UNIVERSIDADE ESTADUAL PAULISTA
"Júlio de Mesquita Filho"
Pós Graduação em Ciência da Computação

Rodrigo Yuji Mizobe Nakamura

Explorando Abordagens de Aprendizado Sequencial para Floresta de Caminhos Ótimos

UNESP

2014

Rodrigo Yuji Mizobe Nakamura[1]

Explorando Abordagens de Aprendizado Sequencial para Floresta de Caminhos Ótimos

Orientador: Prof. Dr. João Paulo Papa

Dissertação de Mestrado elaborada junto ao Programa de Pós-Graduação em Ciência da Computação - Área de Concentração em Computação Aplicada, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

UNESP

2014

[1]O projeto recebeu apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) através do processo n^o 2011/14058-5. As opiniões, hipóteses e conclusões ou recomendações expressas neste material são de responsabilidade do(s) autor(es) e não necessariamente refletem a visão da FAPESP.

Nakamura, Rodrigo Yuji Mizobe.
Explorando abordagens de aprendizado sequencial para floresta de caminhos ótimos / Rodrigo Yuji Mizobe Nakamura. -- São José do Rio Preto, 2014
55 f. : il., gráfs., tabs.

Orientador: João Paulo Papa
Dissertação (mestrado) – Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências, Letras e Ciências Exatas

1. Computação. 2. Processamento de imagens - Técnicas digitais. 3. Floresta de caminhos ótimos. 4. Markov, Campos aleatórios de. 5. Árvores (Teoria dos grafos) I. Papa, João Paulo. II. Universidade Estadual Paulista "Júlio de Mesquita Filho". Instituto de Biociências, Letras e Ciências Exatas. III. Título.

CDU – 518.72:76

Ficha catalográfica elaborada pela Biblioteca do IBILCE
UNESP - Câmpus de São José do Rio Preto

Agradecimentos

Eu gostaria de agradecer a minha mãe, pelo amor, apoio, incentivo e educação dados em todos esses anos de vida.

Ao meu orientador João Paulo Papa pela confiança e amizade, por todos os ensinamentos e companhia ao longo desses anos de aprendizado.

À Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pelo auxílio financeiro e à UNESP por oferecer a infra-estrutura necessária para o desenvolvimento deste projeto.

Ao Paulo e aos meus irmãos Yuzo e Jéssika pelo companherismo e pensamento positivo, mesmo estando longe.

Ao pessoal do laboratório LCAD pelos trabalhos realizados em conjunto.

Ao pessoal de minha república: Renato, Anderson, Bruno, Thiago e Rodrigo por terem sido grandes amigos durante todos esses anos.

Resumo

A modelagem do problema de classificação como um problema de busca em um grafo fornece uma estrutura elegante, rica em algoritmos eficientes e comprovadamente corretos. A abordagem Floresta de Caminhos Ótimos reduz o problema de classificação para o cálculo de uma floresta de caminhos ótimos relativa a uma função de conectividade, a qual atribui um valor a qualquer caminho no grafo. Considerando o valor máximo entre todos os caminhos possíveis com término em cada vértice, o caminho ideal é trivial para alguns vértices, chamados raízes, e para os vértices restantes, a minimização da função de conectividade atribui a cada vértice um caminho de custo mínimo a partir de sua raiz mais fortemente conectada. Não obstante, para a classificação de novos conjuntos de dados, assume-se que cada amostra compõe um vértice pertecente ao grafo e calcula-se a afinidade deste vértice às árvores geradoras mínimas respectivas a cada classe. Este procedimento não utiliza a estrutura inerente da aplicação que pode ser fundamental para uma melhor precisão dos resultados. Dentro desse contexto, este trabalho avalia a contribuição de técnicas de modelagem contextual como os campos aleatórios Markovianos e as abordagens de empilhamento de classificadores. A modelagem do campo aleatório sumariza o comportamento global do sistema através de suas interações locais. Os métodos baseados em empilhamento de classificadores interpretam as interações entre as amostras como uma análise no espaço escala, capturando as interações de longa distância de forma eficiente através da definição das regiões de vizinhança em múltiplas escalas. Resultados obtidos para a classificação de estruturas anatômicas do cérebro em imagens de ressonância magnética e de coberturas do solo em imagens multi-espectrais de sensoriamento remoto monstram que a inclusão da informação contextual é de fato capaz de melhorar significativamente o desempenho do classificador.

Palavras-chave: Floresta de Caminhos Ótimos, Campos Aleatórios Markovianos, Aprendizado Sequencial Empilhado, Modelagem Contextual

Abstract

The interpretation of classification problem as a graph search provides a rich framework with correct and efficient algorithms. The Optimum-Path Forest classifier can reduce classification to the the computation of an optimum-path forest according to a connectivity function, which assigns a value to any path in the graph. Considering the maximum value among all possible paths with terminus at each node, the optimum path is trivial for some nodes, called roots, and the remaining nodes will have an optimum path coming from their most strongly connected root, partitioning the graph into an optimum-path forest (disjoint sets of optimum-path trees). Notwithstanding, to classify out-of-sample, we assume that each sample in the new dataset composes one node in the graph and we compute their most strongly connected root within all spanning trees. As one can see, this procedure do not take advantage of the problem structure information, which can be fundamental for a better precision of the results. In this context, the purpose of this work is to evaluate the contribution of contextual modelling techniques, such as Markov random fields and stacked classifiers. The first approach, called Markov random fields, sumarizes the system overall behavior through its local interactions. The second approach, based on combination of classifiers, model the interaction between samples in the space scale, which provides efficient implementations of long interaction by defining neighborly relations in multiple scales. The results for brain tissue segmentation of magnetic resonance images and land-cover classification of multi-spectral satellite images show that the contextual information can improve the effectiveness of the Optimum-Path Forest classifier.

Keywords: Optimum-Path Forest, Markov Random Fields, Stacked Sequential Learning, Contextual Modelling

Lista de Figuras

1 Cadeia de Markov de primeira ordem relativa às observações $\{x_n\}$, na qual a distribuição $p(x_n|x_{n-1})$ de uma observação em particular x_n é condicionada ao valor da observação anterior x_{n-1}. 7
2 Modelo de cadeia Markoviana de segunda ordem. Neste caso, a distribuição de uma observação em particular x_n é condicionada ao valor das duas últimas observações (x_{n-1} e x_{n-2}). 8
3 Exemplo de modelo de grafo não direcionado que não pode ser modelado perfeitamente por um grafo direcionado. 9
4 Um simples modelo de grafo direcionado. 14
5 Grafo completo. 18
6 Árvore geradora mínima e os protótipos escontrados. 20
7 Floresta de Caminhos Ótimos após execução do algoritmo de treinamento. 21
8 Propagação dos rótulos baseado no processamento local da floresta de caminhos ótimos. 22
9 Atribuição do rótulo respectivo a raiz mais fortemente conectada. . 23
10 Imagem de ressonância magnética de um cérebro humano ponderada em T1. 30
11 Identificação dos tecidos relativos a substância branca, substância cinzenta (em cinza claro) e líquido cérebro-espinhal (em cinza escuro). 31
12 Resultado da etapa inicial para a identificação dos tecidos utilizando o algoritmo Floresta de Caminhos Ótimos. 32
13 Resultado final após a execução do método proposto ($\beta = 0.54$). . . 32
14 Resultado do processo de identificação dos tecidos após as etapas de pré-processamento utilizando o classificador Floresta de Caminhos Ótimos. 33
15 Resultado do processo de identificação dos tecidos após as etapas de pré-processamento utilizando o método proposto. 34
16 Resultados em termos de κ obtidos aplicando o sistema de classificação proposto para os pacientes 01 e 02. 35
17 Resultados em termos de κ obtidos aplicando o sistema de classificação proposto para os pacientes 03 e 04. 35
18 Resultados em termos de κ obtidos aplicando o sistema de classificação proposto para os pacientes 05 e 06. 36
19 Resultados em termos de κ obtidos aplicando o sistema de classificação proposto para os pacientes 07 e 08. 36
20 Resultados em termos de κ obtidos aplicando o sistema de classificação proposto para os pacientes 09 e 10. 37

21 Resultados em termos de κ obtidos aplicando o sistema de classificação proposto para os pacientes 11 e 12. 37
22 Resultados em termos de κ obtidos aplicando o sistema de classificação proposto para os pacientes 13 e 14. 38
23 Resultados em termos de κ obtidos aplicando o sistema de classificação proposto para os pacientes 15 e 16. 38
24 Resultados em termos de κ obtidos aplicando o sistema de classificação proposto para os pacientes 17 e 18. 39
25 Imagem composta colorida obtida pelo sensor CBERS-2B CCD (20m) (R2G3B4) sobre a área de Itatinga, SP - Brasil (731 x 683 pixels). 44
26 Figura obtida de forma manual por um especialista identificando as regiões de pastagens (verde claro), reflorestamento (verde escuro), culturas (salmão), estradas (cinza), barragens (vermelho) e arbustos (verde). 44
27 Assinaturas espectrais da soja de plantio direto e alface romana, presentes nas base de dados Indian Pines e Salinas, respectivamente [35]. 45
28 Resultado do processo de classificação utilizando o algoritmo Floresta de Caminhos Ótimos sob a pressuposição de independência dos *pixels*. 46
29 Fronteiras entre as regiões indentificadas utilizando o algoritmo Floresta de Caminhos Ótimos. 46
30 Resultado utilizando o método baseado em Aprendizado Sequencial Empilhado Multi-Escala com decomposição multi-resolução. 47
31 Fronteiras das regiões identificadas utilizando o método baseado em Aprendizado Sequencial Empilhado Multi-Escala com decomposição multi-resolução. 48
32 Resultado utilizando o método baseado em Aprendizado Sequencial Empilhado Multi-Escala com decomposição piramidal. 48
33 Fronteiras das regiões identificadas utilizando o método baseado em Aprendizado Sequencial Empilhado Multi-Escala com decomposição piramidal. 49

Lista de Abreviações

CBERS *China-Brazil Earth-Resources Satellite*

CCD Câmera Imageadora de Alta Resolução

ICM *Iterated Conditional Modes*

LBP *Loopy Belief Propagation*

MRF Campos Aleatórios de Markov (*Markov Random Fields*)

MST Árvore Geradora Mínima (*Minimum spanning tree*)

OPF Floresta de Caminhos Ótimos (*Optimum-Path Forest*)

Pixel Elemento de imagem (*Picture Element*)

SA *Simulated Annealing*

SVM Máquinas de Vetores de Suporte (*Support Vector Machines*)

Voxel *Volume Element*

Sumário

1 Introdução **1**

2 Referencial Teórico **6**
2.1 Modelos de Grafos Direcionados 13
2.2 Campos Aleatórios Markovianos 14
2.2.1 Modelo de Potts 16
2.3 Floresta de Caminhos Ótimos 17
2.3.1 Treinamento 18
2.3.2 Classificação 22
2.3.3 Classificação Nebulosa 23

3 Classificação Contextual combinando Floresta de Caminhos Ótimos e Campos Aleatórios de Markov **25**
3.1 Trabalhos relacionados 25
3.2 Abordagem proposta 27
3.3 Aplicação em classificação de tecidos cerebrais 29

4 Aprendizado Sequencial Empilhado **40**
4.1 Decomposição Multi-resolução 42
4.2 Decomposição Piramidal 43
4.3 Análise de imagens de sensoriamento remoto 43

5 Conclusão **50**

1 Introdução

Técnicas de aprendizado de máquina estão transformando o modo com que as áreas de conhecimento organizam e analisam os dados, desde a engenharia e ciência até a medicina e economia. Novos algoritmos desenvolvidos na literatura aperfeiçoaram significativamente os sistemas de reconhecimento de fala, tradução automática, navegação robótica e demais aplicações oriundas das mais variadas áreas das ciências experimentais. Em geral, o objetivo é desenvolver algoritmos capazes de inferir conjecturas gerais e conhecimentos a partir de dados e experiências específicas, com base em princípios estatísticos e computacionais. O procedimento de inferência corresponde em selecionar uma classe de hipóteses para a qual uma função estimada $\hat{f}$ relacione as características observadas e o resultado esperado, $\hat{y} = \hat{f}(x)$. No entanto, sempre existe alguma incerteza remanescente sobre qual o modelo mais correto dentre todas as hipóteses possíveis.

Do ponto de vista matemático, uma comum suposição, devido à simplificação subjacente a muitos modelos estatísticos, para a escolha das hipóteses é assumir que as observações ocorrem independentemente e são identicamente distribuídas, ou seja, são obtidas independentemente de uma mesma distribuição. Note que, quanto menor o espaço amostral definido pela hipótese, mais preciso o método se torna ao generalizar as observações, desde que a verdadeira estrutura seja capaz de ser modelada neste espaço, mais precisamente, que a variabilidade das amostras do problema possa ser representada neste espaço. A habilidade do modelo em realizar inferências para novas amostras, as quais não pertencem ao conjunto utilizado para o aprendizado, caracteriza esta capacidade de generalização.

Muito embora as técnicas tradicionais de reconhecimento de padrões tenham sido empregadas com êxito em diversas áreas de pesquisa [18], existem muitas outras aplicações, por exemplo, séries temporais de problemas da área de finanças e observações metereorológicas, as quais não se enquadram perfeitamente nesses modelos [43]. Em alguns casos, a estrutura origina-se naturalmente de uma ordenação temporal - por exemplo, o processamento de áudio e fala e a série temporal de vendas de uma livraria. Em outros casos, a ordem pode ser apenas tangencialmente relacionada com o tempo, como as tarefas de processamento de uma sequência de caracteres em uma determinada língua (processamento de linguagem natural), ou mesmo completamente alheios a noção temporal, como a análise de sequências biológicas. Todas as tarefas de previsão nestas áreas requerem mais do que apenas respostas verdadeira-falsa ou de múltiplas escolhas, mas possuem um número exponencial de respostas para se considerar. Independentemente do fator de origem, a estrutura inerente ao problema pode ser explorada para a criação de modelos mais precisos e métodos computacionais mais eficientes [24].

Sabe-se que diversos problemas nas áreas de processamento de imagens e visão computacional são mal condicionados por natureza simplesmente porque o espaço

de soluções é extremamente vasto [51]. Mais formalmente, o mal condicionamento surge em decorrência de problemas que não apresentam uma das seguintes condições: existência e unicidade da solução, continuidade e estabilidade [4, 3].

Uma das maneiras de se reduzir o espaço de soluções e, como consequência, atenuar o efeito do mal condicionamento, é a incorporação de conhecimento a *priori* sob a forma de restrições aplicadas à solução desejada. Na classificação contextual de imagens, uma maneira de introduzir conhecimento a *priori* na formulação do problema é através da utilização de restrições de suavidade considerando o contexto espacial dos dados [31, 22, 29]. Ao observar uma imagem ou um vídeo, nós podemos observar que os *pixels* (*picture elements*) variam suavemente com exceção das regiões de borda, também conhecidas como regiões de alta frequência e de descontinuidades.

Dentro desse contexto, o aprendizado sequencial descarta a condição de independência dos dados e, portanto, modela o conjunto de interações entre os elementos, representados por pares $(\vec{x}, y)$, onde $\vec{x}$ representa a amostra como um ponto no espaço $\mathbb{R}^n$ e y a variável dependente relacionada a amostra. Neste caso, a idéia consiste em modelar cada amostra não apenas como um único ponto no espaço de características, mas explorar a interdependência entre as variáveis de entrada $\mathcal{X}$ e as de saída †. Tais dependências podem refletir estruturas sequênciais, espaciais, recursivas ou combinatórias, dependendo do domínio do problema.

Na literatura, o aprendizado sequencial têm sido endereçado através de diferentes perspectivas: do ponto de vista de abordagens de meta-aprendizado através de técnicas de janelas deslizantes, janelas deslizantes recorrentes [17] e aprendizado sequencial empilhado [50, 12, 20]. Do ponto de vista de modelos baseados em grafos, cabe ressaltar a importância dos modelos ocultos de Markov [1], modelos de Markov parcialmente ordenados [16], campos aleatórios de Markov [8] e campos aleatórios condicionais [28].

Métodos de janelas deslizantes oferecem a habilidade de converter um problema de aprendizado supervisionado sequencial em um formato de reconhecimento de padrões clássico. Assume-se que, dada uma sequência de L observações $\langle x_1, x_2, x_3, \ldots, x_L \rangle$ e uma sequência correspondente às variavéis dependentes $\langle y_1, y_2, y_3, \ldots, y_L \rangle$, podemos definir uma janela de comprimento fixo d, representando o contexto local, que desliza ao longo da sequência de observações, capturando algumas informações contextuais ao transformar cada amostra em um sequência de características observadas. Por exemplo, dada uma janela de tamanho igual a três, a janela deslizante irá produzir as seguintes amostras $(\langle \mathcal{B}, x_1, x_2 \rangle, y_1)$, $(\langle x_1, x_2, \ x_3 \rangle, y_2)$,$(\langle x_2, x_3, x_4 \rangle, y_3)$,. . .,$(\langle x_{w-1}, x_w, \mathcal{B} \rangle, y_w)$, onde $\mathcal{B}$ representa o início/fim de uma sequência.

Nos modelos probabilísticos baseados em grafos, cada vértice representa uma variável aleatória (ou um grupo de variáveis aleatórias) e as arestas expressam as

relações de probabilidade entre essas variáveis. O grafo, portanto, captura o modo o qual a distribuição conjunta sobre todas as variáveis aleatórias pode ser decomposta no produto de fatores, cada um dependendo em apenas um subconjunto de variáveis.

Uma das grandes vantagens dos modelos probabilísticos baseados em grafos é que um grafo específico é capaz de representar as relações de probabilidade para uma ampla classe de distribuições. Baseados na propriedade dos processos Markovianos, o qual especificam que, dado uma relação vizinhança, as amostras não adjacentes determinam uma relação estatística de independência condicional, os métodos de minimização da função de energia, a qual representa a função objetivo para inferência destes modelos, podem ser processados localmente de forma eficiente, uma vez que o comportamento global do sistema consiste de suas interações locais, determinadas pelas suposições de independência condicional. Não obstante, os métodos tradicionais de otimização utilizados para a minimização da função de energia, como o método Iterado da Moda Condicional (*Iterated Conditional Modes* - ICM) [6] e a técnica de Recozimento Termicamente Simulado (*Simulated Annealing* - SA), possuem um custo computacional exponencialmente proporcional ao tamanho da clique máxima no grafo, restringindo a utilização prática a grafos com cliques de cardinalidade máxima restrita (por exemplo, um reticulado regular bidimensional, cuja vizinhança representa os elementos ao norte, sul, leste e oeste).

