sid.inpe.br/mtc-m21b/2014/03.11.18.59-TDI

UMA EXTENSÃO DO CLASSIFICADOR K-NN PARA MÚLTIPLOS ESPAÇOS

Flávia de Toledo Martins Bedê

Tese de Doutorado do Curso de Pós-Graduação em Computação Aplicada, orientada pelos Drs. Sandra Aparecida Sandri, e Luciano Vieira Dutra, aprovada em 26 de fevereiro de 2014.

URL do documento original:
<http://urlib.net/8JMKD3MGP5W34M/3FT6D55>

INPE
São José dos Campos
2014

PUBLICADO POR:

Instituto Nacional de Pesquisas Espaciais - INPE
Gabinete do Diretor (GB)
Serviço de Informação e Documentação (SID)
Caixa Postal 515 - CEP 12.245-970
São José dos Campos - SP - Brasil
Tel.:(012) 3208-6923/6921
Fax: (012) 3208-6919
E-mail: pubtc@sid.inpe.br

sid.inpe.br/mtc-m21b/2014/03.11.18.59-TDI

UMA EXTENSÃO DO CLASSIFICADOR K-NN PARA MÚLTIPLOS ESPAÇOS

Flávia de Toledo Martins Bedê

Tese de Doutorado do Curso de Pós-Graduação em Computação Aplicada, orientada pelos Drs. Sandra Aparecida Sandri, e Luciano Vieira Dutra, aprovada em 26 de fevereiro de 2014.

URL do documento original:
<http://urlib.net/8JMKD3MGP5W34M/3FT6D55>

INPE
São José dos Campos
2014

Dados Internacionais de Catalogação na Publicação (CIP)

Bedê, Flávia de Toledo Martins.

B39u Uma extensão do classificador k-NN para múltiplos espaços / Flávia de Toledo Martins Bedê. – São José dos Campos : INPE, 2014.

xxvi + 126 p. ; (sid.inpe.br/mtc-m21b/2014/03.11.18.59-TDI)

Tese (Doutorado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2014.

Orientadores : Drs. Sandra Aparecida Sandri, e Luciano Vieira Dutra.

1. vizinhos mais próximos (VMP). 2. vizinhos mais próximos em múltiplos espaços(ms-NN). 3. relações difusas. I.Título.

CDU 528.854

Aprovado (a) pela Banca Examinadora em cumprimento ao requisito exigido para obtenção do Título de *Doutor(a)* em

Computação Aplicada

Dr. Rafael Duarte Coelho dos Santos

Presidente / INPE / SJCampos - SP

Dra. Sandra Aparecida Sandri

Orientador(a) / INPE / SJCampos - SP

Dr. Luciano Vieira Dutra

Orientador(a) / INPE / SJCampos - SP

Dr. Stephan Stephany

Membro da Banca / INPE / SJCampos - SP

Dra. Ana Carolina Lorena

Convidado(a) / UNIFESP / São Paulo - SP

Dr. Getúlio Teixeira Batista

Convidado(a) / UNITAU / Taubaté - SP

Este trabalho foi aprovado por:

( ) maioria simples

(X) unanimidade

Aluno (a): *Flávia de Toledo Martins Bedê*

São José dos Campos, 26 de Fevereiro de 2014

"Você não tem o direito de sair da presença de uma pessoa sem fazê-la melhor e mais feliz.".

Madre Tereza de Calcuta

À minha filha Francesca.

AGRADECIMENTOS

Agradeço a Deus acima de tudo e a todos que contribuíram de alguma forma para que esse trabalho pudesse ser concluído, em especial:

Ao meu esposo Fred, pela ajuda com as implementações e pelo amor, torcida e apoio.

À meus pais Clemente e Cândida, por todo o amor, carinho, apoio incondicional, suporte logístico, afetivo e espiritual, além da grande torcida.

À minha orientadora, Dra. Sandra Sandri, pela contribuição pessoal e intelectual e pelo enorme apoio à conclusão dessa tese.

Ao meu orientador, Dr. Luciano Vieira Dutra, pela sabedoria, paciência, confiança, incentivo e apoio no desenvolvimento do trabalho.

Aos amigos da Senzala que compartilharam todos os momentos comigo. À Eliana, que além do incentivo, contribui significativamente em todas as etapas do doutorado. Ao Henrique pelo empréstimo do computador para processar os resultados. Ao Leonardo Torres por sanar minhas dúvidas do latex. À Maria Antônia pela geração dos mapas. À Mariane pelos dados cedidos e pela ajuda com o processamento dos resultados. À Rafaela pela ajuda com o processamento dos resultados. Ao Rogério pelas leituras da tese antes da entrega para a banca.

Aos Professores Dr. Camilo e Guaraci pela atenção dispensada a todas as minhas dúvidas.

Aos professores do curso de Computação Aplicada por compartilhar seus conhecimentos e experiências em suas aulas.

Aos funcionários da biblioteca, em especial à Yolanda e André, pela forma atenciosa e eficiente de trabalhar.

Ao Instituto Nacional de Pesquisas Espaciais pela oportunidade e por todo suporte concedido.

À CAPES pelo auxílio financeiro.

RESUMO

O algoritmo dos k vizinhos mais próximos (k-NN, do inglês k Nearest Neighbors) é uma técnica de classificação bastante popular em reconhecimento de padrões. Essa técnica consiste em atribuir uma classe a um elemento com rótulo desconhecido usando a classe da maioria de seus vizinhos mais próximos, segundo uma determinada distância no espaço de atributos. A sua versão estendida, proposta neste trabalho, é identificada como vizinhos mais próximos em múltiplos espaços (ms-NN, do inglês: *multiple space Nearest Neighbors*). Essa versão incorpora na construção do modelo a utilização de múltiplos espaços de atributos semanticamente distintos. Para cada espaço, é definido um número de vizinhos e um tipo de distância. Além disso, a localização geográfica dos objetos pode ser incluída como um espaço, desde que estes estejam representados por polígonos ou pontos. A construção do modelo ms-NN baseia-se na procura de vizinhos para cada distância utilizada. Neste trabalho, também é proposto o uso de relações difusas para avaliar a proximidade entre casos. O método ms-NN proposto foi usado para mapear o risco da prevalência da esquistossomose no estado de Minas Gerais e para classificação da cobertura da terra na região amazônica (Tapajós). As classificações resultantes do ms-NN foram comparadas com as classificações do k-NN e de outras duas técnicas, árvore de decisão e máquinas de vetores de suporte (SVM, do inglês: Support Vector Machine), comumente usadas em problemas de reconhecimento de padrões. As acurácias das classificações de cada modelo foram avaliadas usando o método Monte Carlo. Também foi feito o teste T pareado para avaliar a diferença estatística entre as classificações. Os resultados demonstram uma melhora significava das acurácias do método proposto quando comparadas com as acurácias obtidas pelo k-NN.

AN EXTENSION OF K-NN CLASSIFIER FOR MULTIPLE SPACES

ABSTRACT

The k Nearest Neighbors model (k-NN) is a popular classification method used in pattern recognition. This technique assigns to an unlabeled pattern the class from the majority of its nearest neighbors, according to a given distance in the attribute space. The extended version of k-NN, proposed in this work, is named multiple space Nearest Neighbors (ms-NN). This version includes the use of multiple semantically distinct attribute spaces in the model generation. For each space, it is defined a number of neighbors and a distance metric. Besides, the geographic information of the objects can be included as a separate space, provided that they are represented as points or polygons. In this work, we also propose the use of fuzzy relations to evaluate the proximity among cases. The proposed ms-NN method was used to map the risk of schistomiasis prevalence in the state of Minas Gerais and for land cover classification in the Amazon region (Tapajós). The results were compared with classifications obtained with k-NN and two other classifiers, decision tree and support vector machine (SVM), commonly used in pattern recognition problems. The global accuracy of each model was evaluated using Monte Carlo method. The paired T test was used to evaluate the statistical difference among the accuracies. The results show a significative improvement of the classifications when compared to k-NN.

LISTA DE FIGURAS

Pág.

2.1 Exemplo de árvore de decisão. 6
2.2 Exemplo de classificação usando k-NN. 8

4.1 Localização das áreas de estudo. 21
4.2 Distribuição da esquistossomose no estado de Minas Gerais. 22
4.3 Dados matriciais Tapajós 26
4.4 Amostras Tapajós 28

5.1 Distribuição da esquistossome em municípios do estado de Minas Gerais. 35
5.2 Distribuição das amostras em treinamento e teste. 36
5.3 ms-NN com até 4 espaços ATR, usando todos os conjuntos de atributos com a abordagem com 3 classes para Schisto por município. 39
5.4 ms-NN com até 4 espaços ATR e 1 espaço Geo, usando todos os conjuntos de atributos com a abordagem com 3 classes, para Schisto por município. 40
5.5 Comparação dos melhores resultados do ms-NN com os métodos da literatura, quando se usam todos os conjuntos de atributos com a abordagem com 3 classes, para Schisto por município. 40
5.6 Classificação selecionada para abordagem com 3 classes. 43
5.7 Comparação dos resultados do ms-NN com o k-NN, quando se usam 2 conjuntos de atributos com a abordagem com 4 classes usando 25% dos casos indene, para Schisto por município. 43
5.8 Comparação dos resultados do ms-NN com o k-NN, quando se usam 3 conjuntos de atributos com a abordagem com 4 classes usando 25% dos casos indene, para Schisto por município. 44
5.9 ms-NN com até 4 espaços ATR, usando todos os conjuntos de atributos com a abordagem com 4 classes usando 25% dos casos indene, para Schisto por município. 45
5.10 ms-NN com até 4 espaços ATR e 1 espaço Geo, usando todos os conjuntos de atributos com a abordagem com 4 classes usando 25% dos casos indene, para Schisto por município. 46
5.11 Comparação dos melhores resultados do ms-NN com os métodos da literatura, quando se usam todos os conjuntos de atributos com a abordagem com 4 classes usando 25% dos casos indene, para Schisto por município. . 47
5.12 Classificação selecionada para abordagem com 4 classes, usando 25% dos casos indene. 49

5.13 Comparação dos resultados do ms-NN com o k-NN, quando se usam 2 conjuntos de atributos com a abordagem com 4 classes usando 25% dos casos indene, para Schisto por município. 49
5.14 Comparação dos resultados do ms-NN com o k-NN, quando se usam 3 conjuntos de atributos com a abordagem com 4 classes usando 25% dos casos indene, para Schisto por município. 50
5.15 ms-NN com até 4 espaços ATR, usando todos os conjuntos de atributos para a abordagem com 4 classes usando todos os casos indene, para Schisto por município. 51
5.16 ms-NN com até 4 espaços ATR e 1 espaço Geo, usando todos os conjuntos de atributos para a abordagem com 4 classes usando todos os casos indene, para Schisto por município. 51
5.17 Comparação dos melhores resultados do ms-NN com os métodos da literatura, quando usam-se todos os conjuntos de atributos para a abordagem com 4 classes com todos os casos indene, para Schisto por município. . . 52
5.18 Classificação selecionada para abordagem com 4 classes, usando todos os casos indene. 54
5.19 Comparação dos resultados do ms-NN com o k-NN, quando se usam 2 conjuntos de atributos com a abordagem com 3 classes, para Schisto por município. 55
5.20 Comparação dos resultados do ms-NN com o k-NN, quando se usam 3 conjuntos de atributos com a abordagem com 3 classes, para Schisto por município. 55
5.21 ms-NN com até 3 espaços ATR, usando todos os conjuntos de atributos, para Schisto por localidade. 57
5.22 ms-NN com até 3 espaços ATR e 1 espaço GEO, usando todos os conjuntos de atributos, para Schisto por localidade. 57
5.23 Comparação dos melhores resultados do ms-NN com os métodos da literatura, quando se usam todos os conjuntos de atributos, para Schisto por localidade. 58
5.24 Classificação selecionada (ms-NN3+) para Schisto por localidade. 60
5.25 Comparação dos melhores resultados do ms-NN com o k-NN, quando se usam 2 conjuntos de atributos, para Schisto por localidade. 60

