUJITA, eds., *Representations of Discrete Functions*, Springer US, Boston, MA, 1996.

[101] R. SCHOWENGERDT, *Techniques for image processing and classification in remote sensing*, (1983), p. 249.

[102] C. SHAW AND G. KING, *Using cluster analysis to classify time series*, Physica D: Nonlinear Phenomena, (1992).

[103] Y. SHI AND R. EBERHART, *A modified particle swarm optimizer*, ...The 1998 IEEE International Conference on, (1998).

[104] H. SPATH, *Cluster analysis algorithms for data reduction and classification of objects*, (1980).

[105] M. STEINBACH, G. KARYPIS, V. KUMAR, AND OTHERS, *A comparison of document clustering techniques*, in KDD workshop on text mining, vol. 400, Boston, 2000, pp. 525–526.

[106] R. STORN AND K. PRICE, *Differential Evolution - A Simple and Efficient Heuristic for global Optimization over Continuous Spaces*, Journal of Global Optimization, 11 (1997), pp. 341–359.

[107] W. STREET, *Oblique multicategory decision trees using nonlinear programming*, INFORMS Journal on Computing, 17 (2005), pp. 25–31.

[108] V. ČERNÝ, *Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm*, Journal of optimization theory and applications, 45 (1985), pp. 41–51.

[109] H. D. VINOD, *Integer Programming and the Theory of Grouping*, Journal of the American Statistical Association, 64 (1969), pp. 506–519.

[110] M. VLACHOS, J. LIN, E. KEOGH, AND D. GUNOPULOS, *A wavelet-based anytime algorithm for k-means clustering of time series*, In Proc. Workshop on Clustering . . . , (2003).

[111] K. WAGSTAFF, C. CARDIE, S. ROGERS, S. SCHRÖDL, AND S. SCHROEDL, *Constrained K-means Clustering with Background Knowledge*, ICML, (2001), pp. 577–584.

[112] S. WANG AND D. ZHU, *Research on selecting initial points for k-means clustering*, Machine Learning and Cybernetics, 5 (2008), pp. 2673–2677.

[113] W. WANG, *Hierarchical clustering*, 2011.

[114] J. WILPON AND L. RABINER, *A modified K-means clustering algorithm for use in isolated work recognition*, Acoustics, Speech and Signal . . . , (1985).

[115] E. XING AND M. JORDAN, *Distance metric learning with application to clustering with side-information*, Advances in neural . . . , (2002), pp. 505–512.

[116] W. XU, X. LIU, AND Y. GONG, *Document clustering based on non-negative matrix factorization*, in Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, ACM, 2003, pp. 267–273.

[117] T. ZHANG, R. RAMAKRISHNAN, AND M. LIVNY, *BIRCH: an efficient data clustering method for very large databases*, ACM SIGMOD Record, (1996).

[118] J. ZHOU, L. JIA, G. HU, AND M. MENENTI, *Evaluation of Harmonic Analysis of Time Series (HANTS): impact of gaps on time series reconstruction*, in 2012 Second International Workshop on Earth Observation and Remote Sensing Applications, IEEE, June 2012, pp. 31–35.

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.