Ao longo dos últimos anos, algoritmos mais eficientes de otimização da função de energia para modelos baseados em grafos foram propostos na literatura, dentre os quais podem-se destacar os baseados em cortes em grafos [7, 26], propagação de crenças (*Loopy Belief Propagation* - LBP) [53] e os métodos poliédricos e combinatórios que resolvem um relaxamento de programação linear do problema de minimização de energia discreto [25, 27, 45].

A técnica de propagação de crenças consiste de um algoritmo exato para a inferência em grafos direcionados sem ciclos, e é equivalente a um caso especial do algoritmo soma-produto (*sum-product*). O algoritmo soma-produto permite, a partir a fatorização da distribuição conjunta em um grafo, calcular eficientemente as probabilidades marginais respectivas a cada variável. Para identificar o estado do grafo com maior probabilidade, o problema é endereçado utilizando um algoritmo similar, conhecido como máximo-produto (*max-product*), o qual pode ser interpretado como a aplicação do algoritmo de programação dinâmica em modelos de grafos. Ao contrário da estratégia de encontrar o estado mais provável através do algoritmo soma-produto, o qual retorna o conjunto de valores mais provável individualmente, o algoritmo máximo-produto é capaz de identificar o conjunto de variáveis que possuem maior probabilidade conjuntamente. A abordagem *Loop Belief Propagation* consiste em aplicar o algoritmo soma-produto múltiplas vezes até que um critério de convergência seja atendido, apesar de não existir garantias

de que o método produzirá bons resultados. Para a aplicação desta abordagem, é necessária a definição de uma sequência de envio de mensagens. Vamos assumir que cada mensagem é enviada em um determinado tempo a partir de um determinado vértice e com uma certa direção. Cada mensagem enviada a um vértice substitui qualquer mensagem recebida previamente para uma mesma direção através da mesma aresta e será unicamente uma função das mensagens recebidas por este vértices nos passos anteriores do algoritmo. Toda mensagem só pode ser enviada através da aresta de um vértice quando todas as mensagens tenham sido recebidas pelo vértice através das suas demais arestas. Devido a possibilidade do grafo conter ciclos, a informação, nesses casos, percorrerá o grafo múltiplas vezes, e desta forma, o algoritmo pode não convergir. Em algumas aplicações, este método pode ser altamente efetivo, além disso, algoritmos do estado da arte para decodificação de certos tipos de códigos de correção a erros (*error-correcting codes*) são equivalentes ao algoritmo *Loop Belief Propagation*.

Szeliski et al. [47] apresentaram um estudo comparativo dessas técnicas para diversas aplicações da área de visão computacional. Porém, esse estudo restringiu os modelos a grafos em um reticulado regular com quatro vértices conexos com apenas potenciais unários e pareados. Consequentemente, a capacidade para modelar interações de longo alcance é limitada, geralmente, resultando na suavização excessiva dos limites do objeto no contexto de processamento de imagens. Recentemente, Kappes et al. [23] apresentaram uma nova comparação dos modelos de inferência na qual os parâmetros dos modelos de grafos são estimados, permitindo modelos de ordem superior e com estrutura de conectividade espacial não uniforme.

Nesta dissertação, o objetivo não foi desenvolver métodos de otimização para o processo de inferência dos parâmetros dos grafos de fatores[2], por outro lado, este projeto demonstrou a contribuição de duas técnicas de modelagem contextual para os métodos baseados em caminhos ótimos. A primeira abordagem avalia a combinação do classificador Floresta de Caminhos Ótimos (*Optimum-Path Forest* - OPF) e os Campos Aleatórios Markovianos (*Markov Random Fields* - MRF). A segunda abordagem consiste na técnica de combinação de classificadores conhecida como aprendizado sequencial empilhado (*Stacked Sequential Learning*). Os métodos baseados em combinação de classificadores, como a técnica de empilhamento, são projetados para melhorar as previsões através da combinação de múltiplos modelos de aprendizado de máquina. Alguns trabalhos recentes monstraram que o uso de meta-características, isto é variáveis adicionais descrevendo cada amostra do conjunto de dados, pode auxiliar o desempenho dos métodos de aprendizado

[2]Grafo de fatores corresponde a uma das nomenclaturas dos potenciais de interação entre os vértices do grafo. Dentro desse contexto, os fatores são funções de probabilidade (impróprias), cujo produto é a distribuição posterior (não normalizada).

de máquina e reconhecimento de padrões, possibilitando inclusive a especificação de relações não lineares através de um ajuste fino dos modelos. Neste trabalho, a estratégia foi incluir meta-atributos descrevendo a estrutura espacial da aplicação com o objetivo de melhorar a identificação de coberturas do solo em imagens de sensoriamento remoto. A compreensão do alcance e limitações destes algoritmos é investigada em situações reais, fornecendo subsídios sólidos para o seu correto entendimento.

Além desta seção introdutória, contendo a motivação e os objetivos do projeto em questão, essa dissertação apresenta outras cinco seções, descritas a seguir: a Seção 2 revisa o referencial teórico das modelagens contextuais baseadas em grafos, destacando a modelagem baseada em campos aleatórios de Markov e o classificador Floresta de Caminhos Ótimos. A Seção 3 apresenta a combinação do modelo baseado em caminhos ótimos com as ferramentas matemáticas disponíveis para os campos aleatórios através da definição de uma função de energia composta por um potencial unário definido pela afinidade em relação à árvore geradora mínima (*Minimum spanning tree* - MST) [13] respectiva a cada classe, e um potencial de clique fornecido pelo modelo de Potts, o qual reflete o comportamento global do sistema através das iterações locais. Neste estudo, o modelo de Potts sumariza as interações entre os vértices definidos em uma clique do campo aleatório representando o contexto local do sistema. Experimentos utilizando imagens de ressonância magnética com o intuito de classificar tecidos cerebrais foram realizados comparando a abordagem utilizando o classificador baseado em caminhos ótimos sob a pressuposição de independência dos dados e o método proposto, cujo objetivo é capturar a relação estrutural das imagens. Na Seção 4, as abordagens de aprendizado empilhado baseadas na interpretação das interações entre as amostras como uma análise no espaço-escala, capturando as interações de longa distância de forma eficiente através da definição das regiões de vizinhança em múltiplas escalas, são apresentadas no contexto da análise de sensoriamento remoto de coberturas do solo. Finalmente, a Seção 5 apresenta as conclusões e considerações finais, bem como as perpectivas futuras para a continuidade das pesquisas relacionadas ao projeto.

2 Referencial Teórico

Aprendizado de máquina é a ciência que estuda e desenvolve técnicas capazes de realizar inferências a partir de dados incompletos e ruidosos. Com o intuito de abordar problemas complexos, podem-se empregar dois princípios para a simplificação do problema: o primeiro é o princípio da modularidade, cujo objetivo é dividir o problema em subproblemas os quais são solucionados independentemente por um módulo específico, e o segundo é o princípio da abstração, o qual consiste em desvincular as questões subjacentes ao problema que não afetam o cumprimento do conjunto de restrições. A teoria das probabilidades oferece esses princípios através da fatorização da probabilidade conjunta e da sumarização dos dados através das estatísticas suficientes.

O conjunto de ferramentas fundamentado pela teoria das probabilidades possibilita, consistentemente, a quantificação da incerteza relacionada aos ruídos das observações e da limitação de aprendizado imposta pelo conjunto finito de dados.

Uma comum suposição proveniente da teoria das probabilidades para representar grandes conjuntos de dados é assumir algumas relações de indepêndencia condicional. As propriedades de independência condicional são cruciais em modelos probabilísticos para reconhecimento de padrões por simplificar a estrutura do modelo, e também por reduzir o poder computacional necessário para a inferência e o aprendizado do modelo. Em um conjunto de variáveis $\mathbf{V}$, duas variáveis aleatórias X e Y são condicionalmente independentes dado $\mathbf{Z} \subseteq \mathbf{V} \backslash X, Y$ (denotado $X \perp\!\!\!\perp Y|\mathbf{Z}$), se:

$$\forall x, y, z : P(X = x|Y = y, \mathbf{Z} = \mathbf{z}) = P(X = x|\mathbf{Z} = \mathbf{z}), \tag{1}$$

dado que $\forall \mathbf{z} : P(\mathbf{Z} = \mathbf{z}) > 0$.

A independência condicional é uma generalização da tradicional noção de independência estatística. Se duas variáveis X e Y são independentes, então a distribuição conjunta é determinada pelo produto das probabilidades marginais: $P(X = x, Y = y) = P(X = x)P(Y = y)$.

O conceito de independência condicional é um elemento fundamental nas modelagens baseadas em grafos, pois permite a fatorização da distribuição conjunta, sendo representado pela inexistência de um caminho entre os elementos independentes.

Ao lidar com conjuntos complexos de variáveis dependentes, assume-se que a sequência das variáveis aleatórias constituem um processo Markoviano, o qual especifica que $x_{t+1} \perp\!\!\!\perp x_{1:t-1}|x_t$, ou seja, "o futuro é independente do passado dado o presente". A distribuição conjunta é então definida como o produto dos fatores e o modelo conhecido como cadeias de Markov, as quais são caracterizadas por uma distribuição inicial sobre os estados e uma matriz de transição de estados. As

Figuras 1 e 2 ilustram dois exemplos de cadeias Markovianas, sendo de primeira e segunda ordem, respectivamente.

Em algumas aplicações, a distribuição condicionada $p(x_n|x_{n-1})$ que define o modelo é forçada a ser igual, correspondendo à pressuposição que a série temporal é estacionária[3]. Este modelo é conhecido como cadeia de Markov homogênea. Note que ao estender a ordem do modelo de Markov, torna-se possível capturar informações de tendências e, além disso, se considerarmos modelos de ordem superior, o modelo poderia identificar informações sobre a sazonalidade de uma série temporal. Entretanto, a flexibilidade do modelo induz uma maior complexidade, sendo necessária a estimação de um número maior de parâmetros.

Figura 1: Cadeia de Markov de primeira ordem relativa às observações $\{x_n\}$, na qual a distribuição $p(x_n|x_{n-1})$ de uma observação em particular x_n é condicionada ao valor da observação anterior x_{n-1}.

Pode-se generalizar o pressuposto de Markov de primeira ordem para dimensões arbitrárias utilizando um modelo de representação baseado em grafos, no qual cada vértice representa uma variável aleatória e a ausência de uma aresta, o pressuposto de independência condicional. As arestas definidas no grafo podem ser direcionadas ou não direcionadas: no primeiro caso, os modelos são conhecidos como redes Bayesianas e a aresta define uma relação probabilística não simétrica entre as variáveis aleatórias; em contrapartida, os campos aleatórios de Markov não especificam a orientação das arestas e, portanto, são frequentemente utilizados como restrições suaves em aplicações como análise de imagens e dados relacionais.

Para formalizar a diferença entre os modelos direcionados e não direcionados, observe que um grafo $\mathcal{G}$ é dito um **mapa de independência (I-map)** de uma distribuição $\mathbb{P}$ se I-map($\mathcal{G}$) $\subseteq$ I-map($\mathbb{P}$). Uma equivalência perfeita ocorre apenas quando I-map($\mathcal{G}$) = I-map($\mathbb{P}$), ou em outras palavras, o grafo pode representar

[3]Em distribuições sequênciais estacionárias, os dados apresentam variações com o tempo, entretanto a distribuição, a qual os dados são gerados permanece a mesma $P(x_t|x_{t-1}) = P(x_2|x_1); \{t \in 2 \dots T\}$.

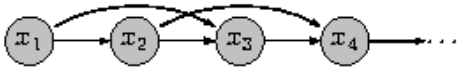

Figura 2: Modelo de cadeia Markoviana de segunda ordem. Neste caso, a distribuição de uma observação em particular x_n é condicionada ao valor das duas últimas observações (x_{n-1} e x_{n-2}).

todas (e somente) as propriedades de independência condicional da distribuição $\mathbb{P}$. Os modelos de grafos com arestas direcionadas e não direcionadas equivalem perfeitamente a diferentes conjuntos de distribuições. Por exemplo, um grafo direcionado consegue modelar perfeitamente uma estrutura V, $A \rightarrow C \leftarrow B$, na qual pode-se afirmar as seguintes relações entre as variáveis aleatórias: $A \perp\!\!\!\perp B$ e $A \not\perp\!\!\!\perp B|C$. Ao descartar as orientações das arestas, obtém-se o grafo representado como $A - C - B$, o qual afirma incorretamente que $A \perp\!\!\!\perp B|C$ e $A \not\perp\!\!\!\perp B$. Por outro lado, um modelo de arestas não direcionadas pode modelar perfeitamente um ciclo com quatro vértices (Figura 3), enquanto os modelos direcionados não podem representar precisamente **todas e apenas** as independências condicionais da distribuição representada por este grafo não direcionado. As distribuições que podem ser modeladas como grafos direcionados e não direcionadas são representadas por grafos denominados **cordais**; para tais grafos, ao agrupar todos os vértices por clique máxima, obtém-se uma representação em forma de árvore para a distribuição, o que permite a utilização de algoritmos eficientes baseados em programação dinâmica. É importante salientar que toda árvore (incluindo as cadeias como casos especiais) representa um grafo cordal.

O conjunto mínimo de vértices que contém informações sobre X, as quais não podem ser obtidas por nenhuma outra variável aleatória é conhecido como *Markov blanket* ($\mathbf{Mb}(X)$, Definição 1). Este conjunto $\mathbf{Mb}(X)$ define as relações de fatorização do grafo e, por conseguinte, as suposições de independência condicional. As Seções 2.1 e 2.2 apresentam detalhadamente tais relações para os grafos com arestas direcionadas e não direcionadas, respectivamente.

Definição 1 ***(Markov blanket).*** *Um Markov blanket $\mathcal{M}\lfloor$ de uma variável $T \in \mathcal{V}$ na distribuição conjunta de probabilidade $\mathbb{P}$ sobre variáveis $\mathcal{V}$ é o conjunto de*

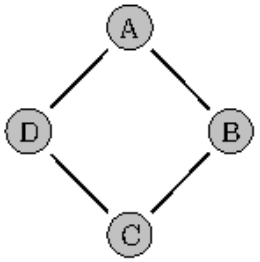

Figura 3: Exemplo de modelo de grafo não direcionado que não pode ser modelado perfeitamente por um grafo direcionado.

variáveis que ao condicionar X as estas observações, todas as demais variáveis são independentes de X, ou seja, para cada $T \in (\mathcal{V} \backslash \mathcal{M}\lfloor \backslash \{X\}), X \perp\!\!\!\perp T|\mathcal{M}\lfloor$.

Definição 2 ***(Estatística suficiente)**. $S(X)$ é uma estatística suficiente para o parâmetro Θ se $p(x|t, \Theta) = p(x|t)$. Assim, dado S, X não traz nenhuma informação adcional sobre o parâmetro Θ.*

Os modelos estatísticos para análise de dados baseiam-se em resumos dos dados observados através das estatísticas descritivas. Por exemplo, utilizam-se estimativas da média e variância (μ, σ) para descrever a distribuição Gaussiana univariada, representando as estatísticas suficientes desta distribuição para descrever o processo generativo dos dados observados. Para os casos multivariados da distribuicção Gaussiana, os parâmetros de interesse correspondem a média e a matriz de covariância, $\mathcal{N}(\mu, \sum)$.

Note que uma estatística sumariza uma medida de probabilidade se a probabilidade de uma amostra é uma função apenas do valor desta estatística. Uma estatística a qual sumariza toda uma distribuição de uma família é uma estatística suficiente para aquela família pelo critério de fatorização de Neyman (Teorema 1).

Teorema 1 ***(Critério de fatoração de Neyman)**. Uma função $T(X)$ é suficiente para Θ se, e somente se,:*

$$p(x|\Theta) = f(t, \Theta)g(x), \tag{2}$$

com f e g sendo funções não negativas.

A partir do modelo dos dados observados, podem-se realizar previsões utilizando o teorema Bayesiano, no qual assume-se que as características são condicionalmente independentes dada a observação da classe:

$$p(x|y=c,\Theta) = \prod_{j=1}^{\mathcal{D}} p(x_j|y=c,\Theta_{jc}), \tag{3}$$

$$p(y,x) = p(y)p(x|y=c,\Theta), \tag{4}$$

onde $\mathcal{D}$ representa o número de características e Θ, as estatísticas descritivas (parâmetros do modelo). Note que as inferências são realizadas através da especificação de funções de densidade e probabilidade que sumarizam cada classe como uma distribuição parametrizada. O elemento principal para a precisão das inferências é especificar adequadamente uma função densidade específica a cada classse, a qual define o perfil dos dados que espera-se pertencer a esta classe, sendo os parâmetros estimados utilizando um critério baseado em máxima verossimilhança. Otimizar uma função de verossimilhança $p(\mathcal{D}|\Theta)$ corresponde em selecionar os valores dos parâmetros Θ para os quais a probabilidade dos dados observadas seja maximizada. Entretanto, apesar da função de verossimilhança mensurar a probabilidade dos dados observados para diferentes valores de parâmetros dos modelos, a verossimilhança não representa uma distribuição de probabilidade sobre os parâmetros, e sua integral em relação aos parâmetros não necessariamente precisa ser igual a um. O denominador do teorema de Bayes pode ser visto como uma constante de normalização a qual garante que a probabilidade condicional definida no lado esquerdo da equação possua soma igual a um.

Portanto, a predição de uma amostra pertencer a uma determinada classe é obtida computando a probabilidade *a posteriori* baseada no produto da probabilidade de ocorrência de um elemento pertencer esta classe e a verossimilhança desta amostra em relação ao modelo parametrizado. Os classificadores baseados no aprendizado de uma probabilidade conjunta $p(x,y)$ das observações e variáveis de resposta são denominados generativos, pois para estes métodos é possível criar dados sintéticos amostrando o modelo apreendido. Tais técnicas realizam suas previsões para novas observações utilizando a probabilidade condicional $p(y|x)$ obtida pela regra de Bayes.