6.1 Acurácias das classificações representantes. 68
6.2 Índice de desempenho para o teste por pixel e segmento. 69
6.3 Imagens classificadas usando melhores classificações da abordagem de teste por pixel. 70

7.1 Gráfico com os conjuntos difuso (A1, A2 e A3). 80
7.2 Acurácias das classifações representantes (a) e índice de desempenho (b), para a abordagem de teste por Pixel 81
7.3 Acurácias das classifações representantes (a) e índice de desempenho (b), para a abordagem de teste por Segmento 82

A.1 Telas do plug-in ms-NN classification: tela principal e escolha dos pesos . 96
A.2 Opção para escolha do tipo de espaço. 97
A.3 Telas para configurar o tipo de distância em cada espaço, seja de atributos ou geográfico. 97

B.1 Média das acurácias para o método SVM e Árvore de decisão para abordagem com 3 classes. 100
B.2 Média das acurácias para o método k-NN para abordagem com 3 classes 101
B.3 Média das acurácias para o método ms-NN para 2 espaços: 2 ATR (a) e (b); 1 ATR e 1 GEO (c), (d), (e) e (f) para abordagem com 3 classes. . . 101
B.4 Média das acurácias para o método ms-NN para 3 espaços: 3 ATR (a) e (b); 2 ATR e 1 GEO (c), (d), (e) e (f) para abordagem com 3 classes. . . 102
B.5 Média das acurácias para o método ms-NN para 4 espaços: 4 ATR (a); 3 ATR e 1 GEO (b), (c) e (d) para abordagem com 3 classes. 103
B.6 Média das acurácias para o método ms-NN para 5 espaços, 4 ATR e 1 GEO para abordagem com 3 classes. 104
B.7 Média das acurácias para o método SVM e Árvore de decisão para abordagem com 4 classes, usando 25% das amostras indenes. 104
B.8 Média das acurácias para o método k-NN para abordagem com 4 classes, usando 25% das amostras indenes. 104
B.9 Média das acurácias para o método ms-NN para 2 espaços: 2 ATR (a) e (b); 1 ATR e 1 GEO (c), (d), (e) e (f) para abordagem com 4 classes, usando 25% das amostras indenes. 105
B.10 Média das acurácias para o método ms-NN para 3 espaços: 3 ATR (a) e (b); 2 ATR e 1 GEO (c), (d), (e) e (f) para abordagem com 4 classes, usando 25% das amostras indenes. 106
B.11 Média das acurácias para o método ms-NN para 4 espaços: 4 ATR (a); 3 ATR e 1 GEO (b), (c) e (d) para abordagem com 4 classes, usando 25% das amostras indenes. 107
B.12 Média das acurácias para o método ms-NN para 5 espaços, 4 ATR e 1 GEO para abordagem com 4 classes, usando 25% das amostras indenes. . 108
B.13 Média das acurácias para o método SVM e Árvore de decisão para abordagem com 4 classes, usando todas as amostras indenes. 108

B.14 Média das acurácias para o método k-NN para abordagem com 4 classes, usando todas as amostras indenes. 108
B.15 Média das acurácias para o método ms-NN para 2 espaços: 2 ATR (a) e (b); 1 ATR e 1 GEO (c), (d), (e) e (f) para abordagem com 4 classes, usando todas as amostras indenes. 109
B.16 Média das acurácias para o método ms-NN para 3 espaços: 3 ATR (a) e (b); 2 ATR e 1 GEO (c), (d), (e) e (f) para abordagem com 4 classes, usando todas as amostras indenes. 110
B.17 Média das acurácias para o método ms-NN para 4 espaços: 4 ATR (a); 3 ATR e 1 GEO (b), (c) e (d) para abordagem com 4 classes, usando todas as amostras indenes. 111
B.18 Média das acurácias para o método ms-NN para 5 espaços, 4 ATR e 1 GEO para abordagem com 4 classes, usando todas as amostras indenes. . 112

C.1 Média das acurácias para o método SVM e Árvore de decisão para schisto em nível local. 113
C.2 Média das acurácias para o método k-NN para Schisto em nível local. . . 113
C.3 Média das acurácias para o método ms-NN para 2 espaços: 2 ATR (a) e (b); 1 ATR e 1 GEO (c), (d), para Schisto em nível local. 114
C.4 Média das acurácias para o método ms-NN para 3 espaços: 3 ATR (a); 2 ATR e 1 GEO (b) e (c) para Schisto em nível local. 115
C.5 Média das acurácias para o método ms-NN para 4 espaços: 4 ATR (a); 3 ATR e 1 GEO para Schisto em nível local. 115

D.1 Média das acurácias para o método SVM 118
D.2 Média das acurácias para o método árvore de decisão 118
D.3 Média das acurácias para o método k-NN 119
D.4 Média das acurácias para o método ms-NN para 2 espaços de ATR, usando distância Euclidiana (a), (c) e (e) e Mahalanobis (b), (d) e (f) . . 120
D.5 Média das acurácias para o método ms-NN para 3 espaços de ATR, usando distância Euclidiana (a), (c) e (e) e Mahalanobis (b), (d) e (f) . . 121
D.6 Média das acurácias para o método ms-NN para 4 espaços de ATR, usando distância Euclidiana (d_E) e Mahalanobis (d_M) 122

E.1 Média das acurácias para o método k-NN, usando partições fuzzy trapezoidal (a), (c) e (e) e partições fuzzy triangular (b), (d) e (f) 123
E.2 Média das acurácias para o método ms-NN para 2 espaços de ATR, usando partições fuzzy trapezoidal (a), (c) e (e) e partições fuzzy triangular (b), (d) e (f) 124

E.3 Média das acurácias para o método ms-NN para 3 espaços de ATR, usando partições fuzzy trapezoidal (a), (c) e (e) e partições fuzzy triangular (b), (d) e (f) 125
E.4 Média das acurácias para o método ms-NN para 4 espaços de ATR, usando partições fuzzy trapezoidal e triangular 126

LISTA DE TABELAS

Pág.

4.1 Número de amostras Schisto em nível municipal. 23
4.2 Número de amostras Schisto em nível local. 23
4.3 Tamanho das amostras de Tapajós 27

5.1 Atributos selecionados pela correlação com Pv – nível municipal. 32
5.2 Atributos selecionados pela correlação com Ip – nível local. 33
5.3 Combinações de atributos por espaço. Schisto em nível municipal. 34
5.4 Combinações de atributos por espaço. Schisto em nível local. 34
5.5 Número de classificações resultantes do modelo ms-NN para Schisto. . . . 37
5.6 Média das acurácias das melhores classificações do ms-NN e métodos da literatura com a abordagem com 3 classes para Schisto por município. . . 41
5.7 Matriz de confusão do método árvore de decisão para a abordagem com 3 classes para Schisto por município (Ac = 0,58). 41
5.8 Matriz de confusão do método SVM para a abordagem com 3 classes para Schisto por município (Ac = 0,57). 41
5.9 Matriz de confusão do método k-NN para a abordagem com 3 classes para Schisto por município (Ac = 0,55). 42
5.10 Matriz de confusão do método ms-NN para a abordagem com 3 classes para Schisto por município (Ac = 0,60). 42
5.11 Matriz de confusão do método ms-NN+ para a abordagem com 3 classes para Schisto por município (Ac = 0,71). 42
5.12 Média das acurácias da melhor classificação do ms-NN e k-NN com a abordagem com 3 classes usando 2 conjuntos de atributos para Schisto por município. 44
5.13 Média das acurácias da melhor classificação do ms-NN e k-NN com a abordagem com 3 classes usando 3 conjuntos de atributos para Schisto por município. 44
5.14 Média das acurácias das melhores classificações do ms-NN e métodos da literatura com a abordagem com 4 classes usando 25% dos casos indene, para Schisto por município. 47
5.15 Acurácias das classificações que originaram os conjuntos para o estudo Monte Carlo para a abordagem com 4 classes, usando 25% das amostras indenes, para Schisto por município. 48

5.16 Resumo das matriz de confusão com os erros considerados ruins para a abordagem com 4 classes, usando 25% das amostras indenes, para Schisto por município. 48
5.17 Média das acurácias da melhor classificação do ms-NN e k-NN com a abordagem com 4 classes usando 25% dos casos indene, e 2 conjuntos de atributos para Schisto por município. 50
5.18 Média das acurácias da melhor classificação do ms-NN e k-NN com a abordagem com 4 classes usando 25% dos casos indene, e 3 conjuntos de atributos para Schisto por município. 50
5.19 Média das acurácias das melhores classificações do ms-NN e métodos da literatura, quando usam-se todos os conjuntos de atributos para a abordagem com 4 classes com todos os casos indene, para Schisto por município. 53
5.20 Acurácias das classificações que originaram os conjuntos para o estudo Monte Carlo para a abordagem com 4 classes, usando todas as amostras da classe Indene, para Schisto por município. 53
5.21 Resumo das matrizes de confusão com os erros considerados ruins para a abordagem com 4 classes, usando todas as amostras indenes, para Schisto por município. 54
5.22 Média das acurácias da melhor classificação do ms-NN e k-NN com a abordagem com 3 classes usando 2 conjuntos de atributos para Schisto por município. 56
5.23 Média das acurácias da melhor classificação do ms-NN e k-NN com a abordagem com 4 classes usando todos os casos indene, com 3 conjuntos de atributos para Schisto por município. 56
5.24 Média das acurácias das melhores classificações do ms-NN e métodos da literatura para Schisto por localidade. 59
5.25 Acurácias das melhores classificações do ms-NN e métodos da literatura que deram origem ao conjuntos para o estudo Monte Carlo Schisto por localidade. 59
5.26 Resumo das matriz de confusão com os erros considerados ruins para Schisto por localidade. 60
5.27 Média das acurácias da melhor classificação do ms-NN e k-NN usando 2 conjuntos de atributos para Schisto por localidade. 60

6.1 Número de casos de teste do modelo ms-NN para Tapajós 63
6.2 Atributos – estudo de caso Tapajós. 64
6.3 Combinações de atributos por espaço - Tapajós 65
6.4 Número de classificações para cada método - Tapajós. 66

6.5 Média das acurácias das classificações selecionadas como representantes para as duas abordagens de teste. 67
6.6 Acurácias das classificações selecionadas como representantes usadas para gerar os conjuntos do estudo Monte Carlo, para abordagens de teste por pixel. 70
6.7 Matriz de confusão para o SVM. 70
6.8 Matriz de confusão para a árvore de decisão. 71
6.9 Matriz de confusão para o k-NN. 71
6.10 Matriz de confusão para o ms-NN. 71
6.11 Matriz de confusão para o ms-NN+. 71

7.1 Acurácia da classificação, para as abordagens global e regional, usando modelos de regressão. 77
7.2 Acurácia da classificação, para as abordagens global e regional, usando CBR difusa. 77
7.3 Média das acurácias das classificações selecionadas como representantes usando todos os conjuntos de atributos e conjuntos difusos trapezoidal e triangular. 83

SUMÁRIO

Pág.