Para clarificar o conceito de verossimilhança, considere o jogo simples de verificar se um número pertence a uma sequência lógica de um conjunto de números observados, proposto por Josh Tenenbaum [49]. Para simplificar o exemplo, assume-se que todos os números são inteiros positivos no intervalo $[1,100]$. Observa-se a sequência $\mathcal{D} = \{16\}$; deste modo, existem várias regras consistentes para definir as hipóteses, dentre elas, a hipótese $h_{dois} \triangleq$ "potências de dois" e a hipótese $h_{par} \triangleq$ "números pares". Para a primeira hipótese h_{dois}, o espaço de ocorrências

possíveis é definido pelo conjunto $\{2, 4, 8, 16, 32, 64\}$, enquanto para a segunda hipótese, h_{par}, $\{2, 4, 6, \ldots, 98, 100\}$. O **raciocínio de Occam** define que deve-se escolher a hipótese mais simples (a qual representa o menor espaço de ocorrências) que seja consistente com os dados observados. Note que $p(\mathcal{D}|h_{dois}) = \frac{1}{6}$, pois existe apenas 6 números que são potência de dois e menores que 100, por outro lado, $p(\mathcal{D}|h_{par}) = \frac{1}{50}$, a Equação 5 define a verossimilhança em função do espaço definido pela hipótese, e o número de observações N:

$$p(\mathcal{D}|h) = \left[\frac{1}{\text{size(h)}}\right]^N = \left[\frac{1}{|h|}\right]^N . \tag{5}$$

Note que a medida que os dados são observados em um espaço discreto, qualquer hipótese inconsistente com as novas informações (observações) deverá ter sua probabilidade *a posteriori* atribuída a zero, e portanto a hipótese é descartada. É importante lembrar sobre o termo de conhecimento *a priori* da regra de Bayes, o qual permite evitar a definição de uma hipótese pouco provável como a "potências de dois com excessão do número 32" para as seguintes observações $\mathcal{D} = \{2, 4, 8, 16, 64\}$. Além disso, ao maximizar a função de verossimilhança, torna-se comum a estimativa imprecisa da variância da distribuição. Este é um fenômeno conhecido como viés e está relacionado ao problema de super ajuste no contexto de interpolação por polinômios.

Por outro lado, os classificadores denominados discriminativos tem por objetivo otimizar diretamente uma função de erro $\Delta : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ que, em princípio, determina o custo de se estimar $\hat{y}$ quando o verdadeiro valor é y, sendo estritamente positiva. Algumas das funções Δ frequentemente utilizadas são: a função de custo quadrática, que penaliza severamente valores extremos de erro; a função de custo de valor absoluto, que penaliza de modo linearmente crescente os erros; a função que apenas classifica como erro ou acerto e a função *hinge loss*, critério de maximização da margem ao hiperplano de decisão. Com base em pressupostos estatísticos típicos, a média minimiza a função de custo quadrática, enquanto a mediana minimiza a função de custo de valor absoluto.

Dentre os principais classificadores, destacam-se as Máquinas de Vetores de Suporte [14], k-vizinhos mais próximos [18], Floresta de Caminhos Ótimos [41, 40], Regressor Logístico, Redes Neurais Artificiais e Árvores de Decisão.

O método popularmente conhecido, Máquinas de Vetores de Suporte, propõe solucionar o problema de classificação de padrões assumindo ser possível definir uma superfície de decisão entre as classes. Para os casos onde não é possível definir a separabilidade linear, a idéia consiste em utilizar uma função para a transformação do espaço de tal forma que o problema de encontrar uma função discriminante não linear seja transformado em um problema de encontrar um hiperplano no novo espaço definido pela transformação. Uma propriedade importante desta técnica é

que a determinação dos parâmetros do hiperplano corresponde a um problema de otimização convexa, e portanto, qualquer solução local representa um ótimo global.

Métodos baseados em vizinhos mais próximos utilizam as observações do conjunto de treinamento para identificar as amostras mais próximas no espaço métrico definido pelo conjunto de variáveis de entrada e a função de distância. Tais métodos envolvem a escolha do parâmetro de vizinhança k, cujo objetivo é atenuar a característica de alta variância do método.

A técnica Regressor Logístico assume uma forma paramétrica da distribuição a posteriori $P(y|x)$, utilizando a função logística sigmóide e estimando diretamente seus parâmetros utilizando o conjunto de treinamento. Em síntese, o objetivo é otimizar diretamente a probabilidade condicional, de tal forma, que as classes possuam um melhor critério de separabilidade em função das variáveis de entrada.

As redes neurais artificiais representam uma classe de técnicas computacionais cujo objetivo é o de modelar matematicamente a estrutura neural de organismos inteligentes, adquirindo conhecimento através da experiência pelo algoritmo de retopropagação com o objetivo de minimizar o erro esperado entre as variáveis de respostas esperadas e as previsões do modelo.

O método não paramétrico de aprendizado supervisionado, conhecido como árvores de decisão, constrói uma base de regras de decisão como modelo, e a previsão para cada amostra pode ser interpretada como uma expressão lógica. Apesar do problema de aprendizado de uma árvore de decisão ótima não possuir um algoritmo polinomial como solução, na prática, utilizam-se algumas heurísticas para associar a cada vértice de decisão o atributo mais informativo, utilizando, por exemplo, o critério de impureza de Gini ou a entropia como uma medida quantitativa do ganho de informação, dentre aquelas variáveis de entrada não utilizadas no caminho desde a raiz até a árvore.

Na literatura de aprendizado de máquina, o negativo do logaritmo da função de verossimilhança é conhecido como a função de erro a ser minimizada. Como o negativo do logaritmo é uma função monotonamente decrescente, maximizar a verossimilhança equivale a minimizar a função de erro. Além de simplicar a análise matemática, aplicar a função logaritmo sobre a verossimilhança contruibui para a estabilidade numérica das operações, uma vez que o produto de um grande número de probabilidades baixas podem facilmente causar probemas de precisão numérica para os computadores.

Note que ambos os métodos de aprendizado, generativo e discriminativo, possuem seus pontos positivos e negativos. A abordagem generativa por envolver a modelagem da probabilidade conjunta sobre as variáveis de entrada e de resposta requer um grande conjunto de dados de treinamento para determinar as densidades condicionadas a variável de resposta, especialmente em aplicações com um

alto número de variáveis de entrada. A vantagem reside na capacidade de detectar amostras cuja probabilidade seja pequena segundo o modelo e, portanto, para as quais as previsões tenham uma baixa precisão, sendo esta habilidade conhecida como detecção de anomalias ou de novos conceitos. A abordagem discriminativa possui como ponto positivo otimizar diretamente a função objetivo, entretanto o ponto negativo é a perda da habilidade de estimar a probabilidade *a posteriori* da amostra. Na literatura existem algumas técnicas para estimar uma medida de probabilidade para os classificadores discriminativos, porém as probabilidades obtidas desta forma, geralmente, não são bem calibradas.

2.1 Modelos de Grafos Direcionados

Um modelo de grafo direcionado em um conjunto de variáveis aleatórias consiste em um grafo orientado acíclico e uma função de probabilidade condicional para cada variável aleatória condicionada à observação dos vértices pais. Dado a existência de uma aresta com origem em x_i até x_j, x_i é então denominado pai de x_j, definindo uma relação anti-simétrica entre os vértices. A partir da ordenação topológica desta representação, definem-se as relações de independência condicional entre as variáveis aleatórias através da afirmação de que um vértice é condicionalmente independente de seus demais predecessores dado as observações de seus vértices pais (Equação 6). Se um dado vértice x_i não possui nenhum pai, pois não possui nenhuma aresta incidente, então a distribuição condicional será apenas a distribuição marginal $p(x_i)$:

$$x_t \perp\!\!\!\perp x_{pred(t)\setminus pais(t)} | x_{pais(t)}. \tag{6}$$

Neste caso, o *Markov blanket* do vértice t é definido como os vértices pais, $pais(t)$; os vértices filhos, $filhos(t)$ e os vértices que compartilham a paternidade de seus filhos, $co-pais(t)$ (por exemplo, os vértices 2 e 3 da Figura 4). Para observar a razão dos vértices $co-pais(t)$ pertencerem ao *Markov blanket*, veja na Equação 8 a derivação de $p(x_t|x_{t-1})$. Os modelos de grafos direcionados desempenham um papel central em ambientes de modelagem com muitas variáveis, especialmente para as redes de crenças, nas quais as arestas direcionadas em um grafo podem ser interpretadas como dependências diretas entre variáveis pais e filhas. Note que uma condição a qual estabeleça que o grafo seja acíclico previne o raciocínio circular. A Figura 4 ilustra a seguinte distribuição como um grafo:

$$\begin{aligned} p(x_{1:5}) &= p(x_1)p(x_2|x_1)p(x_3|x_1,\cancel{x_2})p(x_4|\cancel{x_1},x_2,x_3)p(x_5|\cancel{x_1},\cancel{x_2},x_3,\cancel{x_4}), & (7)\\ p(x_{1:5}) &= p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2,x_3)p(x_5|x_3), & (8) \end{aligned}$$

onde cada termo $p(x_t|x_{pai(t)})$ é uma distribuição de probabilidade condicional.

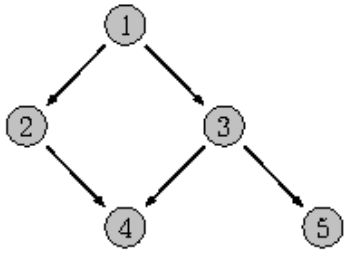

Figura 4: Um simples modelo de grafo direcionado.

2.2 Campos Aleatórios Markovianos

A teoria dos Campos Aleatórios Markovianos baseia-se na relação local entre os vértices $v \in \mathcal{V}$ de um grafo $\mathcal{G}$. Uma vez que não existe uma ordenação topológica associada aos grafos não direcionados, a regra da cadeia não pode ser utilizada para representar a distribuição $p(x)$. Portanto, ao invés de associar uma distribuição de probabilidade condicional a cada vértice, definem-se funções potenciais para cada clique máxima no grafo $\mathcal{G}$, sem perda de generalidade, uma vez que as demais cliques são subconjuntos da clique máxima. Veja a definição de clique a seguir (Definição 3).

Definição 3 *(**Clique**). Uma clique ou conjunto completo num grafo é qualquer conjunto de vértices dois a dois adjacentes. Em outras palavras, $\mathcal{X}$ é uma clique se o grafo induzido $\mathcal{G}[\mathcal{X}]$ é completo.*

A função pontencial por clique c será denotada por $\psi_c(x_c|\Theta_c)$, onde x_c representa as variáveis aleatórias respectivas a concretização particular do campo aleatório para a clique c e Θ_c os parâmetros da função potencial, e pode representar qualquer função não negativa apropriada para modelar o comportamento global do sistema. O Teorema 2 demonstra que para qualquer distribuição, para as quais suas propriedades de independência condicional possam ser representadas por um grafo não direcionado, pode-se então calcular a probabilidade conjunta $p(x)$ através do produto das funções potenciais por clique. Veja que o cálculo da constante de normalização Z para encontrar a probabilidade de uma configuração de x é determinada localmente, e portanto computacionalmente realizavél.

Teorema 2 ***(Hammersley-Clifford)****. Uma distribuição positiva $p(x) > 0$ satisfaz as propriedades de independência condicional de um grafo não direcionado $\mathcal{G}$, se e somente se, p puder ser representado como um produto das funções potenciais, uma por clique máxima:*

$$p(x|\Theta) = \frac{1}{Z(\Theta)} \prod_{c \in \mathcal{C}} \psi_c(x_c|\Theta_c), \tag{9}$$

onde $\mathcal{C}$ é o conjunto de cliques máximas de $\mathcal{G}$ e $Z(\Theta)$, a função de partição, a qual garante que a distribuição $p(x)$ da equação acima seja corretamente normalizada, sendo definida por

$$Z(\Theta) \triangleq \sum_{v} \prod_{c \in \mathcal{C}} \psi_c(x_c|\Theta_c). \tag{10}$$

Ao considerar apenas funções potenciais as quais satisfazem a restrição $\psi_c(x_c) \geq 0$, garante-se que $p(x) \geq 0$. Um outro ponto importante para se observar é que não existe a restrição das funções potenciais à classe de funções as quais existem uma interpretação probabilística como distribuições marginais ou conjuntas. Este é um ponto divergente em relação aos modelos baseados em grafos direcionados, nos quais cada fator representa uma distribuição condicional relativa a variável correspondente, condicionada ao estado de seus vértices pais.

Entretanto, uma das consequências devido a liberdade na escolha das funções potenciais $\psi_c(x_c)$ é que o produto entre as funções não é geralmente corretamente normalizado, e portanto a constante de normalização deve ser explicitamente definida, como vimos acima na definição de $Z(\Theta)$. Relembre que para os modelos baseados em grafos direcionados, a distribuição conjunta é automaticamente normalizada como consequência da normalização de cada distribuição condicional na fatorização do grafo.

A presença do denominador de normalização representa a grande limitação dos modelos de grafos não direcionados, pois para um grafo com M vértices cada um com K estados possíveis, o cálculo do fator de normalização envolve a soma dos K^M estados e portanto, no pior cenário, sua complexidade é exponencial em relação ao tamanho do modelo. Entretanto, para a avaliação da distribuição condicional local, a função de partição não é necessária, pois a condicional é representada como a razão entre as probabilidades marginais, e deste modo, a função de partição se cancela entre o numerador e o denominador para o cálculo da razão.

Analogamente, para o cálculo das probabilidades marginais locais, pode-se utilizar a probabilidade conjunta não normalizada e normalizar as probabilidades marginais explicitamente no final, como as probabilidades marginais envolvem um menor número de variáveis, o cálculo do fator de normalização é computacionalmente realizável. Note que ao definir a função potencial como uma função $\psi_c(x_c|\theta)$

(Equação 11), nós podemos observar uma conexão entre os modelos de grafos não direcionados e a distribuição de Gibbs, a qual pode ser escrita segundo a Equação 12. É importante enfatizar que quanto menor for a energia de uma configuração da distribuição de Gibbs, maior será a sua probabilidade de ocorrência.

A escolha das funções potenciais deve ser analisada como uma função a qual retorna a preferência por um estado em relação aos demais. Assim, as configurações globais do modelo com alta probabilidade são aquelas que encontram um bom balanceamento em satisfazer as (possivelmente conflituosas) influências dos potencias de clique.

$$\psi_c(x_c|\theta) = \exp(-E(x_c|\theta_c)). \tag{11}$$

$$p(x|\theta) = \frac{1}{Z(\theta)} \exp(-\sum_c E(x_c|\theta_c)). \tag{12}$$

2.2.1 Modelo de Potts

O modelo antiferromagnético de Potts é um modelo de campo aleatório Markoviano, o qual emergiu a partir da física estatística com o intuito de generalizar o modelo de Ising para múltiplos estados discretos pertencentes a um intervalo finito [42, 52], sendo um dos modelos mais utilizados para diversos domínios de problemas. Os estados são muitas vezes consideradas como cores, de modo que o estado fundamental consiste em uma configuração em que não há dois vizinhos com a mesma cor [33]. No contexto de processamento de imagens e reconhecimento de padrões, o modelo de Potts é um dos modelos mais utilizados para refletir um conhecimento *a priori* na forma de restrições de suavidade nas imagens, pois os seus *pixels* possuem uma alta correlação espacial devido a presença de regiões homogêneas.

De acordo com o Teorema 2 (**Hammersley-Clifford**), o modelo de Potts pode ser equivalentemente definido de duas formas: por uma distribuição conjunta (Gibbs) ou pelas funções de densidade condicionais locais. Note que a representação por modelos locais é mais adequada no contexto de análise de imagens, pois permite o cálculo direto das probabilidades para cada elemento do campo de maneira individual e não para toda uma ocorrência de um campo aleatório. As probabilidades condicionais locais definidas através dos potenciais de clique permitem medir a ligação estatística entre um vértice (interpretado como uma variável aleatória) e os valores observados para os demais vértices, que pertencem a clique dos quais dependem a distribuição condicional local.

Considerando um sistema de vizinhança $\mathcal{N}$, podemos definir a função densidade

condicional local de um modelo de Potts como:

$$p(x_{ij} = m | x_{\mathcal{N}_{ij}}, \beta) = \frac{\exp\{\beta \mathcal{H}_{ij}(m)\}}{\sum_{l=1}^{|\mathcal{L}|} \exp\{\beta \mathcal{H}_{ij}(l)\}}, \tag{13}$$

onde $\mathcal{H}_{ij}$ representa o número de amostras em $\mathcal{N}ij$ que possuem rótulo igual a classe representada pelo rótulo l, β é o parâmetro representando a dependência espacial entre os vizinhos (também conhecido como o inverso da temperatura, $\beta = \frac{1}{T}$), e m, o valor observado para a amostra central x_{ij} do sistema de vizinhança $\mathcal{N}$. Observa-se que β é um compromisso entre o conhecimento *a priori* e a informação observada e, portanto, quanto maior o valor de β, maior a dependência espacial entre as amostras. No caso de $\beta = 0$, assume-se que as amostras são independentes e, portanto, não existe dependência entre as variáveis.

2.3 Floresta de Caminhos Ótimos

Um conjunto de dados pode ser visto como uma matriz $\mathcal{X} \in \mathbb{R}^{m \times n}$ na qual cada linha identifica uma amostra do conjunto e cada coluna, uma variável aleatória correspondente a uma característica associada à amostra. Técnicas de aprendizado de máquina e reconhecimento de padrões assumem uma classe de hipóteses com o intuito de realizar previsões precisas, $h : \mathcal{X} \to \mathcal{Y}$, associando cada observação a um resultado estruturado (um valor real ou um valor inteiro identificando uma classe). Uma amostra do conjunto de dados representa um ponto no espaço de características $\mathbb{R}^n$ e uma função de distância entre dois pontos, $\Phi(x, y)$, define o espaço métrico ou o espaço de probabilidades no qual o algoritmo ajusta seu modelo aos dados observados. Muitos dos problemas de classificação estudados na área de aprendizado de máquina podem ser interpretados como uma superfície de resposta colorida definida no espaço vetorial $\mathbb{R}^{|\mathcal{Y}|}$. As amostras do conjunto de treinamento consistem de exemplos coloridos no mapa e o objetivo é colorir os demais pontos do mapa. Note que, a partir de amostras diferentes do conjunto de treinamento, obtém-se como resultado superfícies diferentes.

O classificador Floresta de Caminhos Ótimos [41, 40] é uma técnica baseada em grafos a qual modela os problemas de classificação como uma busca por caminhos ótimos em um grafo derivado a partir de uma relação de adjacência entre amostras em um determinado espaço de características. Os vértices são representados por vetores de características e as arestas conectam pares de amostras. Vértices representativos (protótipos) são escolhidos entre as amostras de treinamento para todas as classes e utilizados para classificação dos vértices remanescentes através de um processo de competição entre os protótipos com base nos custos de caminhos no grafo. Portanto, para construir um classificador baseado em Floresta de Caminhos

Ótimos, deve-se estabelecer a relação de adjacência, a função de custo de caminho e a metodologia para estimar os protótipos. Este método têm como vantagem não assumir que: as amostras formam nuvens compactas no espaço métrico; os grupos de amostras não se sobrepoẽm; cada grupo de pontos corresponde a uma classe; e que a função de probabilidade densidade das classes apresentam formas apropriadas para a modelagem paramétrica. Além das propriedades interessantes, este classificador é matematicamente correto e de fácil implementação. A próxima seção descreve o classificador em mais detalhes.

2.3.1 Treinamento

Suponha que exista um conjunto de dados rotulado $\mathcal{Z}(\mathcal{X}, \mathcal{Y}) = \mathcal{Z}_1 \cup \mathcal{Z}_2$, onde $\mathcal{Z}_1$ e $\mathcal{Z}_2$ representam o conjunto de treinamento e validação, respectivamente. Como veremos a seguir, o conjunto de treinamento é utilizado para inferir o modelo de classificação e o conjunto de validação, para estimar a precisão do modelo para previsões de novos dados diferentes do conjunto utilizado para aprender o modelo. Seja $(\mathcal{Z}_1, \mathcal{A})$ um grafo completo onde os vértices representam as amostras em $\mathcal{Z}_1$ e a relação de adjacência $\mathcal{A}$ consiste em todos os pares de vértices no produto cartesiano $\mathcal{A} = \mathcal{Z}_1 \times \mathcal{Z}_1$ (Figura 5).

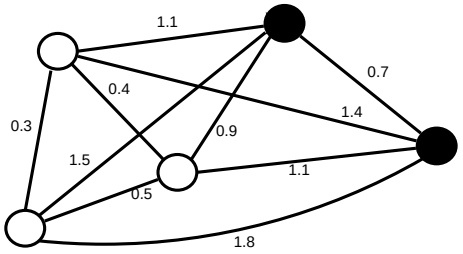

Figura 5: Grafo completo.

Para um grafo $(\mathcal{Z}_1, \mathcal{A})$, um caminho $\pi_s = \langle s_i, s_2, \ldots, s \rangle$ representa uma sequência de vértices adjacentes na qual s representa o estado final. Um caminho é dito trivial quando $\pi_s = \langle s \rangle$. Um caminho $\pi_s = \pi_t \cdot \langle t, s \rangle$ representa uma extensão do caminho π_t por uma aresta (t, s). Note que todos os caminhos considerados neste trabalho são caminhos simples, isto é, caminhos sem vértices repetidos.

Um mapa de predecessores é uma função $\mathcal{P}$ a qual atribui a cada vértice $s \in \mathcal{G}$, um vértice adjacente em $\mathcal{G}$ ou um marcador distintivo $nil \notin \mathcal{G}$. Neste caso, onde não é atribuído um vértice predescessor, s representa uma semente utilizada para a construção de uma árvore relativa a classe representada pela amostra s. Uma floresta geradora consiste de um mapa de predecessores que não contém ciclos, ou seja, para cada vértice $t \in \mathcal{G}$, uma floresta geradora $\mathcal{P}$ define um caminho π_s recursivamente como $\langle s \rangle$ se $P(s) = nil$ ou $\pi_t \cdot \langle t, s \rangle$ se $P(s) \neq nil$.

Uma função de conectividade atribui um custo $\Psi(\pi_s)$ para qualquer caminho π_s. Neste trabalho nós definimos Ψ (Equação 14) como uma função a qual retorna a aresta com maior custo ao longo do caminho com o intuito de evitar as restrições e para melhor representar a conectividade entre as amostras.

$$\begin{aligned} \Psi(\langle s \rangle) &= \begin{cases} 0 & \text{se } P(s) = nil \\ +\infty & \text{em outro caso,} \end{cases} \\ \Psi(\pi_s \cdot \langle s, t \rangle) &= \max\{\Psi(\pi_s), d(s,t)\}. \end{aligned} \tag{14}$$

Seja $\Pi(\mathcal{Z}_1, \mathcal{A}, t)$ o conjunto de todos os caminhos no grafo $\mathcal{G}$ com estado final em s. Um caminho é dito ótimo de acordo com a seguinte definição.

Definição 4 *(**Caminho Ótimo**). Um caminho π_s é ótimo se $\Psi(\pi_s) \leq \tau_s$ para qualquer caminho $\tau_s \in \Pi(\mathcal{Z}_1, \mathcal{A}, t)$.*

O algoritmo Floresta de Caminhos Ótimos resolve o problema de otimização, minimize $\Psi(\pi_s)$, $\forall s \in \mathcal{Z}_1$, através do algoritmo modificado de caminho mínimo de Dijkstra para múltiplas fontes. Dado o grafo $\mathcal{G}$, a função de conectividade Ψ e a relação de adjacência $\mathcal{A}$, o método atribui um caminho ótimo π_s para todo vértice $s \in \mathcal{G}$, obtendo uma floresta geradora $\mathcal{P}$ onde todos os caminhos são ótimos. É importante salientar que a prova matemática de corretude do algoritmo restringe Ψ às funções suaves, as quais satisfazem a seguinte Definição 5.

Definição 5 *(**Função de conectividade suave**). Uma função de conectividade Ψ é suave se para todo vértice $s \in \mathcal{G}$, existe um caminho ótimo π_s o qual é trivial ou possui a forma $\tau_t \cdot \langle t, s \rangle$ onde*

- $\Psi(\tau_t) \geq \Psi(\pi_s)$.
- *τ_t é ótimo.*
- *Para qualquer caminho ótimo τ'_t, $\Psi(\tau_t \cdot \langle t, s \rangle = \Psi(\pi_s)$.*

Note que em uma floresta de caminhos ótimos, qualquer caminho com estado inicial em uma raiz é um caminho ótimo completo, se a função de conectividade for suave, segundo a Definição 6.

Definição 6 *(**Caminho ótimo completo**). Um caminho $\pi_{s_n} = \langle t_1, t_2, \ldots, t_n \rangle$ é ótimo completo se todos os caminhos $\pi_{t_i} = \langle t_1, t_2, \ldots, t_i \rangle$, $i \in [1, n]$ forem caminhos ótimos.*

Um conjunto ótimo de raízes S^* pode ser encontrado explorando a relação teórica entre árvores geradoras mínimas e árvores de caminhos ótimos para um grafo acíclico onde os vértices representam todas as amostras de $\mathcal{Z}_1$ e as arestas são não direcionadas e ponderadas pela função de distância entre as amostras. A árvore geradora mínima é ótima no sentido que a soma dos pesos de suas arestas é mínima se comparada a qualquer outra árvore geradora no grafo completo. Na árvore geradora mínima, cada par de amostras é conectada por um único caminho, que é ótimo segundo Ψ. Consequentemente, cada árvore geradora mínima contém uma árvore de caminhos ótimos para cada raiz. Os protótipos ótimos são definidos como as amostras de diferentes classes mais próximas em uma árvore geradora mínima (i.e, elementos que estão na fronteira das classes, como podemos observar na Figura 6).

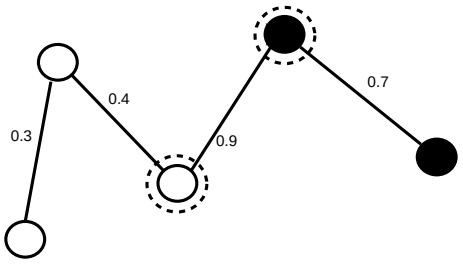

Figura 6: Árvore geradora mínima e os protótipos escontrados.

Ao selecionar raízes em torno das regiões de fronteira, o algoritmo conquista primeiro as regiões de fronteira do espaço métrico definido, bloqueando as passagens para os caminhos com origem a partir de outras raízes. Portanto, os possíveis caminhos de transição entre as fronteiras recebem custos mais elevados do que os caminhos de mesma região em relação a cada uma das raízes. Cada raiz $s \in \mathcal{S}$ define uma zona de influência (árvore de caminho ótimo enraizada em s) composta das amostras que são mais fortemente conectadas à s em relação a qualquer outra raiz. A propagação dos custos pelo caminho a partir das raízes descreve uma região ordenada crescente, semelhante a inundação de um relevo. Ao final deste procedimento, nós obtemos uma floresta de caminhos ótimos, que representa uma coleção de árvores de caminhos ótimos enraizadas em cada protótipo (Figura 7).

O Algoritmo 1 abaixo implementa a fase de treinamento.

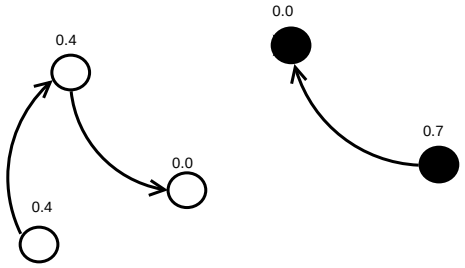

Figura 7: Floresta de Caminhos Ótimos após execução do algoritmo de treinamento.

Algoritmo 1 – Treinamento por Floresta de Caminhos Ótimos

Entrada:	Um conjunto de treinamento $\mathcal{Z}_1$ λ-rotulado, conjunto de protótipos $\mathcal{S}$ e o espaço métrico $\phi(x)$ composto pelo vetor de características e a função de distância entre dois pontos.
Saída:	Floresta de Caminhos Ótimos $\mathcal{P}_1$, mapa de custos de caminho $\mathcal{C}_1$, mapa de rótulos $\mathcal{L}_1$, e o conjunto ordenado $\mathcal{Z}_1'$.
Auxiliares:	Fila de prioridade $\mathcal{Q}$ e variável de custo *cst*.

1. $\mathcal{Z}_1' \leftarrow \emptyset$
2. *Para cada* $s \in \mathcal{Z}_1 \backslash \mathcal{S}$, *faça*
3. $\quad \mathcal{C}_1(s) \leftarrow +\infty$.
4. *Para cada* $s \in \mathcal{S}$, *faça*
5. $\quad \mathcal{C}_1(s) \leftarrow 0$
6. $\quad \mathcal{P}_1(s) \leftarrow nil$
7. $\quad \mathcal{L}_1(s) \leftarrow \lambda(s)$
8. $\quad$ *Insira* s *em* $\mathcal{Q}$.
9. *Enquanto* $\mathcal{Q} \neq \emptyset$, *faça*
10. $\quad$ *Remova de* $\mathcal{Q}$ *uma amostra* s *tal que* $\mathcal{C}_1(s)$ *e minimo.*
11. $\quad$ *Insira* s *em* $\mathcal{Z}_1'$.
12. $\quad$ *Para cada* $t \in \mathcal{Z}_1$ *tal que* $\mathcal{C}_1(t) > \mathcal{C}_1(s)$, *faça*
13. $\quad\quad$ *Compute* $cst \leftarrow \max\{\mathcal{C}_1(s), d(s,t)\}$.
14. $\quad\quad$ *Se* $cst < \mathcal{C}_1(t)$, *então*
15. $\quad\quad\quad$ *Se* $\mathcal{C}_1(t) \neq +\infty$, *então remova* t *de* $\mathcal{Q}$.
16. $\quad\quad\quad$ $\mathcal{P}_1(t) \leftarrow s$, $\mathcal{L}_1(t) \leftarrow \mathcal{L}_1(s)$, $\mathcal{C}_1(t) \leftarrow cst$.
17. $\quad\quad\quad$ *Insira* t *em* $\mathcal{Q}$.
18. *Retorne o classificador* $[\mathcal{P}_1, \mathcal{C}_1, \mathcal{L}_1, \mathcal{Z}_1']$.

Em síntese, as amostras sementes, também conhecidas como protótipos, são inicializadas com um custo zero e inseridas em uma fila de prioridades $\mathcal{Q}$, enquanto que as demais amostras recebem um valor de custo máximo. Enquanto a fila de prioridades não estiver vazia, o vértice s removido é sempre o vértice de custo

acumulado mínimo em $\mathcal{Q}$, pois a fila é mantida ordenada na ordem crescente de custo acumulado. Esta característica faz com que o processo de classificação termine em um número menor de iterações. Para todos os elementos t do conjunto de dados, verifica-se se s oferece um caminho a t com um custo menor do que o custo atual de t. Caso a amostra t receba um novo custo ótimo, a amostra é então reinserida na fila de prioridades com o custo atualizado e o vértice s se torna seu novo predecessor. O tempo de processamento do algoritmo depende principalmente da estrutura de dados utilizada para a fila $\mathcal{Q}$ e do mecanismo de ordenação dos vértices de $\mathcal{Q}$. Ao final deste procedimento, uma floresta de caminhos ótimos $\mathcal{P}_1$ é encontrada relativa a função de conectividade.

2.3.2 Classificação

A propagação dos rótulos para novas amostras $t \in \mathcal{Z} \backslash \mathcal{T}$ é eficientemente executada baseado no processamento local da floresta de atributos e as distâncias entre os nós $s \in \mathcal{T}$ e t. Para todas as amostras $t \in \mathcal{Z}_2$, considera-se todas as arestas conectando o vértice t a todas as amostras $s \in \mathcal{Z}_1$, como se amostra t pertencesse ao grafo de treinamento (Figura 8). Considerando todos os caminhos possíveis a partir de $\mathcal{S}^*$ até o vértice t, encontra-se o caminho ótimo $\mathcal{P}^*(t)$ a partir de $\mathcal{S}^*$ e a amostra t é classificada com a classe $\lambda(\mathcal{R}(t))$ relativa ao protótipo mais fortemente conexo, $\mathcal{R}(t) \in \mathcal{S}^*$ (Figura 9). Esse caminho pode ser identificado avaliando o custo ótimo $\mathcal{C}(t)$:

$$\mathcal{C}(t) = \min\{\max\{\mathcal{C}(s), d(s,t)\}\}, \ \forall s \in \mathcal{Z}_1. \tag{15}$$

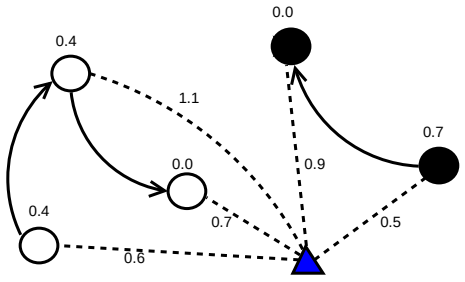

Figura 8: Propagação dos rótulos baseado no processamento local da floresta de caminhos ótimos.

Seja o vértice $s^* \in \mathcal{Z}_1$ aquele que satisfaça a Equação 15 (i.e, o vértice predecessor $\mathcal{P}(t)$ no caminho ótimo $\mathcal{P}^*(t)$). Dado que $\mathcal{L}(s^*) = \lambda(\mathcal{R}(t))$, a classificação simplesmente atribui $\mathcal{L}(s^*)$ como a classe do vértice t. Um erro ocorre quando $\mathcal{L}(s^*) \neq \lambda(t)$.

O Algoritmo 2 apresenta o processo de classificação.

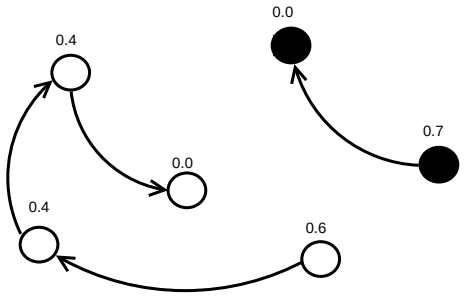

Figura 9: Atribuição do rótulo respectivo a raiz mais fortemente conectada.

Algoritmo 2 – Classificação por Floresta de Caminhos Ótimos

Entrada:	Classificador $[\mathcal{P}_1, \mathcal{C}_1, \mathcal{L}_1, \mathcal{Z}_1']$, conjunto de validação $\mathcal{Z}_2$, e o par (v, d) composto pelo vetor de características e as distâncias computadas.
Saída:	Rótulo $\mathcal{L}_2$ e predecessor $\mathcal{P}_2$ mapas definidos por $\mathcal{Z}_2$, e valor de acurácia *Acc*.
Auxiliares:	Variáveis de custo *tmp* e *mincost*.

1. *Para cada* $t \in \mathcal{Z}_2$, *faça*
2. $\quad i \leftarrow 1$, $mincost \leftarrow max\{\mathcal{C}_1(k_i), d(k_i, t)\}$.
3. $\quad L_2(t) \leftarrow \mathcal{L}_1(k_i)$ *e* $\mathcal{P}_2(t) \leftarrow k_i$.
4. $\quad$ *Enquanto* $i < |\mathcal{Z}_1'|$ *e* $mincost > \mathcal{C}_1(k_{i+1})$, *faça*
5. $\quad\quad$ *Compute* $tmp \leftarrow \max\{\mathcal{C}_1(k_{i+1}), d(k_{i+1}, t)\}$.
6. $\quad\quad$ *Se* $tmp < mincost$, *entao*
7. $\quad\quad\quad$ $mincost \leftarrow tmp$.
8. $\quad\quad\quad$ $\mathcal{L}_2(t) \leftarrow \mathcal{L}(k_{i+1})$ *e* $\mathcal{P}_2(t) \leftarrow k_{i+1}$.
9. $\quad\quad$ $i \leftarrow i + 1$.
10. *Calcule a acuracia acc de acordo com Papa* et al. *[41].*
11. *Retorne* $[\mathcal{L}_2, \mathcal{P}_2, acc]$.

2.3.3 Classificação Nebulosa

A classificação através do algoritmo Floresta de Caminhos Ótimos consiste em encontrar os caminhos ótimos progressivamente a partir das raízes, conectando cada amostra $t \in \mathcal{Z}_2$ com todas os vértices $s \in \mathcal{Z}_1$. Assim sendo, nós podemos definir uma função de afinidade (força de conectividade) ($\mathcal{C}_{\mathcal{Z}_2,\mathcal{Y}} : \mathcal{Z}_2 \rightarrow [0, 1]$), a qual define a capacidade de cada árvore geradora mínima reivindicar a adesão daquela amostra a classe a qual representa.

Algoritmo 3 – Classificação Nebulosa por Floresta de Caminhos Ótimos

Entrada: Classificador $[\mathcal{P}_1, \mathcal{C}_1, \mathcal{L}_1, \mathcal{Z}_1']$, conjunto de teste $\mathcal{Z}_2$, e o par (v, d) composto pelo vetor de características e as distâncias computadas.
Saída: Rótulo $\mathcal{L}_2$, campos de rótulos $\mathcal{C}_i 2, i \in \{1, |\mathcal{L}|\}$ e valor de acurácia *Acc*.
Auxiliares: Variáveis de custo *tmp*, *totcost* e *mincost*. Vetor $weight \in \mathbb{R}^{|\mathcal{L}|}$.