1 INTRODUÇÃO 1
1.1 Trabalhos Relacionados 2
1.2 Contextualização 3
1.3 Objetivos 4

2 MÉTODOS DE CLASSIFICAÇÃO 5
2.1 SVM 5
2.2 Árvore de decisão 6
2.3 k-NN 7

3 MÉTODO PROPOSTO 9
3.1 Formalização do k-NN 9
3.2 Método dos vizinhos mais próximos em múltiplos espaços (ms-NN) . . . 11
3.3 Formalização do ms-NN 15

4 MATERIAIS E MÉTODOS 21
4.1 Dados do estudo de caso Schisto 21
4.2 Dados do estudo de caso Tapajós 25
4.3 Metodologia dos experimentos 28

5 EXPERIMENTOS DO ESTUDO DE CASO SCHISTO 31
5.1 Metodologia dos experimentos 31
5.2 Resultados em escala municipal 38
5.2.1 Abordagem com 3 classes 38
5.2.2 Abordagem com 4 classes usando 25% dos casos indenes 45
5.2.3 Abordagem com 4 classes usando todos os casos indenes 50
5.3 Resultados em escala local 56
5.4 Conclusões 61

6 EXPERIMENTOS DO ESTUDO DE CASO TAPAJÓS 63
6.1 Descrição dos experimentos 63
6.2 Resultados e análises 67
6.3 Conclusões 72

7 ms-NN COM RELAÇÕES DIFUSAS 73
7.1 Conceitos básicos da Teoria dos Conjuntos Difusos 73
7.2 Classificação com relações difusas parametrizadas 75
7.3 Classificação com relações difusas compatíveis com ordem 78
7.3.1 Proposta de função de classificação difusa 78
7.3.2 Uso da função f^+ no caso de estudo do Tapajós 79
7.4 Conclusões 82

8 CONCLUSÕES E TRABALHOS FUTUROS 85

REFERÊNCIAS BIBLIOGRÁFICAS 89

APÊNDICE A - PLUG-IN IMPLEMENTADO 95

APÊNDICE B - GRÁFICOS SCHISTO 99

APÊNDICE C - GRÁFICOS SCHISTO LOCALIDADES 113

APÊNDICE D - GRÁFICOS TAPAJÓS 117

APÊNDICE E - GRÁFICOS TAPAJÓS COM RELAÇÕES DIFUSAS 123

1 INTRODUÇÃO

O classificador k vizinhos mais próximos (k-NN, do inglês *k Nearest Neighbors*), introduzido por Fix e Hodges (1951), é uma das técnicas mais populares de reconhecimento de padrões. Essa técnica é muito simples e poderosa, consistindo em atribuir uma classe a um elemento com rótulo desconhecido usando, como função de predominância, a classe da maioria de seus vizinhos mais próximos. No k-NN tradicional, os k vizinhos mais próximos são determinados segundo a distância Euclidiana no espaço de atributos (WEBB, 2002). A sua versão estendida, identificada como k-NN em múltiplos espaços (ms-NN, do inglês *multi space Nearest Neighbors*), proposta neste trabalho, incorpora na construção do modelo a utilização de múltiplos espaços de atributos semanticamente distintos.

No ms-NN, os espaços podem ser de dois tipos: espaço de atributos e espaço geográfico. Uma função de distância é associada a cada espaço, bem como um tipo de vizinhança (fixa ou variável). A classificação de uma amostra não rotulada é feita a partir da união dos vizinhos calculados a partir de todos os espaços, utilizando uma função de predominância. Essa função pode ser a maioria simples, como nos tradicionais k-NN, ou outras, em particular, as ponderadas.

Em tese, para cada espaço pode ser utilizado um número de vizinhos e uma distância diferente. O grande diferencial do modelo ms-NN está na procura por vizinhos para cada espaço utilizado. É possível definir diferentes funções de distância em cada espaço, como Euclidiana, Mahalanobis, Hamming e função $f+ = 1 - S^+$ (SANDRI et al., 2014), onde $S+$ é uma relação difusa com base em partições fuzzy, proposta em Sandri e Martins-Bedê (2014). A função $f+$ é uma pseudométrica quando na partição fuzzy, os conjuntos são trapezoidais e é uma métrica quando são triangulares.

Neste estudo será dada particular atenção à versão que envolve o espaço geográfico, que incorpora a informação espacial (ou geográfica) na classificação de um novo elemento não rotulado. Os vizinhos deste espaço são obtidos a partir da matriz de proximidade generalizada, proposta por Aguiar et al. (2003). A possibilidade de utilizar o espaço geográfico é muito útil em aplicações que envolvem objetos georreferenciados. No ms-NN proposto também é possível utilizar a localização do objecto ou a sua geometria real como atributos de classificação, o que também permite a utilização de associações topológicas.

O modelo ms-NN é aplicado a dois estudos de caso, um em GeoSaúde e um em classificação do uso e cobertura do solo (LUCC, do inglês *Land Use and Cover Clas-*

sification). Na aplicação em GeoSaúde, são usados dados do programa de controle da esquistossomose do Estado de Minas Gerais em nível municipal e local. Na aplicação LUCC, são usados dados de sensoriamento remoto na região amazônica da Floresta Nacional do Tapajós. Por convenção, foi denominado como *Schisto* o estudo de caso com os dados da esquistossomose e como *Tapajós* o estudo de caso com dados de Tapajós. Em ambos casos, é importante considerar separadamente dados de diferentes fontes. Os resultados resultados obtidos pelo ms-NN são comparados com k-NN, Árvore de Decisão e SVM. Para isto, um método de Monte Carlo foi adotada para o teste e foi feito um número exaustivo de parametrizações para os métodos.

Este trabalho está dividido em oito seções principais. Nesta Seção 1 é feita uma breve introdução, são apresentados trabalhos relacionados à abordagem proposta, as inovações e os objetivos deste trabalho. Na Seção 2 são apresentados os métodos de reconhecimento de padrões usados neste trabalho, os quais são comparados com o método proposto. Na Seção 3 é apresentada a formalização da técnica k-NN juntamente com a formalização do modelo ms-NN. A Seção 4 apresenta uma breve descrição das áreas de estudo, dos dados disponíveis e da metodologia dos experimentos. Nas Seções 5 e 6 são apresentados os experimentos do Schisto e Tapajós, respectivamente. Na Seção 7, são apresentadas as relações difusas proposta por Sandri e Martins-Bedê (2014) com os experimentos e resultados. E finalmente na última Seção, são apresentadas conclusões.

1.1 Trabalhos Relacionados

Recentemente, muitas pesquisas têm sido realizadas com o objetivo de aumentar a acurácia da classificação pelo k-NN culminando em duas linhas de pesquisa principais. A primeira é uma linha de trabalho que usa 1 k-NN (1 espaço) e diferentes métricas de distância, com função de predominância maioria (BAY, 1999; YAMADA et al., 2006). Em uma outra linha de trabalho (WANG et al., 2005; SHRIVASTAVA; MEWADA, 2011), são gerados diferentes subconjuntos de atributos com a esperança de melhorar a eficiência final. Nesta linha de trabalho, para cada subconjunto é aplicado o k-NN e os vizinhos mais próximos de cada subconjunto são agregados usando o voto da maioria.

Bay (1999) propôs um método de classificação por vizinhos mais próximos usando vários subconjuntos de atributos. Neste método, a procura pelo vizinho mais próximo usa apenas a distância Euclidiana e o voto da maioria como função de predominância. A ideia geral do método é encontrar a melhor classificação por vizinho mais próximo, a partir de combinações aleatórias de atributos. O método proposto tem

como objetivo basicamente fazer uma seleção de atributos.

Em Bao et al. (2004), foi proposto um método para combinar classificadores k - NN com base em diferentes funções de distância com pesos. Os autores propõem uma nova distância, baseada na distância Euclidiana ponderada, e usam outras distâncias propostas por Bao et al. (1997). No método Dk-NN são introduzidas várias funções de distância e cada uma delas é usada para gerar k amostras mais próximas nos dados de treinamento. Em seguida, as amostras mais próximas são combinadas e a classe de objeto desconhecido é determinada com base na votação da maioria. O algoritmo Dk-NN foi implementado e testado em 7 conjuntos de dados do repositório da Universidade da Califórnia (UCI Machine Learning Repository) com bons resultados.

Wang et al. (2005) propôs o método FC-CMNN (do inglês *Feature subset Clustering for Multiple Nearest Neighbor Classifiers*), baseado em agrupamento de subconjuntos de atributos que são usados para melhorar o desempenho de múltiplos classificadores k-NN. No método FCMNNC, a seleção dos subconjunto de atributos é semelhante à seleção de atributos, onde algoritmos genéticos (MITCHELL, 1996) são utilizados para formar diferentes subconjuntos de atributos, selecionados de acordo com a acurácia das classificação. A decisão final é obtida pela maioria dos votos.

Shrivastava e Mewada (2011) também combinam vários classificadores k-NN, cada qual utilizando um subconjunto de atributos diferentes. Estes subconjuntos de atributos são selecionados através de métodos de busca baseados em otimização da colônia de formigas (ACO do inglês *Ant Colony Optimization*). O objetivo desta abordagem é selecionar os melhores subconjuntos possíveis de atributos a partir do conjunto original usando ACO. Para cada subconjuntos selecionado, é aplicado o k-NN. No final, os classificadores são combinados usando a maioria simples.

1.2 Contextualização

Este trabalho inova em relação aos antecessores em diferentes pontos. É proposto a combinação de classificadores k-NN em múltiplos espaços, de tal forma que se possa atender as seguintes características dos problema em GeoSaúde e Sensoriamento Remoto:

a) Os atributos de entrada da análise provêm de múltiplas fontes;

b) Os dados provenientes de cada fonte podem ser tratados separadamente (possibilitando o acréscimo de novos dados como espaços ao longo do

tempo);

c) Possibilita prover diferentes pesos de acordo com a importância da fonte;

d) Possibilita o uso de uma, ou mais, função de distância de acordo com as características da fonte. Isto é particularmente importante, quando a informação é espacializada quando entra a distância por contiguidade (toca, não toca), por exemplo. Essa é a chamada classificação contextual;

e) Possibilita o uso de funções de predominância mais complexas com a objetivo de manipular vizinhanças com número de variável de vizinhos, como é o caso da distância por contiguidade;

f) Possibilita a integração teórica de esquemas de classificação por vizinhos mais próximos baseados em um raio de influência, junto com outros tipos de vizinhança.

1.3 Objetivos

O objetivo principal deste trabalho é formalizar o método de classificação de vizinhos mais próximos que trabalha com múltiplas fontes (diferentes espaços), múltiplas funções de distâncias, múltiplos tipos de vizinhança e múltiplos métodos de predominância.

Os objetivos secundários incluem:

a) Produzir uma ferramenta dentro do ambiente TerraLib;

b) Testar uma instanciação da teoria desenvolvida para o caso de GeoSaúde;

c) Testar uma instanciação para área de caracterização do uso e cobertura da terra;

d) Incluir relações difusas como base para calcular a distância entre os casos.

2 MÉTODOS DE CLASSIFICAÇÃO

Em reconhecimento de padrões, existem diversos métodos para classificar um caso (i.e. padrão, objeto). Esta tarefa consiste em associar a cada caso um rótulo a partir de um vetor de atributos. A diferença básica entre os classificadores está no método usado para verificar o quanto um caso é semelhante a outro. Este capítulo contém uma breve descrição de três métodos de classificação que servirão de base de comparação com a abordagem proposta, são eles: Máquina de Vetores Suporte (SVM – *Support Vector Machine*), árvore de decisão e k-vizinhos mais próximos (k-NN – *k-Nearest Neighbor*).

2.1 SVM

Máquina de Vetores Suporte é um método de aprendizagem que procura encontrar a superfície de separação ótima entre duas classes, tendo como base um conjunto de padrões selecionados entre as amostras de cada classe (vetores suporte). Considera-se que a melhor superfície de separação é aquela que tem a maior distância em relação aos elementos mais próximos das duas classes distintas (VAPNIK et al., 1996).