1. *Para cada* $k \in |\mathcal{L}|$, *faça*
2. $\quad weight(k) \leftarrow +\infty$
3. *Para cada* $t \in \mathcal{Z}_2$, *faça*
4. $\quad$ *Para cada* $i \in \mathcal{Z}_1'$, *faça*
5. $\quad\quad$ *Compute* $tmp \leftarrow \max\{\mathcal{C}_1(i), d(i,t)\}$.
6. $\quad\quad$ *Se* $tmp < weight(L(i))$, *entao*
7. $\quad\quad\quad weight(L(i)) \leftarrow tmp$.
8. $\quad totcost \leftarrow 0$.
9. $\quad$ *Para cada* $k \in |\mathcal{L}|$, *faça*
10. $\quad\quad totcost \leftarrow totcost + weight(k)$
11. $\quad\quad$ *Se* $mincost < weight(k)$
12. $\quad\quad\quad \mathcal{L}_2(t) \leftarrow k$
13. $\quad\quad\quad mincost \leftarrow weight(k)$.
14. $\quad$ *Para cada* $k \in |\mathcal{L}|$, *faça*
15. $\quad\quad \mathcal{C}_{2,k}(t) \leftarrow \frac{weight(k)}{totcost}$.
16. *Calcule a acuracia acc de acordo com Papa* et al. *[41].*
17. *Retorne* $[\mathcal{L}_2, \mathcal{P}_2, acc]$.

3 Classificação Contextual combinando Floresta de Caminhos Ótimos e Campos Aleatórios de Markov

Muitos métodos de aprendizado de máquina não exploram a estrutura inerente ao problema durante processo de identificação, descrição e reconhecimento de padrões. Entretanto, amostras em uma vizinhança espacial/temporal possuem uma alta probabilidade de exibir comportamentos similares. As seções subsequentes apresentam uma abordagem que unifica o algoritmo de aprendizado supervisionado Floresta de Caminhos Ótimos e um modelo de interdependência dos dados através dos Campos Aleatórios Markovianos, no sentido de modelar o conhecimento *a priori* assegurando a suposição de suavidade espacial/temporal dos dados. A seção 3.3 apresenta os resultados obtidos pelo algoritmo proposto para a classificação de tecidos cerebrais em imagens de ressonância magnética.

3.1 Trabalhos relacionados

Ao longo dos últimos anos, com a proliferação de câmeras digitais e a ampla disponibilidade dos computadores modernos, o campo de análise de imagens tem sido desenvolvido extensivamente, resultando em significantes progressos em segmentação, descrição e recuperação de imagens. Não obstante, cada imagem exibe um diferente grau de iluminação, ruído, perspectiva e escala tornando o processo de identificação dos conteúdos de uma imagem extremamente difícil.

O processo de segmentação de imagens consiste em subdividir o domínio da imagem em regiões com base nos príncipio de similaridade e descontinuidade, ou seja, os pixels pertencentes a uma mesma região compartilham certa semelhança com respeito a alguma propriedade, como cor e textura, e regiões adjacentes são significativamente diferentes com respeito às mesmas características. De um modo geral, nós podemos categorizar as técnicas de segmentação a partir de um ponto de vista teórico em: limiarização [37, 44, 38], detecção de bordas [9, 39, 30] e crescimento de regiões [2, 10, 19, 32].

Técnicas de limiarização podem ser baseadas na definição de um limiar global, onde um único valor é utilizado para particionar todo o domínio da imagem, ou na definição de um limiar local, definindo um limiar diferente para cada subconjunto da imagem com base no histograma local, como, por exemplo o algoritmo de Otsu.

O segundo conjunto de métodos os quais têm por objetivo a deteção de bordas são normalmente baseados em bancos de filtros de orientação e identificam a orientação que corresponde à resposta máxima de um banco de filtros, identificando as regiões de alta frequência, onde existe uma descontinuidade da região.

O terceiro grupo de métodos baseados em crescimento de regiões seleciona sementes iniciais as quais são expandidas segundo um critério de homogeneidade.

A principal vantagem dessas técnicas consiste na preservação da informação contextual durante o processo de crescimento das sementes. Entretanto, a precisão é altamente dependente da escolha apropriada das sementes iniciais e sensível aos ruídos, sombras e iluminação desigual presentes em uma mesma imagem, como, por exemplo, o ruído de baixa frequência presente em imagens de ressonância magnética denominado heterogeneidade do campo magnético e os ruídos de alta frequência decorrentes da movimentação do indíviduo durante o processo de aquisição da imagem.

Uma possível solução consiste na utilização de algoritmos de aprendizado de máquina com o intuito de aprender propriedades globais do conjunto de imagens e, portanto, mais robustos em relação aos problemas descritos. Ademais, tais técnicas permitem a modelagem de um certo conjunto de imagens, abstraindo os conceitos específicos a cada imagem e, portanto, inferindo conjecturas gerais sobre as variáveis extraídas para cada região de interesse. Deste modo, torna-se possível a utilização deste modelo para segmentar um novo conjunto de imagens com características similares.

Entretanto, muitos métodos de aprendizado de máquina pressupõem que as amostras são independentemente distribuídas no espaço das características e, portanto, não exploram as interdependências existentes entre os dados. No entanto, amostras adjacentes possuem uma maior probabilidade de pertencerem a uma mesma região, tornando crucial a exploração da informação contextual para um resultado mais preciso.

Alguns trabalhos recentes propuseram versões contextuais de técnicas de reconhecimento de padrões tradicionais, isto é, variações que consideram a correlação entre amostras adjacentes. Tarabalka et al. [48] propuseram uma abordagem híbrida, conhecida como SVM-MRF, composta por Máquinas de Vetores de Suporte (*Support Vector Machines* - SVM) e Campos Aleatórios de Markov para classificação de imagens de sensoriamento remoto. A idéia consiste em adicionar informações contextuais em um segundo estágio, após uma classificação probabilística pixel a pixel utilizando o SVM sob a pressuposição de independência das amostras. Neste segundo passo, uma regularização baseada em campos aleatórios, o qual adiciona um conjunto de restrições relativas a suavidade espacial, é executada sobre o resultado inicial em conjunto com uma máscara para detecção dos *pixels* de borda computada a partir do filtro de Sobel. Para a estimativa da solução final, os autores propuseram a utilização do algoritmo de Metrópolis, baseado em relaxamento estocástico e recozimento, este método é baseado em uma abordagem Bayesiana e tem por objetivo minimizar a função de energia global através da minimização das energias locais iterativamente.

Moser and Serpico [34] também propuseram uma abordagem SVM-MRF similar para classificação contextual. Entretanto, o trabalho apresentou uma inovação

por utilizar uma única formulação para ambos SVM e MRF, na qual a classificação contextual é executada em um único passo. Os autores provaram que sob um certo conjunto de pressuposições, a energia correspondente a um modelo de campo aleatório sobre o espaço associado a transformação baseada na função kernel do classificador SVM, pode ser expressada como uma expansão da função kernel definindo um kernel Markoviano. Em síntese, a idéia principal destes classificadores contextuais consistem em modelar um conhecimento *a priori* como um campo localmente dependente, a fim de realizar uma suposição de suavidade espacial dos *pixels* [5, 21, 15].

Ambos os trabalhos tem em comum o classificador SVM, sendo o MRF utilizado para a regularização do mapa de rótulos baseado na pressuposição de coerência espacial da imagem. Recentemente, uma técnica de classificação proposta por Papa et al. [41, 40] tem sido empregada com sucesso em diversos domínios, incluindo aplicações de análise de imagens as quais demandam etapas de retreinamento do classificador em tempos iterativos [46]. O classificador Floresta de Caminhos Ótimos aborda o reconhecimento de padrões como um problema de particionamento de um grafo, no qual uma competição entre amostras principais de todas as classes produzem uma coleção de árvores de caminhos ótimos enraizadas nestas amostras principais. Neste trabalho, nós apresentamos uma versão contextual conhecida como OPF-MRF, a qual executa rapidamente várias iterações do OPF seguido por melhorias na descrição dos *pixels* com base nos Campos aleatórios de Markov e informações contextuais a partir de segmentações intermediárias.

3.2 Abordagem proposta

A principal idéia que fundamenta o método proposto é especificar uma função de energia a qual contempla potenciais unários, os quais capturam a evidência local da amostra pertencer a uma certa classe e potenciais de clique, os quais impoẽm um conjunto de restrições em relação ao contexto local das amostras, como por exemplo, a suposição de que amostras vizinhas tendem a pertencer a mesma classe definida pelo modelo de Potts, que penaliza os nós vizinhos que possuem uma classe (cor) diferente. A minimização da energia do sistema pode ser realizada através do algoritmo ICM, o qual utiliza uma estratégia determinística gulosa para encontrar uma solução correspondente a um mínimo local, ou seja, corresponde a simplesmente a aplicação do método gradiente ascendente coordenado. O Algoritmo 4 consiste em, a partir de uma estimativa de segmentação inicial de uma imagem, atribuir a cada amostra o rótulo que minimiza a função densidade condicional local definida pelo modelo de campo aleatório utilizado, neste caso o modelo de Potts. Em síntese, o método define se a amostra, neste caso o *voxel* (*volume element*) da imagem, deve ser rotulado como uma classe diferente do estado atual ou não baseando-se no cálculo da probabilidade condicional local relativa a cada classe.

Na prática, como estamos utilizando o modelo de Potts, o qual penaliza uniformemente a diferença entre os estados, o cálculo da probabilidade condicional consiste apenas no cálculo da razão entre o número de ocorrências relativas a cada classe no determinado clique e o número de vértices da clique. Nós podemos observar que os resultados são extremamente sensíveis à estimativa inicial, especialmente em espaços de alta dimensão e energias não convexas.

Algoritmo 4 – OPF-MRF

Entrada: Um conjunto de treinamento rotulado, $\mathcal{Z}_1$, um conjunto de avaliação, $\mathcal{Z}_2$, e um descritor, (v, d), representando as características e a métrica de distâncias.
Saída: Mapa de rótulos, $\mathcal{L}$.

1. *Definir o modelo MRF, $\mathcal{M}$ (por exemplo, modelo de Potts)*
2. *Treinamento do OPF utilizando o conjunto $\mathcal{Z}_1$.*
3. *Classificação de $\mathcal{Z}_2$ produzindo um mapa inicial, $\mathcal{L}$.*
4. *Para cada iteração i ($i = 1, \ldots, \mathcal{N}_{ite}$), faça*
5. *Calcule a função de energia, $\mathcal{E}$,*
6. *associada ao modelo $\mathcal{M}$ para o mapa $\mathcal{L}$.*
7. *Treinamento de uma nova instância do OPF estendendo*
8. *o conjunto original de caracteristicas com*
9. *($\mathcal{E}_j, j = 1, \ldots, \mathcal{N}_{classes}$).*
10. *Atualize o mapa $\mathcal{L}$ aplicando o novo classificador sobre $\mathcal{Z}_2$.*

O algoritmo OPF-MRF inicia o processo inferindo um modelo de classificação baseado em floresta de caminhos ótimos (não contextual) utilizando $\mathcal{Z}_1$ e, ao classificar o conjunto $\mathcal{Z}_2$, obtém-se como resultado um mapa de segmentação inicial (Linhas $2-3$ do Algoritmo 4). O laço nas linhas linhas $4-10$ aprimora os resultados do mapa de classificação como uma função fornecida pela densidade local do modelo de Potts (Equação 13). O mapa de rótulos é atualizado usando o algoritmo ICM, que visa maximizar a probabilidade local (linhas 5-6). Posteriormente, o conjunto original de características para cada amostra é estendido com um conjunto de meta-características associadas a probabilidade respectivas a cada classe (linhas $7-8$). Em seguida, o modelo de classificação baseado em caminhos ótimos é re-treinado e aplicado a fim de aperfeiçoar o mapa de classificaçao resultante (linhas $9-10$). O laço principal do algoritmo é executado repetitivamente até que um critério de convergência seja satisfeito, por exemplo, o número de trocas de estados realizadas em relação ao tamanho do modelo.

3.3 Aplicação em classificação de tecidos cerebrais

A análise quantitativa das estruturas cerebrais a partir de imagens de ressonância magnética auxilia o processo de diagnóstico e tratamento de doenças relacionadas a atrofia das regiões do hipocampo e do córtex entorrinal tais como Alzheimer, epilepsia e esquisofrenia. A Figura 10 ilustra um dos exames cerebrais fornecidos pelo *Center for Morphometric Analysis - Massachusetts General Hospital*.

Os exames coronais cerebrais foram ponderados com as características micro-estruturais locais da difusão de água do tipo T1 após serem centralizados e normalizados. Os exames são ponderados para ajustar o contraste da imagem de modo a demonstrar estruturas ou patologias anatômicas diferentes. Cada tecido retorna ao seu estado de equilíbrio após a excitação pelos processos independentes de relaxamento de T1 (*spin*-reticulado) e T2 (*spin-spin*).

Para criação de uma imagem ponderada em T1, temos que esperar pelos diferentes valores de magnetização antes de medir o sinal de ressonância magnética, alterando o tempo de repetição. Essa ponderação da imagem é útil para avaliar o córtex cerebral, identificando tecido adiposo, caracterização das lesões hepáticas focais e para geração de imagens pós-contraste. A identificação dos tecidos de substância cinzenta (*Gray Matter* - GM), substância branca (*White Matter* - WM) e líquido cérebro-espinhal (*Cerebrospinal Fluid* - CSF), conforme Figura 11, possibilita um estudo de populações comparando pacientes que apresentam as patologias supracitadas e indivíduos saudáveis.

Para imagens de ressonância magnética do tipo T1 de imagens cerebrais, espera-se que as intensidades dos tecidos de substância cinzenta no cortex sejam maiores que as intensidades do líquido cerebrospinal e menores que dos tecidos de substância branca. Para a extração de características relativas ao conjunto de variáveis de entrada para o aprendizado do modelo de classificação, cada amostra é representada pela intensidade do brilho de seus 26 *voxels* vizinhos. Portanto, cada *voxel* é utilizado para compor o conjunto de dados de sua imagem correspondente, para então ser dividida em conjuntos de treinamento e avaliação.

Para avaliar a precisão do método, nós utilizamos o procedimento de validação cruzada com proporção $0.05 - 99.95\%$ para o conjunto de treinamento e avaliação, respectivamente. A Figura 12 ilustra a etapa inicial de identificação dos tecidos da imagem utilizando um modelo de classificação baseado no algoritmo Floresta de Caminhos Ótimos assumindo que os *voxels* são independemente distribuídos.

A Figura 13 ilustra o resultado obtido pelo método proposto. Como podemos observar, a imposição de restrições acerca da suavidade espacial dos *voxels* resultou em uma solução mais coerente com o resultado esperado (Figura 11).

Note que as técnicas de delineamento do cérebro que visam selecionar apenas as regiões do telencéfalo e cerebelo, removendo os conteúdos intra-craniais e o tronco auxiliam os algoritmos de identificação dos tecidos (Figuras 14 e 15, κ igual a 0.8812

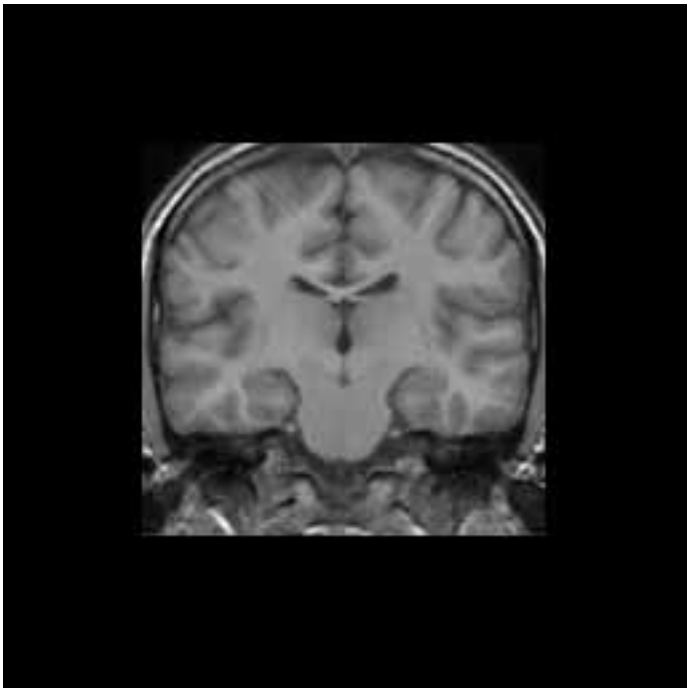

Figura 10: Imagem de ressonância magnética de um cérebro humano ponderada em T1.

e 0.8995). As técnicas de correção de homogeneidade visam atenuar o efeito do ruído de baixa frequência denominado heterogeneidade do campo magnético, o qual varia de acordo com o transdutor de ressonância magnética e com indivíduo sujeito a ressonância. Este efeito determina que *voxels* (*volume elements*) contendo substâncias brancas e cinzentas de diferentes partes do cérebro possam ter brilhos similares, influenciando significativamente na precisão dos métodos de classificação de tecidos. Vale ressaltar que uma das principais dificuldades com técnicas de ressonância magnética é a inconsistência das intensidades dos voxels mesmo se o procedimento utilizar o mesmo protocolo, região de interesse, transdutor e paciente [36]. Acrescenta-se ainda, os ruídos devido a movimentação do paciente durante a aquisição da imagem, a falta de contraste e a variabilidade anatômica.

As Figuras a seguir ilustram os resultados em termos do κ de Cohen [11] para cada imagem utilizando três diferentes valores para a variável de restrição espacial β (Equação 13), a qual controla a quantidade de informação contextual utilizada no processo de aprendizado. O índice de κ representa uma medida de concordância entre o dado de referência e o resultado do classificador, sendo computada pela seguinte fórmula:

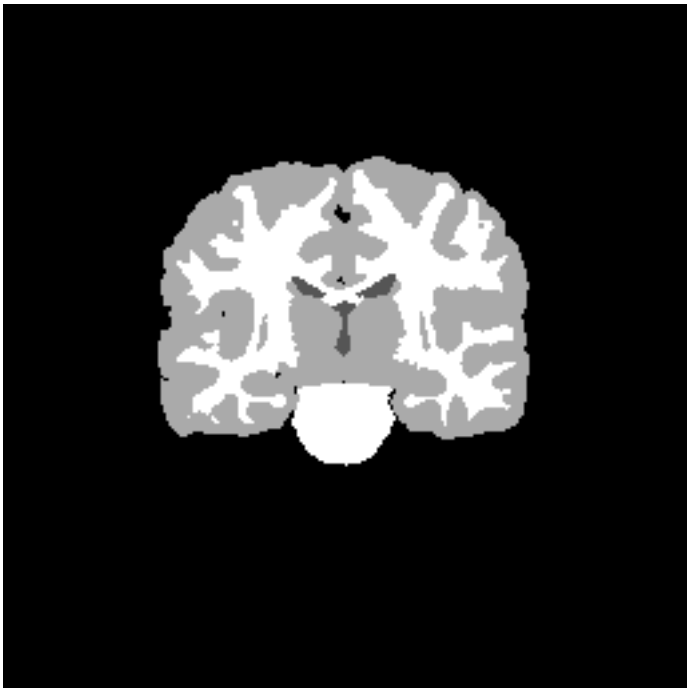

Figura 11: Identificação dos tecidos relativos a substância branca, substância cinzenta (em cinza claro) e líquido cérebro-espinhal (em cinza escuro).

$$\kappa = \frac{N \sum_{i=1}^{m} x_{ii} - \sum_{i=1}^{m} (x_{i+} \times x_{+i})}{N^2 - \sum_{i=1}^{m} (x_{i+} \times x_{+i})}, \tag{16}$$

onde m representa o número de linhas da matriz de confusão (número de classes), x_{ii} é o número de observações na linha i e coluna i, x_{i+} e x_{+i} são totais marginais da linha i e coluna i, respectivamente; e N, o número total de observações. Em síntese, um índice de κ negativa significa que não existe concordância entre o dado de referência e o resultado do classificador, e para $\kappa = 1$, uma "concordância perfeita". Experimentos em diferentes áreas demonstram que o índice κ pode ter várias interpretações e as orientações dependem da aplicação, entretanto a tabela a seguir ilustra uma interpretação para os intervalos de κ:

Note que o valor crítico do parâmetro β no modelo de Potts é $\beta = \ln(1+\sqrt{|\mathcal{L}|})$, e portanto - para as nossas simulações onde $|\mathcal{L}| = 4$, o valor $\beta = 1.2$ representa um valor próximo ao valor crítico; $\beta = 0.54$, um valor bem abaixo a referência e $\beta = 1.5$, um valor acima a referência. Para cada imagem, nós executamos 10 rodadas para o procedimento de validação cruzada. Deste modo, o valor #1 no

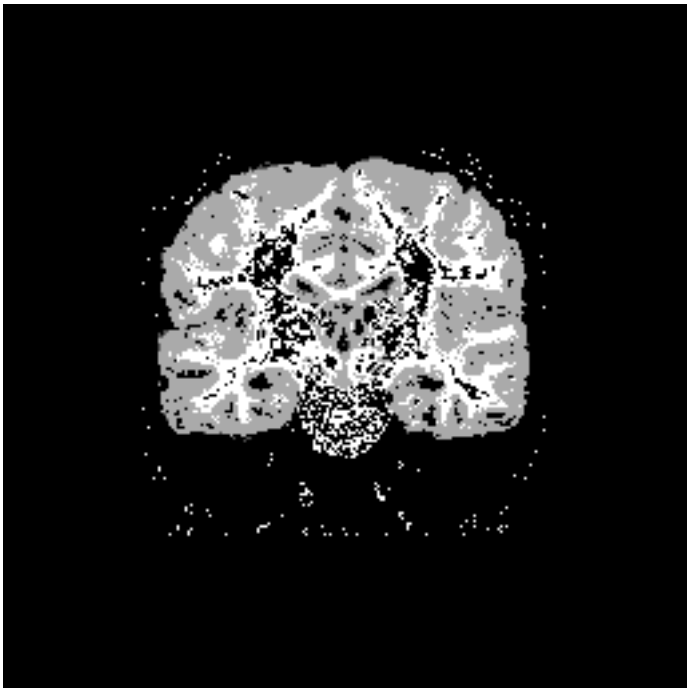

Figura 12: Resultado da etapa inicial para a identificação dos tecidos utilizando o algoritmo Floresta de Caminhos Ótimos.

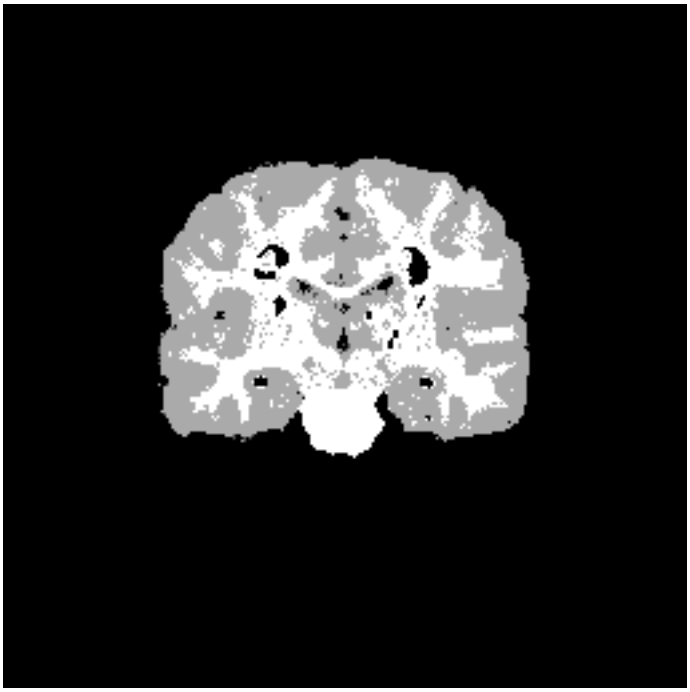

Figura 13: Resultado final após a execução do método proposto ($\beta = 0.54$).

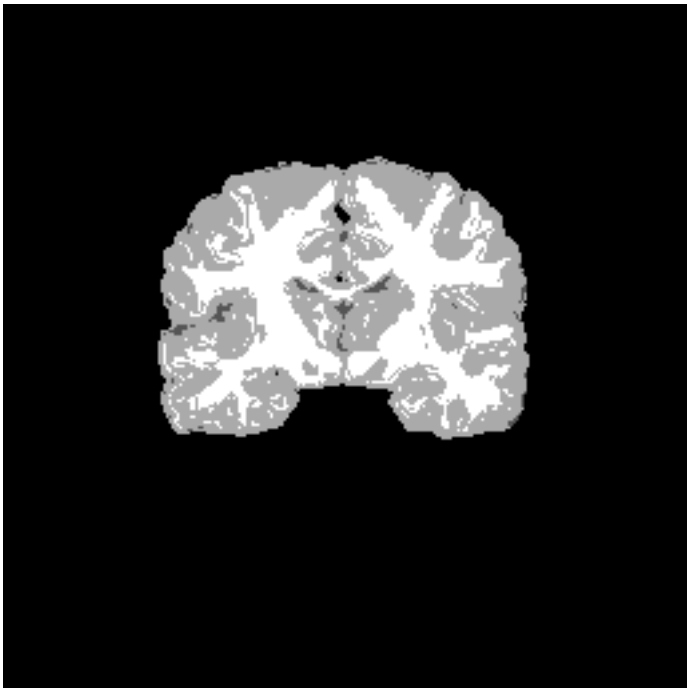

Figura 14: Resultado do processo de identificação dos tecidos após as etapas de pré-processamento utilizando o classificador Floresta de Caminhos Ótimos.

Índice Kappa	**Interpretação**
$\kappa = 1$	Concordância perfeita
$0.8 < \kappa < 1.0$	Concordância quase perfeita
$0.6 < \kappa < 0.8$	Concordância substancial
$0.4 < \kappa < 0.6$	Concordância moderada
$0.0 < \kappa < 0.4$	Concordância baixa
$kappa \leq 0$	Nenhuma concordância

Tabela 1: Possíveis interpretações para os valores de κ.

eixo horizontal para cada β, por exemplo, representa o valor médio para o OPF-MRF em 10 validações cruzadas durante a primeira iteração do algoritmo proposto (nós utilizamos $T = 10$ iterações).

Como podemos observar através dos resultados experimentais obtidos, a inclusão de informações acerca da estrutura do problema pode proporcionar uma solução mais homogênea para a identificação de tecidos em imagens de ressonância magnéticas do que o método de referência (modelo de classificação baseado no algoritmo Floresta de Caminhos Ótimos sob o pressuposto de independência das amostras). Além disso, deve-se salientar que o método proposto é altamente

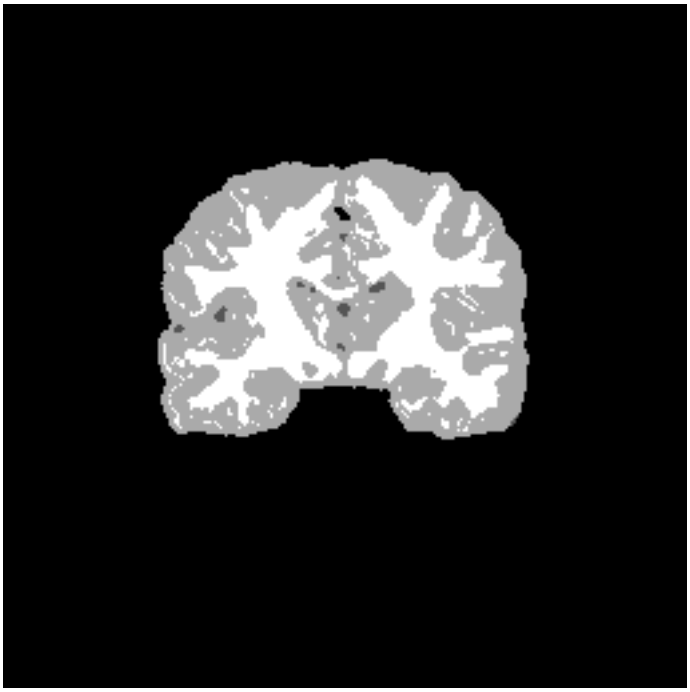

Figura 15: Resultado do processo de identificação dos tecidos após as etapas de pré-processamento utilizando o método proposto.

dependente da solução inicial, deste modo, a utilização de descritores mais robustos para cada amostra, poderá resultar em uma solução ainda mais próxima da solução desejada. O melhor resultado das simulações foi obtido com $\beta = 0.54$, o qual representa o menor valor dentre os avaliados, para representar o compromisso entre as observações e o conhecimento *a priori*.

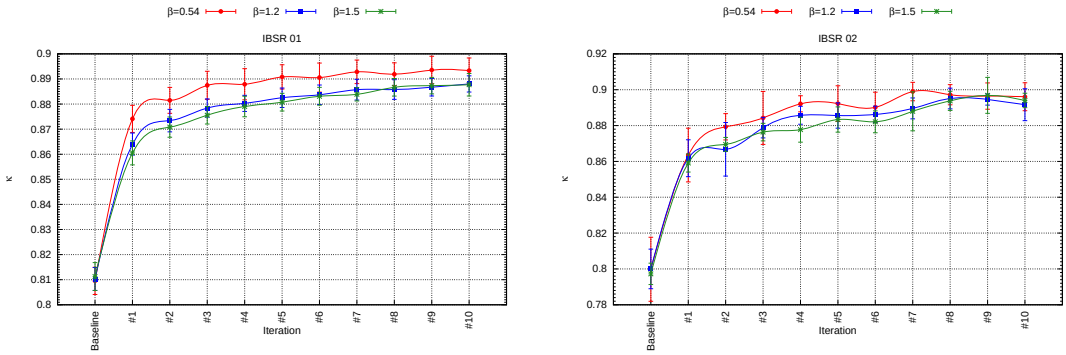

Figura 16: Resultados em termos de κ obtidos aplicando o sistema de classificação proposto para os pacientes 01 e 02.

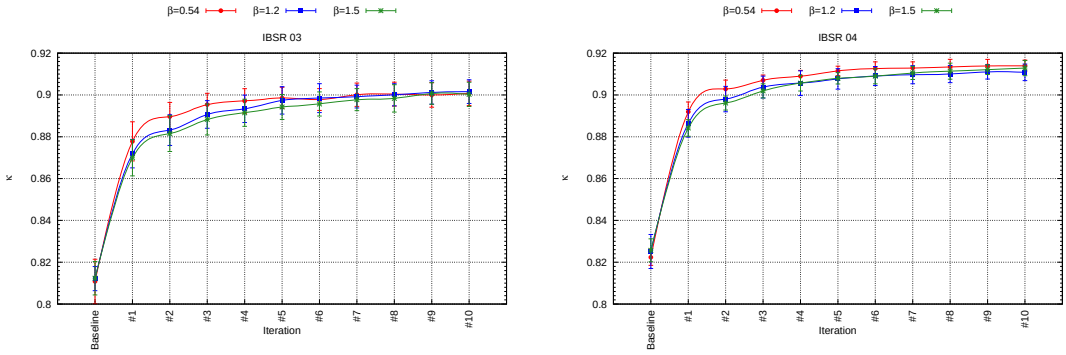

Figura 17: Resultados em termos de κ obtidos aplicando o sistema de classificação proposto para os pacientes 03 e 04.

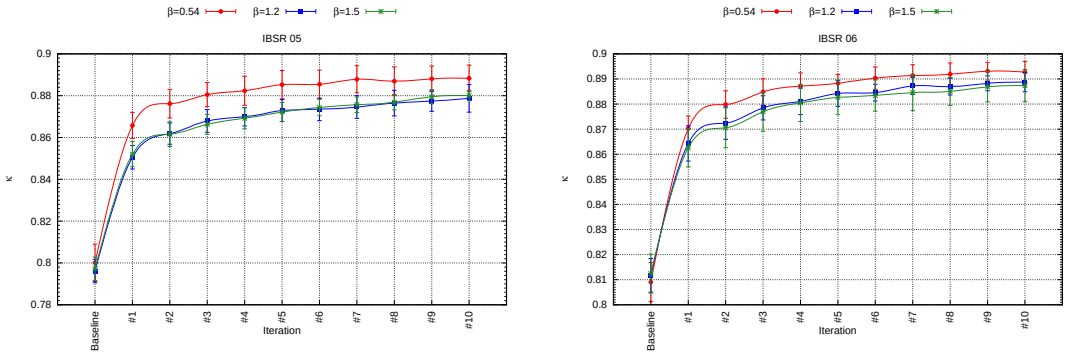

Figura 18: Resultados em termos de κ obtidos aplicando o sistema de classificação proposto para os pacientes 05 e 06.

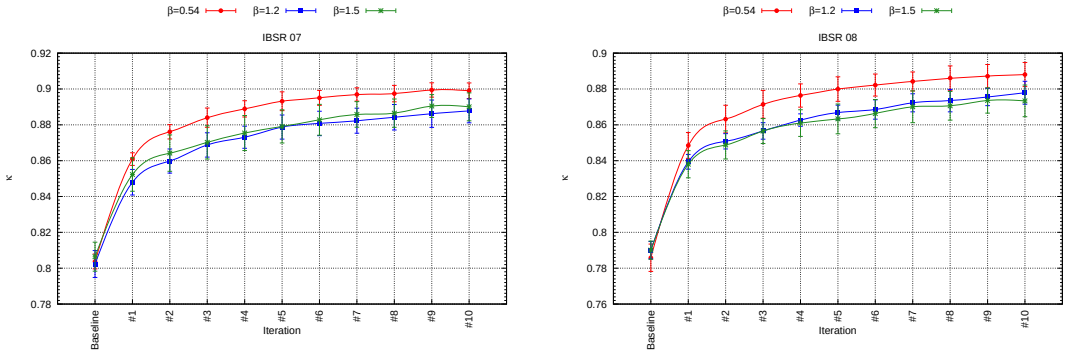

Figura 19: Resultados em termos de κ obtidos aplicando o sistema de classificação proposto para os pacientes 07 e 08.

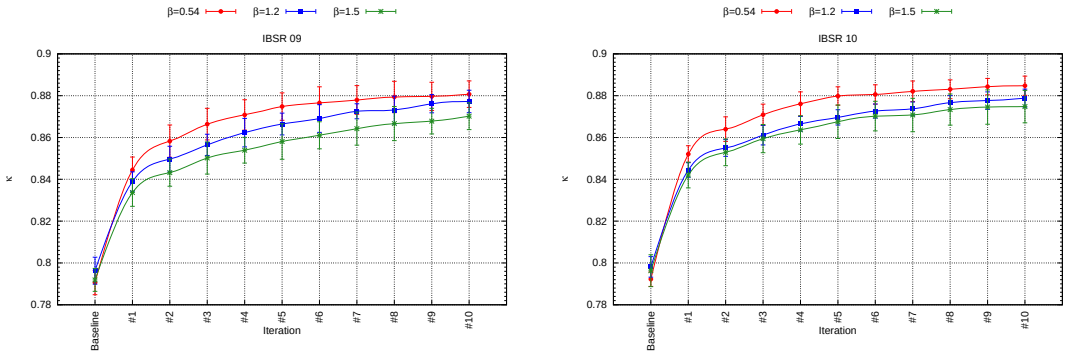

Figura 20: Resultados em termos de κ obtidos aplicando o sistema de classificação proposto para os pacientes 09 e 10.

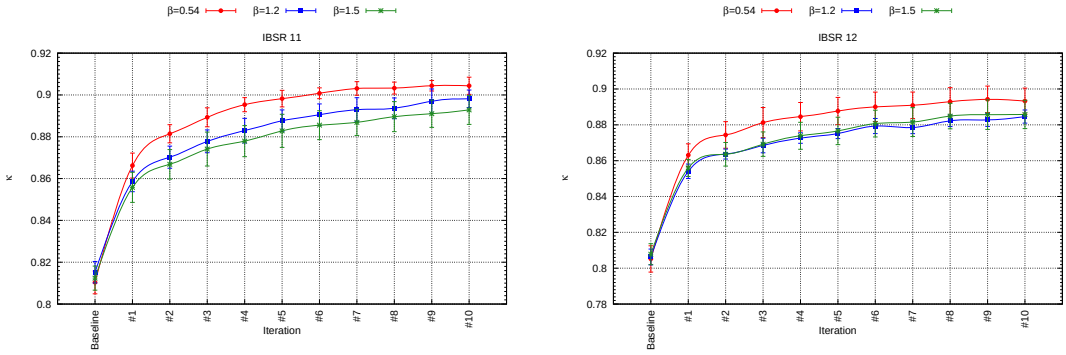

Figura 21: Resultados em termos de κ obtidos aplicando o sistema de classificação proposto para os pacientes 11 e 12.

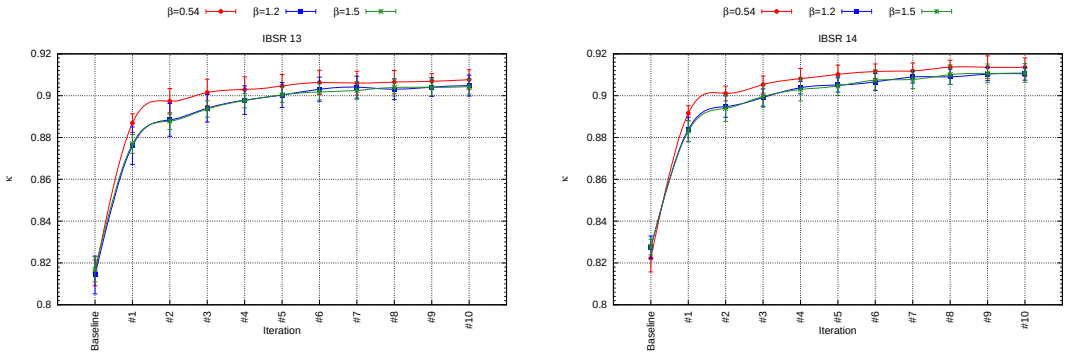

Figura 22: Resultados em termos de κ obtidos aplicando o sistema de classificação proposto para os pacientes 13 e 14.