No modelo linear de SVM, a superfície de separação ótima é um hiperplano $g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$, onde $\mathbf{w}^T$ é o transposto do vetor ortogonal ao hiperplano de separação e w_0 é um escalar real. Estes parâmetros podem ser encontrados pela maximização da função Lagrangeana (Equação 2.1), também conhecida por sua formulação dual (Equação 2.2)

$$L_p = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{N_A} \alpha_i(y_i(\mathbf{w}^T x_i + w_0) - 1) \tag{2.1}$$

$$L_D = \sum_{i=1}^{N_A} \alpha_i - \frac{1}{2}\sum_{i=1}^{N_A}\sum_{j=1}^{N_A} \alpha_i\alpha_j y_i y_j {x_i}^T x_j \tag{2.2}$$

Nas fórmulas acima, N_A é a quantidade de padrões de treinamento, y_i representa o rótulo da classe (-1 ou 1) e cada $\alpha_i, i = 1, \cdots, N_A$ é um multiplicador de Lagrange.

A formulação do método SVM original permite apenas a separação entre duas classes, o que não atende a maioria dos problemas reais. Para isto são adotadas técnicas multiclasses.

2.2 Árvore de decisão

A árvore de decisão é uma técnica de reconhecimento de padrões e um modelo prático de inferência indutiva. Estas árvores são construídas de acordo com um conjunto de casos previamente classificados. Posteriormente, outros casos são classificadas de acordo com essa mesma árvore. A estratégia dos algoritmos baseados em árvores de decisão é particionar sucessivamente o espaço de busca em subespaços de menores dimensões. As partições são feitas até que cada um dos subespaços contemple apenas uma classe ou até que uma das classes demonstre uma clara maioria, não justificando posteriores divisões. Como é evidente, a classificação consiste apenas em seguir o caminho ditado pelos sucessivos testes colocados ao longo da árvore até que seja encontrada uma folha que conterá a classificação correspondente (FONSECA, 1994). A Figura 2.1 mostra um exemplo de árvore binária univariada.

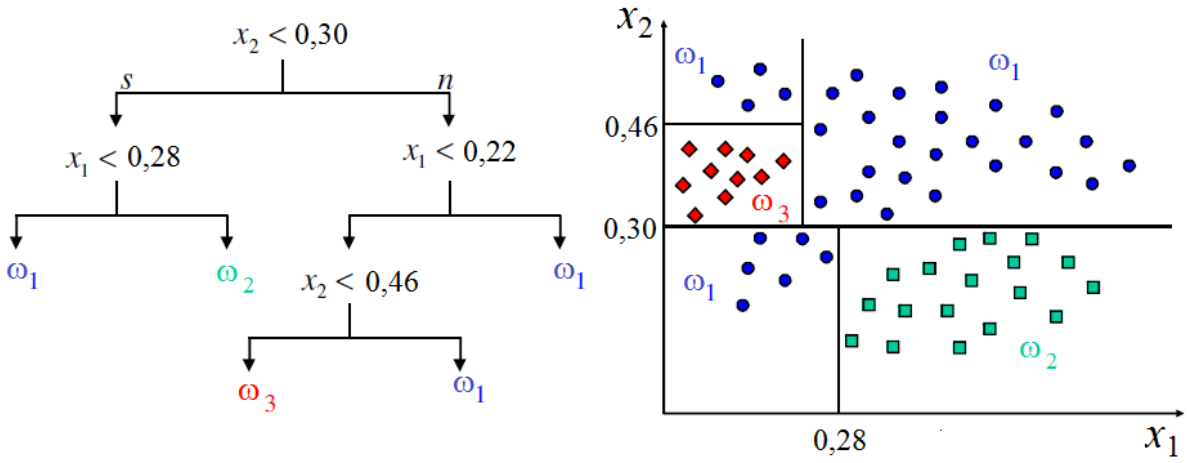


Figura 2.1 - Exemplo de árvore de decisão.
Fonte: Medeiros et al. (2011)

Para a construção de árvores de decisão, são usados algoritmos como o ID3, ASSISTANT e C4.5 (QUINLAN, 1993). O C4.5 não depende de suposições sobre a distribuição dos valores dos atributos. As árvores são formadas por três elementos básicos:

- Nós internos, que representam os atributos;
- Arcos, provenientes dos nós internos e que recebem os valores possíveis para os atributos;

- Nós folha, cada qual contendo a classe majoritária considerando-se o caminho formado pelos arcos entre a raiz e o nó folha em questão.

A árvore de decisão pode ser analisada pelo especialista e, se necessário, pode ser modificada, para então ser convertida em regras que formam a base de conhecimento de um sistema. Cada caminho da raiz até um nó folha corresponde a uma regra de decisão ou classificação.

Uma desvantagem do método é a necessidade da introdução de um critério para a poda da árvore que, caso contrário, poderá se tornar demasiadamente grande e especializada nos exemplos de treinamento. Um critério muito usado é o número mínimo de amostras por folha, definido experimentalmente. Quanto menor o número de amostras em cada folha, maior será o número de regras de decisão. No caso extremo, cada nó folha contém apenas uma amostra.

2.3 k-NN

Em reconhecimento de padrões, o k-NN é um dos algoritmos de aprendizado de máquina mais simples, no qual um objeto é classificado segundo a classe de maior frequência, dentre os k casos vizinhos mais próximos no espaço de atributos. Este processo de classificação pode ser computacionalmente exaustivo para um conjunto de dados muito grandes. Por isso, uma das grandes desvantagens deste método é o tempo de computação para a obtenção dos k vizinhos mais próximos. Por esta razão, a maioria dos estudos envolvendo k-NN tem o objetivo de aumentar a eficiência computacional e diminuir a taxa de erro de generalização deste método (BISHOP, 2006; MICHIE; SPIEGELHALTER, 1994; WEBB, 2002).

k-NN possui apenas um parâmetro livre, o número de vizinhos k, que é controlado pelo usuário com o objetivo de obter uma melhor classificação. Normalmente, os valores de k escolhidos são 1, 2, 3 até $\sqrt{n}$, onde n é o tamanho da base de treinamento (BISHOP, 2006; WEBB, 2002). O melhor valor de k pode também ser determinado experimentalmente. Inicia-se com $k = 1$, e utilizando um conjunto de testes, é feita uma estimativa da taxa de erros do classificador. Para cada k, classificam-se os casos do conjunto de testes e verifica-se quantos casos foram classificados corretamente. O valor de k que resultar na classificação com a menor taxa de erro será o escolhido.

Na Figura 2.2 pode ser visto um exemplo de k-NN com $k = 5$. Neste exemplo, quatro objetos pertencem à classe ω_1, um pertence à classe ω_3 e nenhum objeto pertence à classe ω_2. Neste exemplo o objeto x_j será classificado como a classe ω_1.

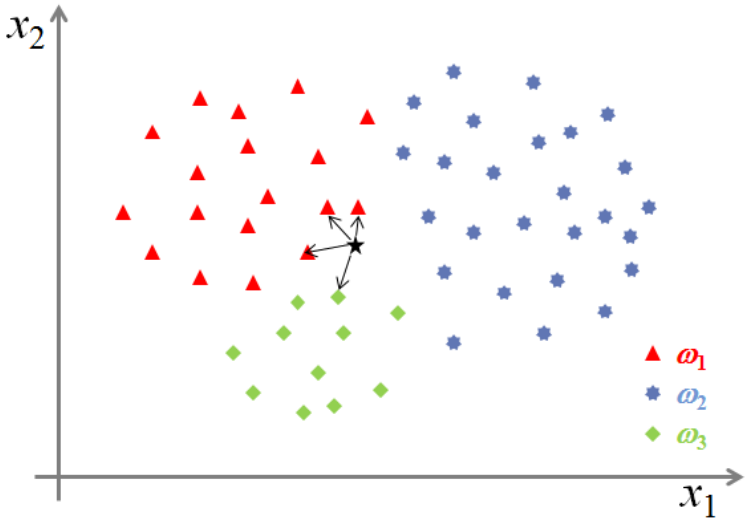


Figura 2.2 - Exemplo de classificação usando k-NN.

No k-NN, pode-se atribuir pesos às contribuições dos vizinhos, de modo que os vizinhos mais próximos contribuam mais para a média dos votos. O mais comum é dado um caso x, atribuir a cada caso vizinho y, o peso de $1/d(x, y)$, onde d é uma função de distância. Também é possível atribuir um peso relativo à importância de cada atributo (BISHOP, 2006; WEBB, 2002).

O objetivo deste trabalho é propor um método que calcule os vizinhos mais próximos em múltiplos espaços, denominado ms-NN. A formalização matemática dos métodos k-NN e ms-NN são apresentadas na Seção 3.1 e 3.3, respectivamente.

3 MÉTODO PROPOSTO

Nesta seção são apresentadas as formalizações do métodos k-NN e ms-NN, além dos algoritmos para cada modelo.

Seja C um conjunto finito e não vazio de casos. Os conjuntos de casos T, P e N formam uma partição do conjunto C, i.e. e $T \cap P \cap N = \emptyset$. O conjunto T é definido como o conjunto de treinamento e possui tr casos ($T = \{c_t \colon t = 1, \cdots, tr\}$). O conjunto P é definido como conjunto de teste e possui te casos ($P = \{c_p \colon p = 1, \cdots, te\}$). O conjunto N é definido como o conjunto de casos não rotulados. Seja $A = \{a_i \colon i = 1, \cdots, ta\}$ o conjunto de atributos, em que cada a_i está definido no universo U_i, sendo $U = \{U_i \colon i = 1, \cdots, ta\}$ e $CL = \{cl_1, \cdots, cl_m\}$, o conjunto das classes.

3.1 Formalização do k-NN

A *hipótese básica do k-NN tradicional* é de que um caso (objeto, padrão, pixel, município, segmento, regiões, etc.) com rótulo desconhecido pode ser classificado de acordo com a maioria das classes dos seus k vizinhos mais próximos.

No k-NN, cada caso $c \in C$ é caracterizado por um par $c = (\mathbf{a}, \omega)$, em que $\mathbf{a} \in U_1 \times \cdots \times U_{ta}$ é um vetor com ta atributos e $\omega \in CL$ é a classe do caso. Neste modelo, a classe $\omega \in CL$ mais frequente é atribuída a cada caso c_0 ($c_0 \in C$) usando os k casos de T mais próximos de c_0 em U. Para tanto, calcula-se a distância, no espaço de atributos, do caso c_0 a todos os casos $c_t \in T$, e produz-se o conjunto $V_0 = \{c_v \colon c_v \in T,\ v = 1, \cdots,\ k\}$, com os k vizinhos mais próximos de c_0 de acordo com a distância estabelecida. A classe ω_0 mais frequente em V_0 é atribuída a c_0.

Como visto na Seção 2.3, o único parâmetro do modelo é k, um inteiro positivo. O algoritmo do k-NN tradicional baseado em Bishop (2006) e Webb (2002) é apresentado, de forma sucinta, no Algoritmo 3.1, que contém as definições das variáveis, o programa principal e as sub-rotinas.

No Algoritmo 3.1, o especialista deve fornecer um conjunto de dados C e um valor para k. O conjunto de dados C deve conter obrigatoriamente um conjunto de treinamento T e um conjunto de teste P e pode ou não conter o conjunto de dados não rotulados. O valor definido para k deve ser maior ou igual a 1. Neste algoritmo, é possível selecionar um subconjunto de atributos, dentre os atributos dos casos.

Para classificar um caso $c_0 \in C$ usando o k-NN, são necessários três elementos básicos

Algoritmo 3.1 Algoritmo do modelo k-NN tradicional

Definição das constantes

k – número de casos de treinamento vizinhos definido pelo especialista;
C – base de dados;

Definição das variáveis

$novoCaso$ – contador de casos;
$nCasos$ – número de casos da base de dados;

Programa Principal

calculeDistâncias
para $novoCaso = 1$ **até** $nCasos$ **faça**
encontre os kVizinhos e conte os rótulos de cada classe
classifique o $novoCaso$ de acordo com a maioria das classes dos seus vizinhos
fim para

Sub rotinas

calculeDistâncias: retorna um conjunto D_0 contendo a distância Euclidiana dos atributos de c_0 aos atributos de cada caso na base de treinamento (c_t).
kVizinhos: retorna um conjunto V_0 com os k casos de treinamento mais próximos de c_0;

cos, uma função de distância, um tipo de vizinhança e uma função de predominância, discutidas a seguir.

a) Função de distância $d(c_0, c_t)$:

Seja $d\colon C \times C \to R$ uma função de distância[1]. Seja D_0 o conjunto com as distâncias dos atributos do caso c_0 a todos os casos c_t em C, ou seja:

$$D_0 = \{d(c_0, c_t)\colon t = 1, \cdots, tr\} \tag{3.1}$$

No k-NN tradicional, a métrica usada para definir o conjunto de distâncias D_0 a um dado caso c_0 é a distância Euclidiana:

$$d(c_0, c_t) = d(a_0, a_t) = [(a_0 - a_t)^T (a_0 - a_t)]^{1/2} \tag{3.2}$$

b) Tipo de vizinhança:

No k-NN tradicional, utiliza-se uma vizinhança fixa, ou seja, para todo

[1] A distância entre dois vetores quaisquer, x e y é definida como $d(x, y) = |x - y|$. Uma função de distância possui as seguintes propriedades: $d(x, y) \geq 0$, $d(x, y) = 0 \Leftrightarrow x = y$, $d(x, y) = d(y, x)$ e $d(x, y) \leq d(x, z) + d(z, y)$

caso $c_0 \in C$, determina-se um conjunto V_0 com o número k fixo de casos vizinhos do conjunto de treinamento T, ou seja:

$$V_0 = \{c_v \in T : v = 1, \cdots, k\} \tag{3.3}$$

São elementos de V_0 os k casos c_t que possuem as menores distâncias à c_0, i.e., os k casos que minimizam $d(c_0, c_t)$.

c) Função de predominância:

No k-NN, a classe ω_0 de um caso c_0 é atribuída segundo a função de predominância *maioria simples*, ou seja:

Seja V_i um subconjunto de V_0 com os casos vizinhos de c_0, classificados como pertencentes a $cl_i \in CL$:

$$V_i = \{c_v \in V_0 : \omega_v = cl_i\}, \tag{3.4}$$

em que ω_v é a classe do caso c_v.

Atribui-se a c_0 a classe mais frequente dentre as m classes cl_i encontradas

$$\omega_0 = \arg\max_i |V_i|; \quad i = 1, \cdots, m \tag{3.5}$$

Em caso de empate, atribui-se a c_0 a classe do seu vizinho mais próximo.

3.2 Método dos vizinhos mais próximos em múltiplos espaços (ms-NN)

A *hipótese básica do ms-NN*, é que um caso com rótulo desconhecido pode ser devidamente classificado de acordo com uma função de predominância mensurada a partir de um conjunto de vizinhanças pré-estabelecidas, cada uma definida em um espaço n-dimensional.

Neste método, os atributos de uma base de dados podem ser divididos em tantos conjuntos quanto se queira, o que possibilita ao especialista tratar conjuntos de atributos diferentes de forma distinta. Neste modelo, os conjuntos de atributos são chamados de *espaços*. Para cada espaço, define-se uma função distância para compor a vizinhança de cada caso.

No k-NN tradicional, utiliza-se distância Euclidiana entre os atributos para a classificação de um novo caso (BISHOP, 2006; WEBB, 2002). O uso exclusivo da distância

Euclidiana é um inconveniente do k-NN tradicional, pois pode acarretar problemas na classificação, visto que essa métrica é sensível a mudanças de escala. No entanto, essa desvantagem pode ser contornada normalizando-se os dados. A normalização dos dados pode ser feita por uma abordagem simplista, por exemplo, transformando o intervalo de valores de cada atributo em valores entre 0 e 1, ou padronizando cada atributo, de forma que apresente média 0 e desvio padrão 1 (padronização estatística).

Outra desvantagem do k-NN é o fato do seu modelo considerar todos os atributos juntos no processo de classificação, independentemente da semântica e do valor de cada atributo. A solução proposta neste trabalho é usar diversas distâncias para conjuntos de atributos distintos em diferentes espaços. Desta forma, por exemplo a distância Euclidiana pode ser usada para um conjunto de atributos em um determinado espaço e a de Mahalanobis para outro conjunto de atributos e espaço, dando ao usuário a opção de normalizar ou não os dados. Além da distância Euclidiana e de Mahalanobis, a distância de Hamming poderia ser usada, a critério do usuário, em espaços contendo atributos que possuem valores como números binários, inteiros e até mesmo texto.

Considerando uma base de dados censitários, em que cada caso é uma residência, podemos ter atributos sobre a característica da família, sobre o responsável pelo domicílio e sobre o saneamento. Dependendo da aplicação, esses atributos podem ser usados em um único espaço com uma única relação de vizinhança, como ocorre no k-NN. Mas em algumas aplicações, usar uma única função de distância para determinar quem são os casos similares a um novo caso pode não ser tão conveniente. No ms-NN, o especialista tem a liberdade de optar por dividir o conjunto de atributos, por exemplo, em três espaços, um com as informações relacionadas com a família, um com as informações relacionados com o responsável pelo domicílio e outro com as informações que refletem as condições sanitárias do domicílio, cada espaço com uma função de distância apropriada e uma relação de vizinhança que julgar mais eficiente.

Além disso, no ms-NN, um caso pode estar associado a objetos geo-referenciados (CASANOVA et al., 2005), tais como pixels numa imagem, municípios num mapa e segmentos obtidos para uma imagem. Neste trabalho, é dada maior atenção a esses casos, em que a posição geográfica e a geometria do objeto são atributos do caso ao qual está associado. Por exemplo, quando o caso está associado a um pixel, sua geometria é um ponto e sua localização é dada pelo par ordenado com as coordenadas

geográficas do ponto na imagem. Já quando o caso está associado a um município, sua localização é dada por um par ordenado cujas coordenadas geográficas podem indicar, por exemplo, a sede do município ou o centróide com sua geometria podendo ser um ponto (contendo sua própria localização) ou um polígono (representando o conjunto de todos os pontos que o compõem).

Neste trabalho são considerados dois tipos de dados: *vetoriais* e *matriciais* (CASANOVA et al., 2005). Para ambos os tipos, cada caso possui um conjunto de atributos e está associado a uma representação geométrica, ponto ou polígono. Para os dados vetoriais os casos são, por exemplo, os municípios de um Estado, que possuem atributos tais como o nome do prefeito, tamanho da população, renda per capta média, índice de desenvolvimento humano (IDH), entre outros. Esses casos podem ter diferentes representações geométricas como polígono, quando se tem os limites do município, e pontos, quando se tem o centroide ou sede do município.

Os dados matriciais, podem advir de imagens de sensoriamento remoto. Neste tipo de dado, cada caso possui atributos como, por exemplo, um valor proporcional à energia eletromagnética refletida ou emitida pela área da superfície terrestre correspondente, para cada banda da imagem. Esses atributos são extraídos da imagem, para cada caso, usando algum pré-processamento. Dependendo da aplicação, a representação geométrica de um caso pode ser um ponto (pixel da imagem) ou um polígono (segmento da imagem).

Como a forma geométrica e posição geográfica dos casos se referem a atributos diferenciados, optou-se por tratá-los de forma distintas, neste trabalho. Sendo assim, por convenção, definem-se dois tipos de espaço: o *espaço geográfico* (GEO) e o *espaço de atributos* (ATR). O espaço GEO contém a informação da geometria espacial de um caso, o que inclui a coordenada geográfica e a representação geométrica do caso em si, enquanto o espaço ATR contém os outros atributos descritivos do caso.

No processo de classificação, dependendo do conjunto de dados, o especialista tem a opção de usar mais de um espaço ATR e/ou um ou mais espaços GEO. Como espaço ATR, pode-se usar todos os atributos em um mesmo espaço, a exemplo do k-NN, ou dividir os atributos em vários espaços. Simultaneamente, pode-se optar por espaços GEO com diferentes representações geométricas. Por exemplo, quando os casos são municípios pode-se ter um espaço GEO com a representação por ponto e outro cuja representação é um polígono. A representação do caso está diretamente relacionada com a função de distância e com o tipo de vizinhança, que serão discutidas a seguir.

Como dito anteriormente, no ms-NN a função de distância não fica restrita à distância Euclidiana como no k-NN. Podem ser usadas, por exemplo, as distâncias de Hamming, Chebyshev, Mahalanobis, Kulback Lieber, Bhattacharyya, etc. Tais distâncias podem ser calculadas a partir de malhas viárias (rodoviárias ferroviárias, hidroviárias), raio de influência, etc. O tipo de vizinhança é um outro diferencial do modelo e pode estar associado à função de distância, dependendo ainda do tipo de espaço. A vizinhança pode ser fixa ou variável, proporcionando mais opções ao usuário.

Na vizinhança fixa, assim como no k-NN, o especialista define um número fixo de vizinhos. Por exemplo, os k vizinhos mais próximos do caso c_0 no espaço ATR são os k casos da base de treinamento, cujos atributos possuem as k menores distâncias a c_0.

No espaço GEO, por exemplo, são considerados os k vizinhos mais próximos de c_0, quando forem usados pontos como representação geometria dos casos. Assim, nesse espaço, serão considerados como vizinhos de c_0 os k casos de treinamento cujas as distâncias das suas posições geográficas são as k menores em relação à posição geográfica de c_0. As distâncias usadas no espaço geográfico poderiam ser, por exemplo, a distância Euclidiana ou uma distância sobre uma malha de estradas. Particularmente, quando a representação geométrica dos dados for um ponto e o tipo de dado, matricial, ou seja, quando um caso é um pixel de uma imagem, o caso c_0 ou pode ter 4 ou 8 vizinhos.

Já na vizinhança variável, o número de vizinhos de cada caso depende da posição do caso e/ou da função de distância selecionada. No espaço ATR, serão vizinhos de c_0 todos os casos de treinamento cujas distâncias entre seus atributos estiverem dentro de um raio de influência definido previamente pelo usuário. Também serão vizinhos por raio de influência, no espaço GEO, todos os casos de treinamento cujas posições geográficas estão a uma distância preestabelecida da posição do caso c_0. Ainda no espaço GEO, quando polígonos forem usados como representação geométrica, os vizinhos do caso c_0 serão todos os casos da base de treinamento que são vizinhos por contiguidade do caso c_0, i.e., que o tocam.

Definidos o número e o tipo de espaço, a vizinhança e a função de distância, é necessário definir a *função de predominância*, que é mais um diferencial do modelo proposto. A função de predominância não fica restrita à função de maioria simples. No ms-NN, pode-se optar pela função de maioria ou por uma função de predominância ponderada. A função de predominância ponderada pode ser pelo número de

vizinhos, pelo tipo de vizinhança, com o objetivo de equilibrar o cômputo das classes, ou pode ser dado um peso específico para cada espaço.

Na função ponderada pelo número de vizinhos, todos os espaços passam a ter peso igual no cômputo das classes, independentemente do número de vizinhos que o espaço possua. Outra forma de ponderação é atribuir a cada espaço um fator de ponderação, o que resultará em peso igual para os vizinhos dentro do espaço e peso diferente para vizinhos entre espaços. Este segundo caso, em que o fator de ponderação é usado para equilibrar o cômputo das classes pelo tipo de vizinhança, é denominado "função ponderada pelo tipo de vizinhança". Esta função é indicada quando selecionados os dois tipos de vizinhança, já que na função maioria simples, um espaço com vizinhança variável pode ter maior influência no processo de classificação em relação a um espaço com vizinhança fixa. Além disso, o espaço com vizinhança fixa pode não influenciar na classificação, já que um caso pode ter um número muito grande de vizinhos pela vizinhança variável e poucos vizinhos pela vizinhança fixa.