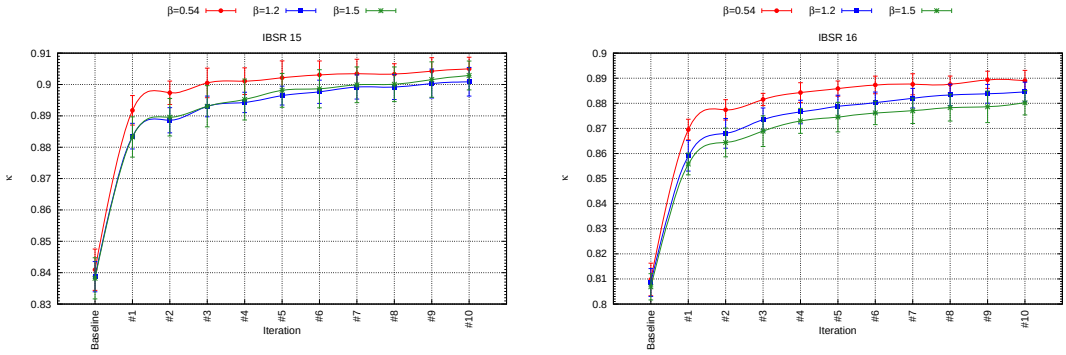

Figura 23: Resultados em termos de κ obtidos aplicando o sistema de classificação proposto para os pacientes 15 e 16.

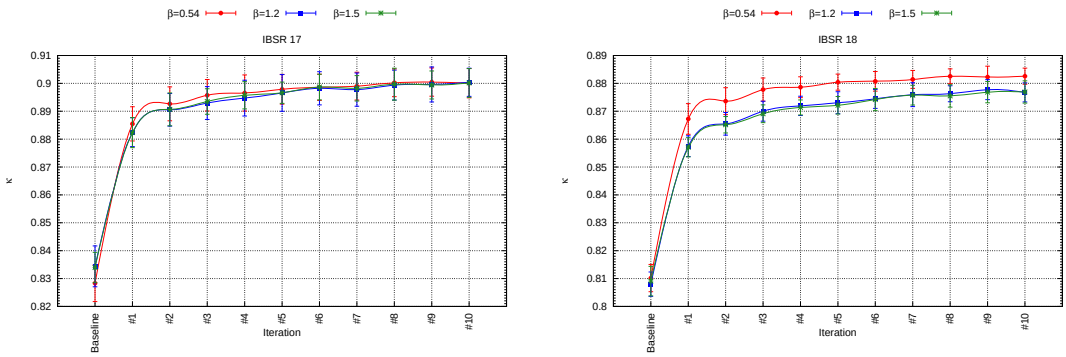

Figura 24: Resultados em termos de κ obtidos aplicando o sistema de classificação proposto para os pacientes 17 e 18.

4 Aprendizado Sequencial Empilhado

Com a disponibilidade atual de grandes quantidades de dados e de recursos computacionais, técnicas de reconhecimento de padrões têm recebido esforços significativos da comunidade científica com o intuito de desenvolver um conjunto de métodos de inferências cada vez mais precisos para a melhor exploração e identificação de novos conhecimentos a partir de análise dos dados. Recentemente, vários trabalhos na literatura propuseram a utilização de informações estruturais do problema (espacial, temporal) para otimizar a precisão do conjunto de previsões, dado que em certos conjuntos de dados existe uma interdependência entre as variáveis de resposta. Por exemplo, um dos problemas fundamentais em visão computacional é a compreensão semântica de uma imagem, isto é, a interpretação de uma imagem em entidades de nosso conhecimento. Na área de processamento de imagens naturais, nós podemos observar que as imagens não são uma coleção aleatória de *pixels* ou regiões, deste modo, a modelagem de uma dependência espacial é importante para uma análise mais precisa. Em um cenário ideal, nós gostaríamos de uma liberdade para modelar complexas e longas interações entre os dados sem nos restringir a pequenas regiões de vizinhança, mas de modo que o algoritmo de inferência se mantenha computacionalmente eficiente.

Cohen e Carvalho [12] propuseram uma abordagem de meta-aprendizado, denominada Aprendizado Sequencial Empilhado, endereçando aplicações de particionamento de sequências, nas quais a informação espacial auxilia no processo de descrição das amostras. Tais tarefas de particionamento sequencial constituem a classe de problemas de classificação caracterizada por longas sequências de rótulos idênticos ou aplicações cujo rótulo de uma amostra é interdependente com o rótulo de um conjunto de amostras próximas: exemplos dessas tarefas incluem a análise de documentos, segmentação de vídeos e a identificação de regiões funcionais de aminoácidos ou cadeias de nucleotídeos para aplicações de bio-informática.

A precisão dos métodos de aprendizado de máquina é altamente dependente da escolha das representações dos dados para as quais o modelo é aplicado. Desta maneira, muito esforços atuais para a construção dos algoritmos de aprendizado de máquina se concentram no projeto dos fluxos de pré-processamento e transformações dos dados, que resultam em uma representação dos dados na qual seja possível o aprendizado de máquina efetivo. O processo de engenharia de características é importante mas requer uma análise intensiva e destaca os pontos negativos dos algoritmos tradicionais de aprendizado: a incapacidade de extrair e organizar informações discriminativas dos dados. A técnica de empilhamento consiste em utilizar um algoritmo de aprendizado de segundo nível para combinar, de forma ótima, uma coleção de previsões realizadas por diferentes modelos. Ao utilizar múltiplos níveis de representações dos dados, o algoritmo de aprendizado torna-se capaz de aprender conceitos abstratos e possivelmente as relações de interdepen-

dência entre as amostras. Muitos cientistas têm obtido sucesso utilizando a técnica de empilhamento e demais métodos de combinação de modelos para melhorar a precisão das previsões realizadas, muito além da precisão obtida por qualquer modelo individualmente. O método amplamente conhecido para fins de detecção de faces (algoritmo de Viola-Jones), conhecido como Adaboost (*Adaptative Boosting*), por exemplo, consiste de um conjunto de classificadores treinados em sequência, onde cada modelo é treinado com uma versão ponderada do conjunto de dados, de modo que o peso do coeficiente associado a cada amostra depende da precisão dos classificadores anteriores. Em síntese, amostras, as quais foram classificadas erroneamente por um dos classificadores, recebem uma maior importância para a construção do classificador seguinte. Após todos os classificadores serem treinados, a previsão final é realizada através de uma votação majoritária entre as previões realizadas individualmente.

A partir do método desenvolvido por Wolpert [50] baseado em empilhamento de classificadores, o aprendizado sequencial empilhado é composto de duas etapas: (i) na primeira etapa, cada amostra do conjunto de treinamento (x, y) representa uma entrada para um modelo de classificação tradicional, construído sob a pressuposição de que as amostras são independentemente distribuídas no espaço de características, definindo o rótulo y'; (ii) em seguida, uma amostra estendida x' é obtida através da expansão do vetor de características de cada amostra com os rótulos de amostras pertencentes a uma certa vizinhança Θ, $x' = \Phi(y', \Theta)$, onde Φ representa a função que modela a interação entre as variáveis pertencentes à vizinhança. Por fim, o novo conjunto de treinamento serve como entrada para a construção de um novo modelo de classificação, no qual a interdependência entre as amostras está representada por este novo conjunto de meta-características. O Algoritmo 5 implementa essa idéia.

Algoritmo 5 – Aprendizado sequencial empilhado

Entrada: Reticulado de vizinhança Θ e o parâmetro para a validação cruzada $\mathcal{K}$

Algoritmo de aprendizado:
  Dado um conjunto $\mathcal{Z} = \{(x_t, y_t)\}$ *e um algoritmo de aprendizado* $\mathcal{H}(x) = \hat{y}$.
  Construa uma amostra estendida com a classe $\hat{y}_t$*, para cada* $x_t \in \mathcal{Z}$*:*
    Divida $\mathcal{Z}$ *em* $\mathcal{K}$ *conjuntos disjuntos* $\mathcal{Z}_1, \ldots, \mathcal{Z}_\mathcal{K}$.
    Para $j \leftarrow 1, \ldots, \mathcal{K}$*:*
      $f_j = \mathcal{H}(\mathcal{Z} \backslash \mathcal{Z}_j)$
    Para $x_t \in \mathcal{Z}_j$*:*
      $\hat{y}_t \leftarrow f_j(x_t)$.
  Construa um conjunto estendido $\mathcal{Z}'$ *de amostras* (x'_t, y_t)*:*
    $x'_t \leftarrow (x_t, \Phi(\hat{y}_i))$*, para todo* $x_i \in \Theta$
  $\mathcal{H}(\mathcal{Z}')$ *construindo o modelo de classificação* $f'(x)$.
Algoritmo de inferência:

13. *Dada uma amostra x e um modelo de classificação $f(x) = \hat{y}$.*
14. $\hat{y} = f(x)$
15. $x' \leftarrow (x, \Phi(\hat{y}_i))$, *para todo* $x_i \in \Theta$
16. *Retorne* $f'(x')$.

Uma das principais vantagens da abordagem de aprendizado sequencial empilhado reside no fato que este procedimento pode ser virtualmente construído sobre qualquer classificador tradicional. Recentemente, Gatta et al. [20] propuseram a utilização de uma função isotrópica[4] Gaussiana como modelo para a interação mútua entre as amostras contidas na região de vizinhança e, portanto, possibilitou a interpretação das interações como uma análise multi-escala do espaço dos rótulos. A possibilidade de representar objetos em múltiplas escalas permite adequar a visualização a distintos níveis de detalhamento. Cabe ressaltar que a resolução espacial da imagem está diretamente relacionada com o conceito de componentes em frequências, que caracterizam o nível de informação de detalhes presentes em uma imagem. Aplicando o filtro Gaussiano com uma evolução do parâmetro variância, a modelagem de longas interações entre as amostras pode ser analisada através das representações obtidas pelas convoluções progressivas com a imagem como a condição inicial de um processo de difusão, podendo ser empregadas as técnicas de decomposição multi-resolução (Seção 4.1) e decomposição pirâmidal do sinal (Seção 4.2).

4.1 Decomposição Multi-resolução

A decomposição multi-resolução deriva diretamente da decomposição multi-resolução em processamento e análise de imagens. Seja $\mathcal{C}_{2,i}$ a probabilidade de uma amostra $\vec{x}$ pertencer a classe i, nós podemos definir a decomposição multi-resolução como:

$$\Theta(\vec{x}, \mathcal{S}) = \mathcal{C}_{2,i}(\vec{x}) * G(0, \gamma^{\mathcal{S}-1}), \tag{17}$$

onde $\mathcal{S}$ representa a escala, $*$, o operador de convolução e G é uma função multidimensional Gaussiana com média zero e $\sigma = \gamma^{\mathcal{S}-1}$. A decomposição multi-resolução produz informações com relação a homogeneidade espacial e a regularidade do campo de rótulos em diferentes escalas. Nós podemos observar, por exemplo, que uma classificação ruidosa em uma escala 1 não influencia os resultados da escala 3. A convolução do sinal com o núcleo Gaussiano tende a eliminar estruturas de uma escala menor, dependendo da variância escolhida. Note que além da linearidade e da invariância por translações, o espaço escala Gaussiano bidimensional é invariante por rotações do domínio.

[4]Uma função densidade $f : \mathbb{R} \to \mathbb{R}_+$ é dita isotrópica, se o seu centróide é a origem, e sua matriz de covariância é a matriz identidade.

4.2 Decomposição Piramidal

A decomposição piramidal é uma decomposição linear multi-resolução na qual o sinal é subdividido baseado em diferentes níveis de escala e orientações. Uma implementação da estrutura piramidal pode ser recursivamente construída através da sub-amostragem dos filtros de passas baixas por um fator igual a 2 ao longo das orientações. Apesar de selecionar subconjuntos da decomposição multi-resolução, teoricamente, não há perda de informação, devido ao fato que, em escalas maiores, os conteúdos de alta frequência são progressivamente filtrados. A Equação a seguir define a decomposição piramidal.

$$\Theta(\vec{x}, \mathcal{S}) = .\mathcal{C}_{2,i}(\mathcal{K}_{\mathcal{S}}\mathcal{S}\vec{x}), \tag{18}$$

onde $\mathcal{K}_{\mathcal{S}}$ representa cada etapa de amostragem e depende de γ, $\mathcal{K}_{\mathcal{S}} = \frac{\gamma^{\mathcal{S}}}{2}$.

4.3 Análise de imagens de sensoriamento remoto

Técnicas de reconhecimento de padrões e mineração de dados aplicadas a imagens de sensoriamento remoto tem sido uma importante ferramenta para a monitoração do meio ambiente em uma escala global, através da identificação de regiões de desmatamento e uso ilegal das terras, bem como para a obtenção de melhores informações acerca dos recursos naturais disponíveis em cada região.

Devido ao crescente aperfeiçoamento dos sensores a bordo, possibilitando a aquisição de imagens com resoluções (espacial, temporal, radiométrica e espectral) cada vez maiores, diversas pesquisas tem se concentrado em desenvolver sistemas de análise automática das imagens, de modo a permitir a criação de mapas temáticos de alta qualidade, estabelecendo inventários precisos dos materiais presentes em cada cena, bem como a detectação em tempo real das mudanças que acontecem na superfície terrestre.

A Figura 25 ilustra uma imagem composta colorida obtida pela câmera imageadora de alta resolução (CCD) do sensor CBERS-2B (*China-Brazil Earth-Resources Satellite*) sobre a área de Itatinga, SP. A câmera CCD fornece imagens com resolução espacial de $20 \times 20m$ e opera em cinco faixas espectrais, sendo elas a faixa pancromática $(0,51 - 0,73\mu m)$, azul $(0,45 - 0,52\mu m)$, verde $(0,52 - 0,59\mu m)$, vermelho $(0,63 - 0,69\mu m)$ e infra-vermelho próximo $(0,77 - 0,89\mu m)$. Para a Figura 26, foram identificadas, de forma manual por um especialista, as áreas de pastagens (verde claro), reflorestamento (verde escuro), culturas (salmão), estradas (cinza), barragens (vermelho) e arbustos (verde).

Em termos das propriedades geométricas de um sistema de sensoriamento remoto, a resolução espacial de um sensor é dada por seu campo de visão e o espectro obtido corresponde à média de reflectância do material dentro deste campo de

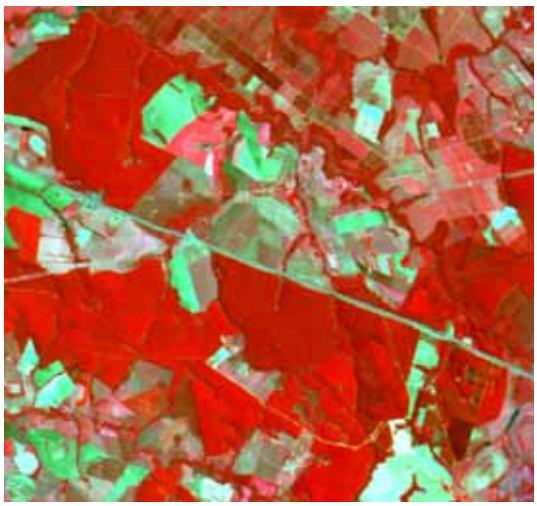

Figura 25: Imagem composta colorida obtida pelo sensor CBERS-2B CCD (20m) (R2G3B4) sobre a área de Itatinga, SP - Brasil (731 x 683 pixels).

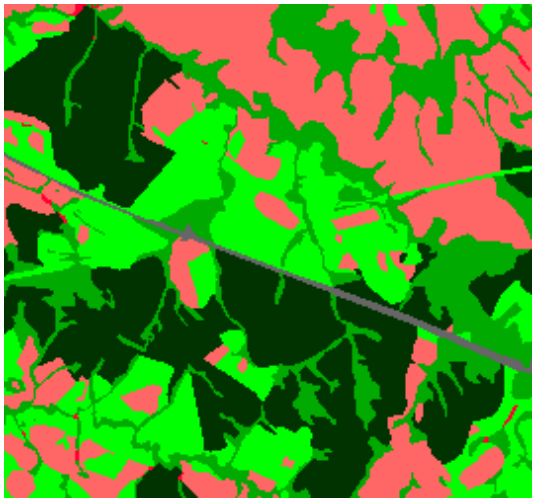

Figura 26: Figura obtida de forma manual por um especialista identificando as regiões de pastagens (verde claro), reflorestamento (verde escuro), culturas (salmão), estradas (cinza), barragens (vermelho) e arbustos (verde).

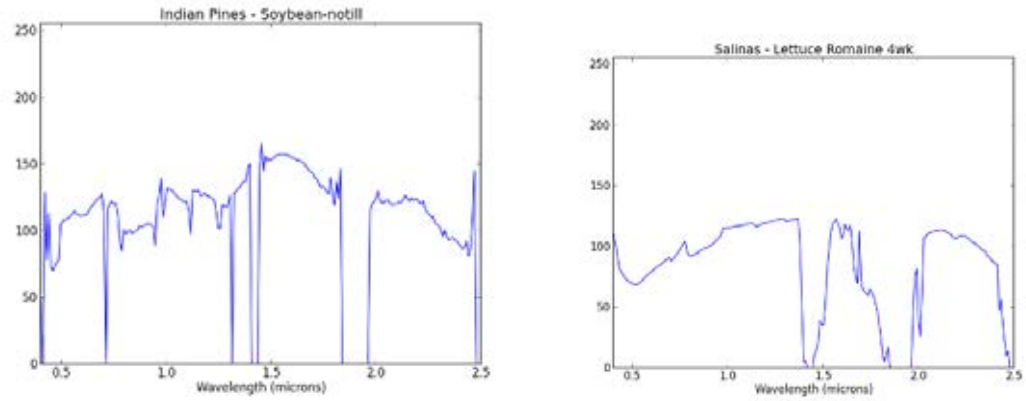

Figura 27: Assinaturas espectrais da soja de plantio direto e alface romana, presentes nas base de dados Indian Pines e Salinas, respectivamente [35].

visão. A resolução espectral é determinada pelo comprimento de onda eletromagnética associado a cada banda espectral. Portanto, a imagem resultado consiste de um hipercubo $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_b}$, contendo $n = n_1 \times n_2$ *pixels* e n_b bandas.