Baseado no Algoritmo 3.1, neste trabalho é proposto o Algoritmo 3.2 que apresenta o ms-NN como uma generalização do modelo k-NN tradicional. O Algoritmo 3.1, o Algoritmo 3.2 apresenta o programa principal do modelo ms-NN, as definições das variáveis e as sub-rotinas do programa principal. Neste algoritmo, além do conjunto de dados C, o especialista deve fornecer primeiramente o número de espaços e o tipo de espaço. O número de espaços deve ser maior ou igual a um. O tipo de espaço poderá ser ATR (de atributos) ou GEO (geográfico). Para ambos tipos de espaço, deve-se selecionar uma função de distância, um tipo de vizinhança e atribuir um peso que define o tipo de função de predominância usada.

Pode-se dizer que o k-NN é um caso particular do modelo ms-NN, em que se estabelece apenas um espaço, do tipo ATR, e selecionam-se a distância Euclidiana e a vizinhança fixa e usa-se a função de predominância maioria.

3.3 Formalização do ms-NN

No ms-NN proposto, são usados um ou mais espaços para atribuir a cada caso $c_0 \in C$, de acordo com uma função de predominância, uma classe $\omega_0 \in CL$. Um caso no ms-NN é representado por $c = (g, \mathbf{a}, \omega)$, em que g representa uma geometria associada a uma posição geográfica $g = (x, y)$, $\mathbf{a}$ é um vetor de atributos e ω a classe na geometria g. Os conjuntos de casos de treinamento (T), teste (P) e não rotulados (N) particiona o conjunto C.

Algoritmo 3.2 Algoritmo do modelo ms-NN proposto

Definição das variáveis

espaço	–	um espaço do conjunto de espaços;
k	–	número de casos de treinamento vizinhos considerado para vizinhança fixa;
r	–	raio de influência definido;
ρ_s	–	peso para cada espaço s;
tipoEspaço	–	tipo de espaço selecionado (atributo ou geográfico);
dist. Atr	–	função de distância selecionada (Hamming, Euclidiana ou Mahalanobis);
dist. Geo	–	função de distância selecionada (Hamming, Euclidiana ou Mahalanobis ou por contiguidade);
tipoVizinhança	–	tipo de vizinhança selecionada (variável ou fixa);
novoCaso	–	um caso da base de dados C (c_0);
nCasos	–	número de casos da base de dados C;

Programa Principal

para cada *espaço* **faça**
  se *tipoEspaço=atributos* **então**
    calculeDistânciasAtributos
  senão
    calculeDistânciasGeograficas
  fim se
fim para
para $novoCaso = 1$ **até** $nCasos$ **faça**
  para cada *espaço* **faça**
    se tipoVizinhança=fixa **então**
      encontre os **kVizinhos** e conte os rótulos de cada classe
    senão
      se $r > 0$ **então**
        encontre os **vizRaioInfluencia** e conte os rótulos de cada classe
      senão
        encontre os **vizContiguidade** e conte os rótulos de cada classe
      fim se
    fim se
  fim para
  Classifique o *novoCaso* de acordo com a **funçãoPredominancia** selecionada
fim para

Sub-rotinas

calculeDistânciasAtributos: retorna o conjunto com a *distânciaEspaçoAtributos* dos atributos do *novoCaso* aos atributos de todos os casos de treinamento (c_t).
calculeDistânciasGeográficas: retorna o conjunto com a *distânciaEspaçoGeográfica* do *novoCaso* aos casos de treinamento (c_t).
kVizinhos: retorna o conjunto com os k casos de treinamento mais próximos de c_0;
vizRaioInfluencia: retorna o conjunto com os casos de treinamento que estão a uma distância menor ou igual a r.
vizContiguidade : retorna o conjunto com os casos de treinamento que são contíguos (que tocam) a c_0.
funçãoPredominância: retorna a classificação do *novoCaso* de acordo com o peso ρ_s definido para cada espaço s.

Como visto anteriormente, espaços distintos podem conter diferentes relações de vizinhança. Então, para cada espaço s separam-se os casos vizinhos $c_v \in T$, de acordo com cada vizinhança, produzindo-se o vetor V_0, com n conjuntos de vizinhos (V_{0s}), um para cada espaço s. Em cada conjunto V_{0s} são separados os casos de cada classe cl_i e, de acordo com a função de predominância selecionada, a classe ω_0 é atribuída a c_0.

Portanto, para classificar um caso c_0 usando o ms-NN é necessário definir o número de espaços (n) e o tipo de espaço, além de definir os três elementos básicos para cada espaço, uma função de distância, um tipo de vizinhança e uma função de predominância, detalhados a seguir.

a) Função de distância $d(c_0, c_t)$: No ms-NN para cada espaço (s) tem-se um conjunto de distâncias D_0^s (em que $D_0^s = D_0$ descrito na Equação 3.1, para o espaço s) com as distâncias do caso c_0 a todos os casos c_t, compondo o vetor $\overline{D_0}$.

$$\overline{D_0} = \{D_0^s \colon s = 1, \cdots, n\} \tag{3.6}$$

A escolha da métrica usada para definir cada conjunto D_0^s depende do tipo de espaço e pode ser uma das distâncias a seguir:

- Distância no espaço de atributos $d(c_0, c_t) = d(a_0, a_t)$:
  i. distâncias Euclidiana mostrada na Equação 3.2
  ii. distância de Mahalanobis;

$$d(a_0, a_t) = [(a_0 - a_t)^T Mcor^{-1}(a_0 - a_t)]^{1/2} \tag{3.7}$$

  em que $Mcor$ é a matriz de covariância de todos os casos de treinamento.

  iii. distância de Hamming;
  Seja H uma matriz l x l com $H(a_0, a_t) = [h_{i,j}]$, $ij = 0, 1, \cdots, l-1$, onde h_{ij} é o número de ocorrências de símbolos i e j em a_0 e a_t, respectivamente. A distância de Hamming é definida como:

$$d(a_0, a_t) = \sum_{i=1}^{l} \sum_{j=1, j \neq i}^{l} a_{ij} \tag{3.8}$$

- Distância no espaço geográfico $d(c_0, c_t) = d(z_0, z_t)$:

i. distâncias por contiguidade;

$$d(z_0, z_t) = \begin{cases} 1, & \text{se } z_0 \text{ e } z_t \text{ são contíguos} \\ 0, & \text{se caso contrário.} \end{cases} \tag{3.9}$$

ii. qualquer uma das distâncias descritas no item da distância no espaço de atributos;

iii. outras, por exemplo, distância sobre a malha de estradas.

b) Tipo de vizinhança:

Como dito anteriormente, no ms-NN o tipo de vizinhança é um dos parâmetros definido inicialmente pelo especialista. A vizinhança pode ser fixa ou variável e deve ser definida para cada espaço s. Assim, para todo caso $c_0 \in C$, existe um vetor $\overline{V_0}$ com n conjuntos V_0^s de casos vizinhos do conjunto de treinamento T, ou seja:

$$\overline{V_0} = V_0^s, \ s = 1, \cdots, n \tag{3.10}$$

Os conjuntos V_0^s podem ser compostos por um número fixo de casos (vizinhança fixa) ou por um número variável de casos (vizinhança variável) que atendam a um critério.

– vizinhança fixa:

$$V_0^s = \{V_0\}, \text{ para o espaço } s \tag{3.11}$$

em que são elementos de V_0^s os casos c_t, que possuem as menores distâncias a c_0.

– vizinhança variável:

$$V_0^s = \{c_v : c_v \in T\} \tag{3.12}$$

i. são elementos de V_0^s todos os casos c_t contidos em um raio de influência r definido pelo especialista, ou seja:

$$V_0^s = \{c_t \colon d(c_0, c_t) \leq r, d(c_0, c_t) \in D_0^s\} \tag{3.13}$$

o raio de influência r pode ser definido pelo especialista, tanto para o espaço de atributos, como para o espaço geográfico.

ii. são elementos de V_0^s todos os casos que são vizinhos por contiguidade no espaço geográfico:

$$V_0^s = \{c_t \colon d(c_0, c_t) = 0, d(c_0, c_t) \ \in \ D_0^s\} \tag{3.14}$$

c) Função de predominância:

No ms-NN a classe ω_0 é atribuída a um caso c_0 tendo em vista as classes que compõem a vizinhança em cada espaço. Seja o conjunto V_i^s, um subconjunto de V_0^s com os casos vizinhos de c_0 da classe cl_i em CL, no espaço s:

$$V_i^s = \{c_v \ \in \ V_{0,s} : \omega_v = cl_i\}, \tag{3.15}$$

em que ω_v é a classe do caso c_v

Define-se IP_i como índice de predominância da classe cl_i conforme a equação:

$$IP_i = \Sigma_s \rho_s |V_{i,s}|, \tag{3.16}$$

em que, $|V_{i,s}|$ é o número total de vizinhos da classe cl_i, no espaço s e ρ_s é o peso dos vizinhos no espaço s. O peso ρ_s é definido como:

$$\rho_s = \begin{cases} 1, & \text{se } a \\ \frac{1}{|\#v_s|}, & \text{se } b \\ p, & \text{se } c \end{cases} \tag{3.17}$$

- (a) função de predominância maioria: a partir da vizinhança definida todos os casos vizinhos terão o mesmo peso na classificação, independentemente do número de casos vizinhos e tipo de vizinhança;
- (b) função de predominância ponderada pelo número de vizinhos ($|\#v_s|$) do espaço s;
- (c) p definido pelo usuário.

  Finalmente, a função de predominância que atribui a c_0 a classe mais frequente dentre as m classes é definida como:

$$\omega_0 = \arg\max_i IP_i, \ i = 1, \cdots, m \tag{3.18}$$

Em caso de empate, atribui-se a classe cujo caso é mais próximo de c_0, dando preferência ao espaço de atributos. Assim como no k-NN o conjunto de casos P é usado

para definir os melhores valores para os parâmetros do modelo ms-NN. Os parâmetros são definidos experimentalmente, e serão escolhidos os que apresentarem a menor taxa de erro em P.

4 MATERIAIS E MÉTODOS

Neste capítulo é apresentado uma breve descrição das áreas de estudo e dos dados usados neste trabalho, bem como o método geral de avaliação das classificações. Foram realizados dois estudos de caso. O primeiro é uma classificação do risco de esquistossomose em Minas Gerais. O segundo é uma classificação de uso e cobertura do solo em uma região amazônica do Tapajós. Por convenção, denomina-se como *Tapajós* o estudo de caso com dados da Floresta Nacional de Tapajós e como *Schisto* o estudo de caso com os dados de esquistossomose. A localização espacial dos estudos de caso pode ser vista na Figura 4.1.

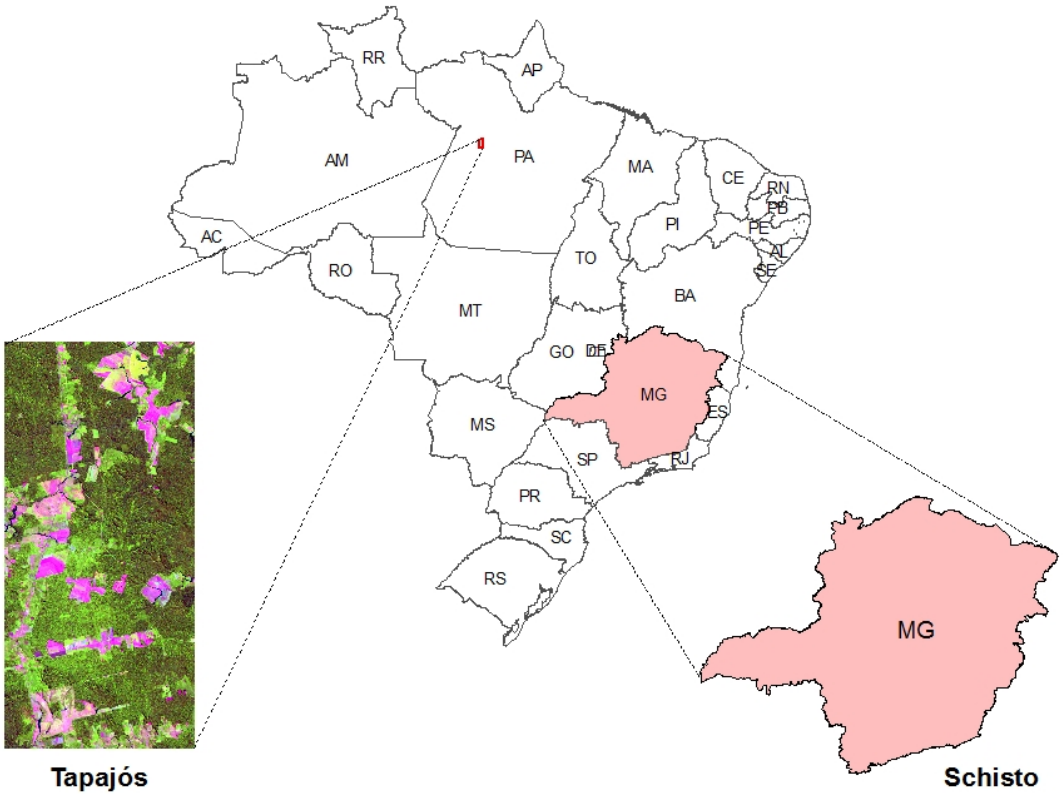


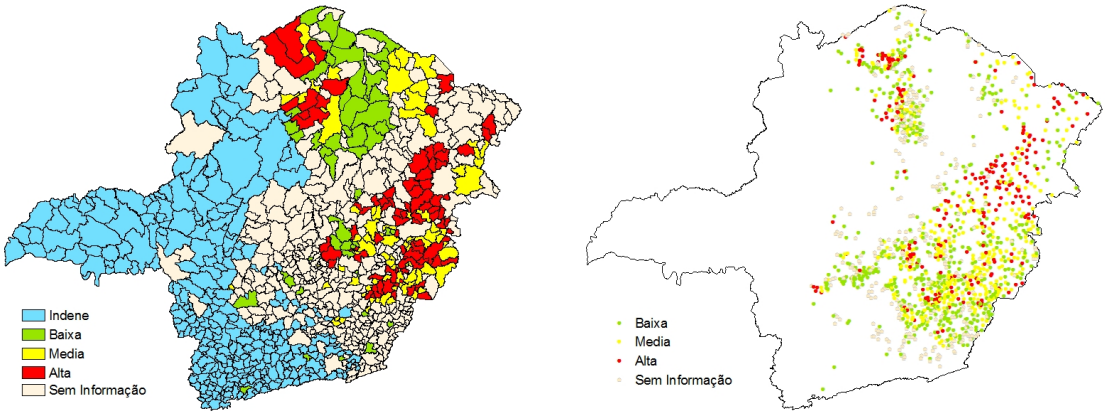
Figura 4.1 - Localização das áreas de estudo.

4.1 Dados do estudo de caso Schisto

O Estado de Minas Gerais localiza-se na região Sudeste do Brasil (Figura 4.1) e é dividido politicamente entre 853 municípios e possui área de aproximadamente 590.000 km^2. A população é de aproximadamente 20 milhões de habitantes e o clima é tropical (IBGE, 2013).

A esquistossomose *mansoni* é um dos graves problemas de saúde publica que afetam milhares de pessoas em todo mundo (WHO, 1985). No Brasil, a esquistossomose é causada pelo agente etiológico *Schistosoma mansoni*, que tem como hospedeiro intermediário caramujos do gênero *Biomphalaria* (AMARAL et al., 2006). O parasita utiliza a água como meio para infectar o homem (hospedeiro definitivo), que através de suas fezes infectadas contamina a água, possibilitando a infecção do caramujo e dando origem a um novo ciclo. Assim, para estudar a transmissão dessa doença, além de combinar fatores ambientais e sociais, relacionados ao caramujo e ao homem, é importante relacionar esses fatores a aspectos espaciais, visto que locais próximos a áreas endêmicas são locais de potencial risco de contaminação.

A distribuição da esquistossomose nos Estado de Minas Gerais é irregular, intercalando-se em áreas de transmissão baixa ou nula com áreas de maior prevalência da doença. Nas Figuras 4.2(a) e 4.2(b), é possível verificar que a doença é endêmica nas regiões norte (compreendendo as zonas do Médio São Francisco e Itacambira), oriental e central (zonas do Alto Jequitinhonha, Metalúrgica, Oeste e Alto São Francisco) e que os maiores índices de infecção são encontrados nas regiões nordeste e leste do Estado, que compreendem as zonas do Mucuri, Rio Doce e da Mata (CARVALHO et al., 2005; CARVALHO et al., 1987).



(a) Prevalência em municípios (b) Índice de positividade em localidades

Figura 4.2 - Distribuição da esquistossomose no estado de Minas Gerais.

Os dados sobre a doença, usados neste trabalho, foram disponibilizados pela Secretaria de Saúde do estado de Minas Gerais em duas escalas: municipal e por localidades. Para diferenciar as duas escalas, foram adotadas as mesmas notações propostas por Guimarães (2010): prevalência da esquistossomose para os dados por municípios

(Pv), apresentada na Figura 4.2(a), e índice de positividade de esquistossomose para os dados de localidades (Ip), Figura 4.2(b). Esses dados foram usados como conjunto de amostras.

Dos 853 municípios apresentados na Figura 4.2(a), 197 possuem informação positiva de prevalência da esquistossomose e 304, apresentados em azul claro, correspondem aos municípios em que a doença é indene (onde não se tem relato de ocorrência). Na Tabela 4.1 pode-se verificar o número de municípios por classe de prevalência. Os pontos indicados na Figura 4.2(b) representam as 1.586 localidades de Minas Gerais. Destas, tem-se informação positiva sobre o índice de positividade da esquistossomose em 1220 localidades. Nas 366 localidades restantes, não existe informação sobre a doença. A quantidade de amostras de cada classe pode ser vista na Tabela 4.2.

Tabela 4.1 - Número de amostras Schisto em nível municipal.

Classe	Municípios
Indene	304
Baixa	46
Média	73
Alta	78

Tabela 4.2 - Número de amostras Schisto em nível local.

Classe	Localidades
Baixa	569
Média	404
Alta	247

As faixas de prevalência e do índice de positividade abaixo de 5%, entre 5% e 15% e acima de 15%, apresentados na Figura 4.2, são definidas como baixa, média e alta, respectivamente, de acordo com classificação da Secretaria de Saúde do Estado de Minas Gerais.

Como conjunto de atributos, são utilizados variáveis extraídas de sensoriamento remoto (SR), obtidas pelos sensores *Moderate Resolution Imaging Spectroradiometer* (MODIS) e *Shuttle Radar Topography Mission*(SRTM), variáveis meteorológicas (chamadas neste trabalhos de climáticas) e variáveis socioeconômicas. Em Martins (2009) pode-se obter uma descrição mais detalhada do conjunto de atributos a nível

municipal. A descrição dos atributos a nível local é retratada em Guimarães (2010).

As variáveis de SR e climáticas em escala municipal foram obtidas em 2003 em um projeto financiado pela FAPEMIG (processo: 1775/03), são compostas pela média por município. Em escala local, as variáveis foram extraídas por Guimarães (2010) e são compostas do pixel correspondente às coordenadas geográficas das localidades. Das variáveis de SR derivadas do MODIS, são usadas: banda azul (Blue), vermelho (Red), infravermelho próximo (NIR), e infravermelho médio (MIR), os índices de vegetação melhorado (EVI), o índice de vegetação da diferença normalizada (NDVI), e as imagens-fração derivadas do MLME, vegetação (Veg), solo (Solo) e sombra (Sombra). Das variáveis obtidas através do SRTM, são usados o modelo digital de elevação (DEM) e a declividade (Dec), derivada do DEM. Das variáveis climáticas (Cli), são usadas a precipitação acumulada (Prec), a temperatura mínima (Tmin) e a temperatura máxima (Tmax).

A nível municipal, as variáveis socioeconômicas foram divididas em dois tipos: IDHs e situação por domicílio (Sit). Essas variáveis foram obtidas do Sistema Nacional de Indicadores Urbanos (SNIU) e foram também usadas no projeto da FAPEMIG em 2003. Além do IDH são usados o IDH de educação (IDHE), de longevidade (IDHL) e de renda (IDHR) dos anos de 1991 e 2000. Os dados referentes à situação por domicílio são do ano 2000 e representam:

- Renda do responsável pelo domicílio;
- Anos de estudo do responsável pelo domicílio;
- Tipo de saneamento (esgoto e água).

Para a escala local foram usadas como variáveis socioeconômica, os dados que descrevem a situação do domicílio (Sit). Essas dados possuem informações sobre:

- Os domicílios (tipo de domicílio, quantidade de moradores (total, por sexo, por faixa etária, por nível de escolaridade, com receita), tipo de abastecimento de água, quantidade de banheiros, tipo de saneamento e coleta de lixo);
- A quantidade de pessoas que vivem no setor (total, casal, filhos, por sexo, por tipo de domicílio, por faixa etária, por nível de escolaridade, com receita, por tipo de abastecimento de água, por quantidade de banheiros, por tipo de saneamento e por tipo coleta de lixo);

- O responsável pelo domicílio (por faixa etária, por escolaridade, por renda, por sexo e por quantidade de dependente).

Para o espaço geográfico (GEO), foi usada a distância por contiguidade (Equação 3.9) em escala municipal e a distância Euclidiana (Equação 3.2) na escala local.

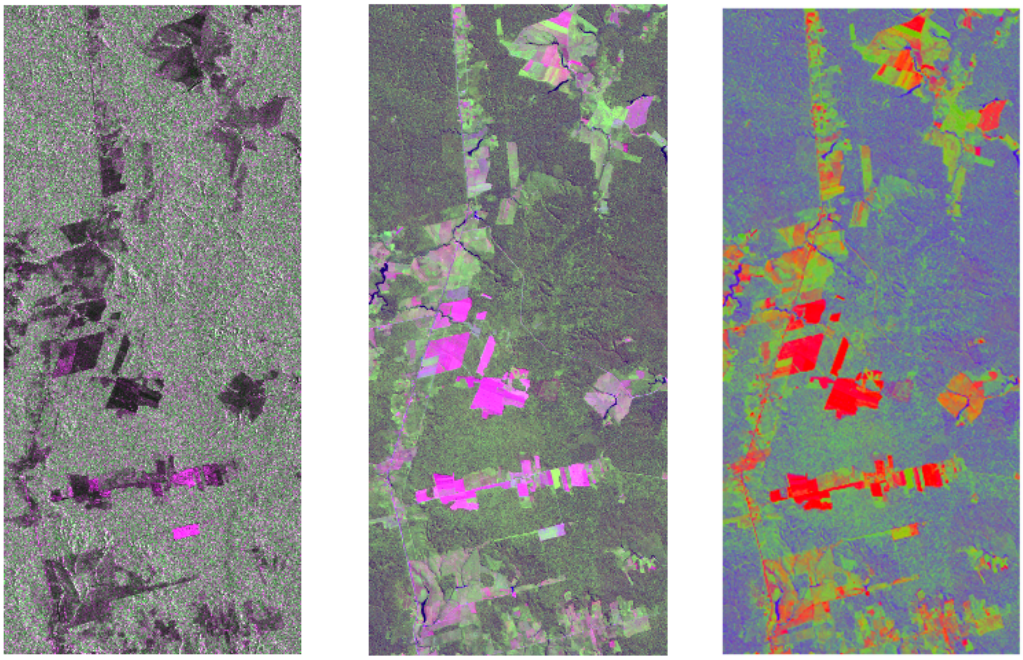
4.2 Dados do estudo de caso Tapajós

A área de estudo tem aproximadamente 411 km^2 e compreende parte da Floresta Nacional do Tapajós e arredores, abrangendo parte do município de Belterra, no Estado do Pará. A vegetação da área é do tipo Floresta Ombrófila Densa, com presença de lianas lenhosas, palmeiras e epífitas. Na matriz de floresta primária, existem mosaicos de vegetação secundária, pastos, áreas desmatadas e agrícolas, de cultura familiar e mecanizada (ESCADA et al., 2009).