Materiais de diferentes composições, ou seja, com diferentes constituições de elementos químicos de átomos e moléculas, presentes em uma cena podem interagir de forma diferente a exposição a radiação de um determinado comprimento de onda eletromagnética (Figura 27), podendo refletir, absorver ou transmitir a energia a qual foi exposta. A reflectância, fator que mede a capacidade de um objeto de refletir a energia radiante, de um mesmo material pode ser diferente para cada tipo de radiação (comprimentos de onda diferentes), de modo que através de análises experimentais, pode-se definir uma assinatura espectral esperada para cada material baseada nas medidas radiométricas associadas a cada radiação que compõe o espectro eletromagnético. Note que ao mensurar a reflectância, indiretamente, avaliamos a quantidade de energia eletromagnética absorvida pelos materiais, as quais definem de fato as informações sobre a composição dos alvos terrestres nas imagens de sensoriamento remoto. Consequentemente, a partir de uma descrição manual do material referente a pequeno sub-conjunto de amostras de uma cena, técnicas supervisionadas de aprendizado de máquina podem ser empregadas para aprender um modelo de classificação automática. A Figura 28 exibe a classificação da imagem apresentada na Figura 25 utilizando o algoritmo Floresta de Caminhos Ótimos.

Um conjunto representando 1% da imagem foi selecionado aleatoriamente para constituir as amostras anotadas para a construção do modelo de classificação. Cada amostra foi representada pelas respostas espectrais de seus 8 vizinhos mais próximos e, portanto, projetada no espaço euclidiano do $\mathbb{R}^{27}$ (relembre que para

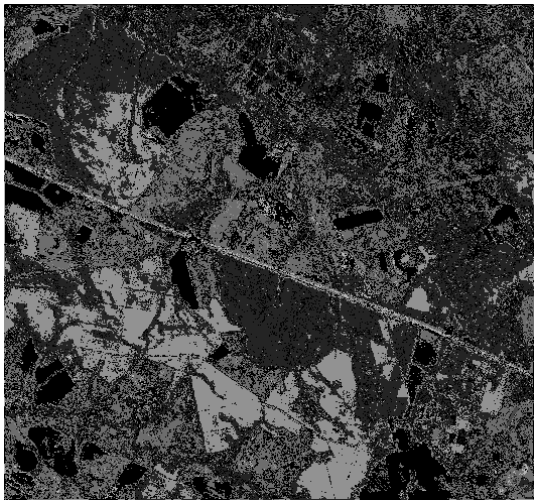

Figura 28: Resultado do processo de classificação utilizando o algoritmo Floresta de Caminhos Ótimos sob a pressuposição de independência dos *pixels*.

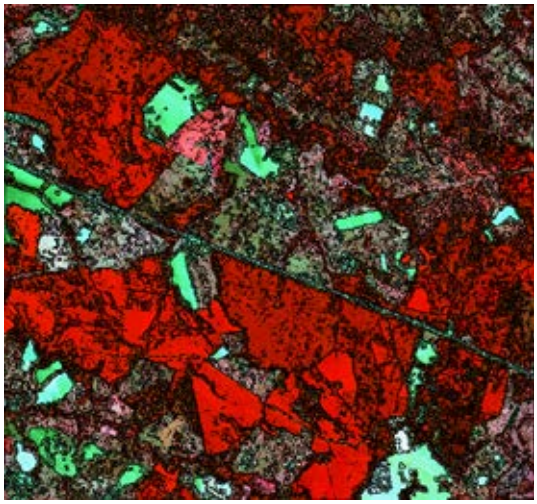

Figura 29: Fronteiras entre as regiões indentificadas utilizando o algoritmo Floresta de Caminhos Ótimos.

cada *pixel*, nós temos a resposta espectral para as faixas verde, vermelha e azul e, portanto cada amostra é representada por $(8 + 1) \times 3 = 27$ características). Nós podemos observar que a interpretação de cada amostra como uma amostra independente resultou em mapas temáticos apresentando um aspecto granulado. Devido a alta correlação espacial, amostras vizinhas tendem a pertencer ao mesmo material, deste modo, torna-se necessário a imposição de restrições de suavidade espacial para estabeler um inventário mais preciso.

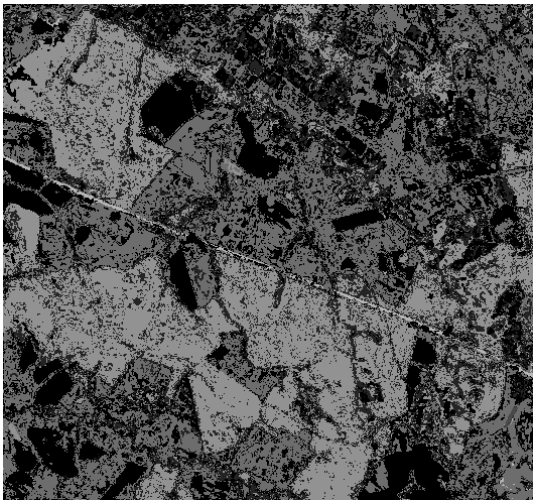

Figura 30: Resultado utilizando o método baseado em Aprendizado Sequencial Empilhado Multi-Escala com decomposição multi-resolução.

As Figuras 30 e 32 ilustram os resultados utilizando uma camada de empilhamento de classificadores com a definição da função de interação entre os componentes do reticulado regular de vizinhança como uma decomposição multi-resolução e piramidal, respectivamente. A escala definida pelo usuário está correlacionada ao tamanho das regiões de interesse, assim é possível balancear o delineamento das regiões e a coerência espacial dos pixels através de um ajuste fino dos parâmetros relacionados ao conjunto de vizinhança e número de representações.

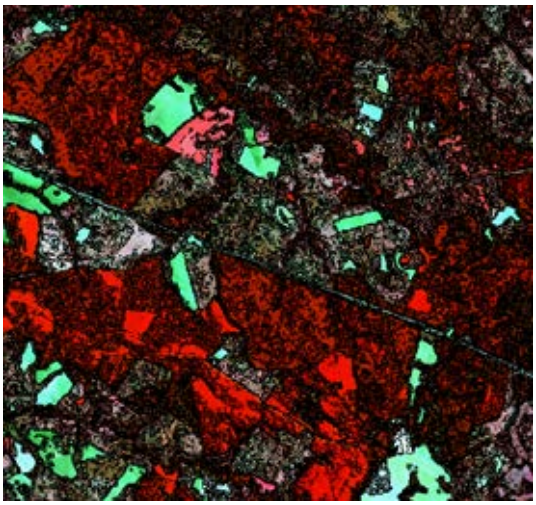

Figura 31: Fronteiras das regiões identificadas utilizando o método baseado em Aprendizado Sequencial Empilhado Multi-Escala com decomposição multi-resolução.

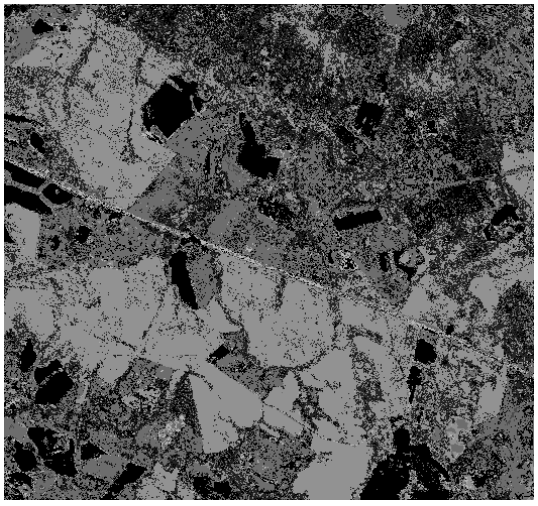

Figura 32: Resultado utilizando o método baseado em Aprendizado Sequencial Empilhado Multi-Escala com decomposição piramidal.

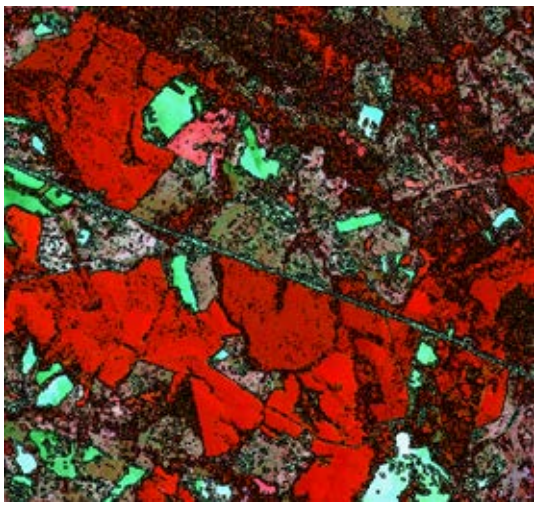

Figura 33: Fronteiras das regiões identificadas utilizando o método baseado em Aprendizado Sequencial Empilhado Multi-Escala com decomposição piramidal.

5 Conclusão

Esta dissertação teve por objetivo o estudo de técnicas sequênciais para representação e classificação de aplicações para as quais a informação estrutural desempenha um papel essencial para uma melhor precisão dos resultados. Dois casos específicos compuseram o cerne deste trabalho: a primeira proposta modela a informação contextual através dos campos aleatórios Markovianos e a segunda abordagem realiza a análise da interdepedência entre as amostras utilizando o espaço-escala através do método de empilhamento de classificadores.

Apesar de recentes trabalhos na literatura investigarem técnicas de otimização para inferência dos parâmetros, como a cardinalidade da clique máxima e interações anisotrópicas, dos modelos em grafos como supracitado na Seção 1, nós investigamos neste trabalho a contribuição de um modelo de campo aleatório definido como um reticulado regular. A Seção 3 introduz o conceito de uma função de energia composta de dois fatores: o primeiro fator representa a classificação baseada no algoritmo Floresta de Caminhos Ótimos, a qual estima a capacidade de cada árvore geradora mínima respectiva a cada classe reivindicar a adesão de uma amostra baseada em uma função de conectividade; o segundo termo sumariza as interações locais dos vértices no modelo de Potts através da penalização dos vértices os quais diferem da moda condicional local. Experimentos realizados para fins de identificação das estruturas anatômicas do cérebro demonstraram que a informação contextual permitiu alcançar uma melhor manutenção da coerência espacial dos rótulos atribuídos a uma mesma região semântica.

A Seção 4 apresenta a modelagem da interdependência entre as variáveis de resposta utilizando uma abordagem de combinação de classificadores. A técnica de empilhamento de classificadores através da definição de uma função de interação entre os vértices proposta por Gatta et al. assemelha-se a definição dos campos aleatórios Gaussianos. Para cada clique definida no grafo, o parâmetro de dispersão da função Gaussiana representa a força da interação espacial entre os vértices do modelo. Experimentos utilizando imagens de sensoriamento remoto foram realizados para demonstrar a capacidade da técnica para a representação da estrutura espacial inerente ao problema de segmentação.

Referências

[1] Kjersti Aas, Line Eikvil, and Ragnar Bang Huseby. Applications of hidden markov chains in image analysis. *Pattern Recognition*, 32(4):703 – 713, 1999.

[2] R. Adams and L. Bischof. Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(6):641–647, 1994.

[3] Richard Aster, Brian Borchers, and Clifford Thurber. *Parameter Estimation and Inverse Problems (International Geophysics)*. Academic Press, San Diego, CA, 2005.

[4] M. Bertero and P. Boccacci. *Introduction to inverse problems in imaging*. Taylor & Francis, Bristol, UK, 1998.

[5] E. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society*, B36:192–236, 1974.

[6] Julian Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society, Series B (Methodological)*, 48(3):259–302, 1986.

[7] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(11):1222–1239, November 2001.

[8] Jinhai Cai and Zhi-Qiang Liu. Pattern recognition using markov random field models. *Pattern Recognition*, 35(3):725 – 733, 2002.

[9] J Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, June 1986.

[10] Yian-Leng Chang and Xiaobo Li. Adaptive image region-growing. *IEEE Transactions on Image Processing*, 3(6):868–872, November 1994.

[11] J. Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37, 1960.

[12] W. W. Cohen and V. R. Carvalho. Stacked sequential learning. In *Proceedings of the 19th international joint conference on Artificial intelligence*, pages 671–676, San Francisco, CA, USA, 2005. Morgan Kaufmann Publishers Inc.

[13] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. The MIT Press, 2 edition, 2001.

[14] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[15] A. H. Seheult D. M. Greig, B. T. Porteous. Exact maximum a posteriori estimation for binary images. *Journal of the Royal Statistical Society*, 51(2):271–279, 1989.

[16] Jennifer L. Davidson, Noel Cressie, and X. Hua. Texture synthesis and pattern recognition for partially ordered markov models. *Pattern Recognition*, 32(9):1475–1505, 1999.

[17] T. G. Dietterich. Machine learning for sequential data: A review. In *Proceedings of the Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition*, pages 15–30, London, UK, 2002. Springer-Verlag.

[18] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.

[19] A.X. Falcão, J. Stolfi, and R.A. Lotufo. The image foresting transform theory, algorithms, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(1):19–29, 2004.

[20] Carlo Gatta, Eloi Puertas, and Oriol Pujol. Multi-scale stacked sequential learning. *Pattern Recognition*, 44(10-11):2414–2426, October 2011.

[21] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transaction on Pattern Analysis and Machine Intelligence.*, 6(6):721–741, November 1984.

[22] Berthold K. P. Horn and Brian G. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[23] Jörg H. Kappes, Bjoern Andres, Fred A. Hamprecht, Christoph Schnörr, Sebastian Nowozin, Dhruv Batra, Sungwoong Kim, Bernhard X. Kausler, Jan Lellmann, Nikos Komodakis, and Carsten Rother. A comparative study of modern inference techniques for discrete energy minimization problem. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Washington, DC, USA, 2013. IEEE Computer Society.

[24] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[25] Vladimir Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(10):1568–1583, October 2006.

[26] Vladimir Kolmogorov and Ramin Zabih. What energy functions can be minimized via graph cuts? *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(2):147–159, 2004.

[27] M. Pawan Kumar, Vladimir Kolmogorov, and Philip H. S. Torr. An analysis of convex relaxations for map estimation of discrete mrfs. *J. Mach. Learn. Res.*, 10:71–106, June 2009.

[28] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[29] Alexandre Luis Magalhães Levada. *Combinação de Modelos de Campos Aleatórios Markovianos para Classificação Contextual de Imagens Multiespectrais.* PhD thesis, University of São Paulo, 2010.

[30] Tony Lindeberg. Edge detection and ridge detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):117–156, November 1998.

[31] D. Marr and T. Poggio. Comparative computation of stereo disparity. *Science*, 194:209–236, 1976.

[32] Paulo A. Miranda and Alexandre X. Falcão. Links between image segmentation based on optimum-path forest and minimum cut in graph. *J. Math. Imaging Vis.*, 35(2):128–142, October 2009.

[33] Cristopher Moore, Mats G. Nordahl, Nelson Minar, and Cosma Rohilla Shalizi. Vortex dynamics and entropic forces in antiferromagnets and antiferromagnetic potts models. *Physical Review E*, 60:5344–5351, Nov 1999.

[34] G. Moser and S. B. Serpico. Combining support vector machines and markov random fields in an integrated framework for contextual image classification. *IEEE Transactions on Geoscience and Remote Sensing*, PP(99):1–19, 2012.

[35] R.Y.M. Nakamura, L.M.G. Fonseca, J.A. dos Santos, R. da S.Torres, X.-S. Yang, and J.P. Papa. Nature-inspired framework for hyperspectral band selection. *IEEE Transactions on Geoscience and Remote Sensing*, PP(99):1–1, 2013.

[36] L. G. Nyúl and J. K. Udupa. On standardizing the MR image intensity scale. *Magnetic resonance in medicine*, 42(6):1072–1081, December 1999.

[37] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man and Cybernetics*, 9(1):62–66, 1979.

[38] Nikhil R Pal and Sankar K Pal. A review on image segmentation techniques. *Pattern Recognition*, 26(9):1277 – 1294, 1993.

[39] P.L. Palmer, H. Dabis, and J. Kittler. A performance measure for boundary detection algorithms. *Computer Vision and Image Understanding*, 63(3):476 – 494, 1996.

[40] J. P. Papa, A. X. Falcão, V. H. C. Albuquerque, and J. M. R. S. Tavares. Efficient supervised optimum-path forest classification for large datasets. *Pattern Recognition*, 45(1):512–520, 2012.

[41] J. P. Papa, A. X. Falcão, and C. T. N. Suzuki. Supervised pattern classification based on optimum-path forest. *International Journal of Imaging Systems and Technology*, 19(2):120–131, 2009.

[42] R.B. Potts. Some generalized order-disorder transformations. volume 48, pages 106–109, 1952.

[43] Daniil Ryabko. Pattern recognition for conditionally independent data. *Journal of Machine Learning Research*, 7:645–664, December 2006.

[44] P. K. Sahoo, S. Soltani, A. K.C. Wong, and Y. C. Chen. A survey of thresholding techniques. *Computer Vision, Graphics, and Image Processing*, 41(2):233–260, February 1988.

[45] Bogdan Savchynskyy. A bundle approach to efficient map-inference by lagrangian relaxation. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR '12, pages 1688–1695, Washington, DC, USA, 2012. IEEE Computer Society.

[46] T. V. Spina, Alexandre X. Falcão, and P. A. V. de Miranda. Intelligent understanding of user interaction in image segmentation. *International Journal of Pattern Recognition and Artificial Intelligence*, 26:1265001–1–1265001–26, 2012.

[47] Richard Szeliski, Ramin Zabih, Daniel Scharstein, Olga Veksler, Vladimir Kolmogorov, Aseem Agarwala, Marshall Tappen, and Carsten Rother. A

comparative study of energy minimization methods for markov random fields with smoothness-based priors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(6):1068–1080, June 2008.

[48] Y. Tarabalka, M. Fauvel, J. Chanussot, and J.A. Benediktsson. SVM- and MRF-based method for accurate classification of hyperspectral images. *IEEE Geoscience and Remote Sensing Letters*, 7(4):736–740, 2010.

[49] Josh Tenenbaum. *A Bayesian Framework for Concept Learning*. PhD thesis, Massachusetts Institute of Technology, 1999.

[50] David H. Wolpert. Stacked generalization. *Neural Networks*, 5:241–259, 1992.

[51] Chee Sun Won and Robert M. Gray. *Stochastic Image Processing (Information Technology: Transmission, Processing, and Storage)*. Kluwer Academic/Plenum Publishers, New York, USA, 2004.

[52] F.Y. Wu. The potts model. *Reviews of Modern Physics*, 54:235–268, Jan 1982.

[53] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Trans. Inf. Theor.*, 51(7):2282–2312, July 2005.