Neste estudo, foram utilizados os dados disponibilizados por Reis (2013), que consistem em duas imagens ortorretificadas da região, três imagens fração, uma imagem segmentada e amostras de áreas cujas classes de cobertura da terra foram previamente identificadas. As imagens disponibilizadas foram uma imagem ótica e uma imagem de micro-ondas. Essas imagens e a imagem fração estão apresentadas na Figura4.3.

A imagem ótica é oriunda do sensor *Thematic Mapper* (TM) do satélite LANDSAT5, imageada no dia 29 de junho de 2010. A composição das bandas 5 (R), 4 (G) e 3 (B) da imagem ótica usadas neste estudo é mostrada na Figura 4.3(b). A imagem de micro-ondas foi obtida no modo *Fine Beam Dual* (FBD) pelo sensor *Phase Array L-Band Synthetic Aperture Radar* (PALSAR), do satélite *Advanced Land Observing System* (ALOS), no dia 21 de junho de 2010. Na Figura 4.3(a), a imagem micro-ondas é mostrada na composição colorida HH (R), HV (G) e HH (B), em amplitude. As imagens foram previamente ortorretificadas e projetadas para UTM WGS84, zona 21 sul, com espaçamento entre pixels de 15m para a imagem ALOS/PALSAR e 30m para a imagem LANDSAT5/TM.

As imagens fração correspondem às frações de vegetação, solo e sombra (Figura 4.3(c)), calculadas usando as bandas 1 a 5 e 7 da referida imagem LANDSAT5/TM, utilizando o modelo de mínimos quadrados ponderados (MQP) (PONZONI; SHIMABUKURO, 2009) implementado no software SPRING 4.3.3 (CÂMARA et al., 1996). A imagem segmentada foi gerada usando o *multiresolution segmentation* do e-Cognition. Esta segmentação foi criada a partir das bandas 3, 4 e 5 da imagem



(a) Imagem ALOS/ PALSAR – composição HH, HV, HH (RGB)

(b) Imagens LANDSAT5/TM 5(R), 4 (G), 3 (B)

(c) Imagens-fração composição solo (R), vegetação (G), sombra (B)

Figura 4.3 - Dados matriciais Tapajós

LANDSAT5/TM, normalizadas para média 127 e desvio padrão 42 e convertidas para byte. Os parâmetros foram: escala = 15, forma = $0,3$ e compacidade = $0,5$. A imagem segmentada resultante é uma supersegmentação.

As amostras de cobertura disponibilizadas foram coletadas na imagem supersegmentada com base no trabalho de campo descrito em Pereira (2012), realizado de 9 a 17 de setembro de 2010, na área em questão. Das amostras de cobertura disponibilizadas, optou-se por utilizar seis classes de cobertura:

- área cultivada (AC);
- solo exposto (SE);
- pasto (PA);
- regeneração inicial e intermediária (RI);
- floresta primária (FP);
- regeneração avançada ou floresta degradada (FA).

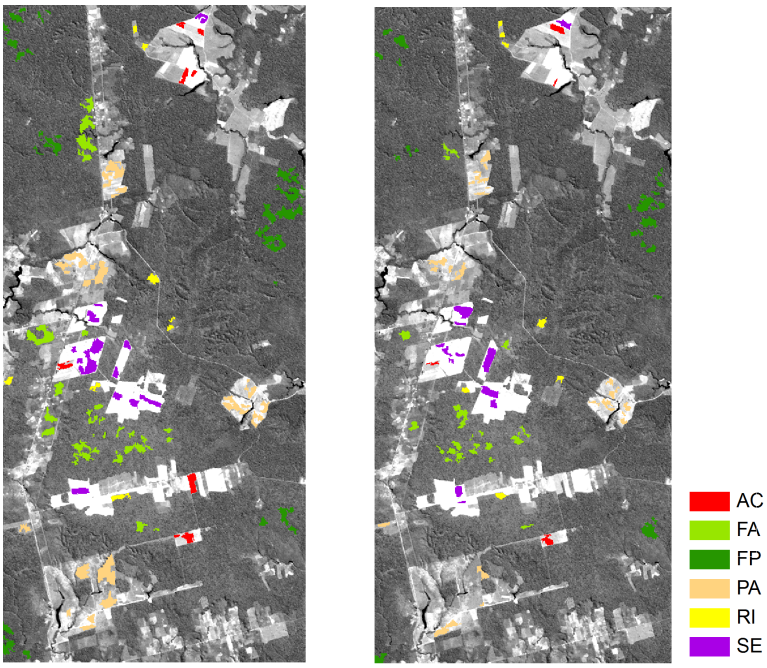
A classe AC corresponde a áreas agrícolas que possuem culturas de grãos. SE são áreas em que a cobertura predominante seja solo nú. A classe PA corresponde a áreas com vegetação típica de pastagens, ou seja, predominantemente gramíneas, com ou sem a presença de pequenos arbustos e espécies invasoras. A classe RI abrange áreas com vegetação secundária em desenvolvimento inicial ou intermediário, compostas principalmente por árvores de pequeno porte, arbustos e gramíneas. FP é a floresta primária em que a ação humana não provocou alterações significativas em suas características originais de estrutura e espécies. A classe FA é composta pela junção de outras duas classes, floresta degradada e regeneração avançada. Entende-se por floresta degradada uma floresta primária que perdeu suas características originais de estrutura e espécies, principalmente pela ação de fogo e extrativismo vegetal. Entende-se por regeneração avançada uma área previamente desmatada, cuja vegetação secundária encontra-se desenvolvida, com predomínio de árvores entre 13 e 17 metros. A quantidade de amostras para cada classe pode ser vista na Tabela 4.3 e a distribuição das amostras na área de estudo na Figura 4.4.

Tabela 4.3 - Tamanho das amostras de Tapajós

Classe	Descrição	Polígonos
AC	Área Cultivada	100
SE	Solo Exposto	47
PA	Pasto sujo e pasto limpo	130
RI	Regeneração Inicial e Intermediária	23
FP	Floresta Primária	104
FA	Floresta Alterada (floresta degradada ou regeneração antiga)	24

A partir das bandas 3, 4 e 5 da imagem LANDSAT5/TM, com base na segmentação disponibilizada, foram calculados atributos de textura (entropia, homogeneidade e dissimilaridade), usando o GEODMA (KORTING, 2012).

Como espaço geográfico (GEO) foi usada a distância por contiguidade (Equação 3.9). O espaço geográfico pôde ser considerado neste trabalho, porque foi usada uma supersegmentação. Em uma supergmentação, os pixels de uma mesma região (polígono) são muito similares e regiões vizinhas podem possuir pixels com valores parecidos. Nestes casos, usar a informação espacial pode ser útil. Cabe ressaltar que em uma segmentação ótima não faz sentido o uso da informação espacial, pois nesses casos os pixels dentro de uma região possuem valores homogêneos e entre regiões possuem valores bem diferentes.



(a) Amostras de Treinamento

(b) Amostras de Teste

Figura 4.4 - Amostras Tapajós

4.3 Metodologia dos experimentos

Neste trabalho, foram geradas classificações usando o SVM, árvore de decisão, k-NN e ms-NN. As classificações foram geradas variando os parâmetros dos classificadores. Para cada classificador, foi selecionada uma *classificação representante*, em um ambiente de simulação. As classificações representantes de cada método foram comparadas de forma pareada, usando o *índice de desempenho*. O índice de desempenho é um termo proposto nesta tese para identificar, em um ambiente de simulação, quantas vezes a acurácia de um classificador é superior a de outro.

As classificações feitas usando SVM foram geradas usando o *SVM multiclass* (CRAMMER; SINGER, 2001) a partir de todos os atributos disponíveis. Como parâmetros do SVM, foi usado o *kernel* polinomial, com grau variando entre 1 a 3. Foram considerados ainda diferentes valores para o parâmetro penalidade (Pen), assumindo valores iguais a $0,01$, $0,1$, 1, 10, 100 e 1000, resultando em 18 classificações. Dessas, foi selecionada uma classificação representante para ser comparada com as classificações representantes de cada método.

As classificações por árvore de decisão foram feitas usando o software WEKA (HALL

et al., 2009). Foram geradas 19 classificações variando o parâmetro número mínimo de amostras por folha. Esse parâmetro foi testado de 2 a 20 amostras. Essas classificações foram geradas a partir de todos os atributos.

Como descrito na seção 2.3, o k-NN possui apenas um parâmetro: o número de vizinhos. Foram testados de 1 a x vizinhos. Esse limite x foi calculado com base em Bishop (2006) e Webb (2002) como uma aproximação da raiz quadrada do número de amostras de treinamento. Então para o k-NN tem-se $x - 1$ classificações para cada conjunto de atributos.

As classificações pelo k-NN e ms-NN foram geradas usando uma implementação feita como plug-in para TerraView (ver Apêndice A). No ms-NN (ver Seção 3.2) foram geradas classificações com 2 a n espaços. O número máximo de espaço n foi definido de acordo com número de conjunto de atributos. Quando usa-se somente espaço do tipo ATR o número de espaços é igual ao número de conjuntos de atributos. Quando se usam espaço ATR e GEO, o número de espaços é igual ao número de conjuntos de atributos mais 1. Para cada espaço ATR foram usados de 1 a x vizinhos (como no k-NN), e foram testados dois tipos de distância (Euclidiana e Mahalanobis).

Em todos os espaços ATR foi adotada a vizinhança fixa. Também foi usada vizinhança fixa para o espaço GEO, no caso de uso de pontos como representação geométrica (no estudo de caso Schisto por localidade). Neste caso, a distância usada entre os pontos geométricos foi a Euclidiana. Quando foi usado polígono, como representação geométrica, foi adotada vizinhança variável e distância por contiguidade (estudo de caso Schisto por município). Nas classificações geradas usando dois tipos de espaço, com vizinhança fixa para os espaços ATR e variável para o espaço GEO, foi usada a função de predominância ponderada pelo número de vizinhos. Para as outras classificações geradas usando vizinhança fixa para todos os espaços, independentemente do tipo, foi usado o voto da maioria como função de predominância.

Para cada classificador, foi selecionada uma classificação representante pela média das acurácias. A seleção das classificações representantes foram feitas em três etapas baseadas em um estudo Monte Carlo. Primeiramente, as amostras do conjunto de teste foram estratificadas e foram selecionados aleatoriamente 1.000 conjuntos com reposição. Esses conjuntos foram usados em todas as comparações. Em seguida, para cada configuração do classificador, foram contabilizadas as acurácias para cada um dos 1.000 conjuntos e foi calculada a média dessas acurácias. A configuração do classificador que apresentou a maior média foi selecionada para ser comparada com as outras configurações do classificador. As 1.000 acurácias da configuração selecionada

foram comparadas, de forma pareada, com as acurácias de cada uma das outras configurações do classificador, visando verificar a diferença estatística pelo teste T (GUIMARÃES, 2008). Em casos em que as acurácias foram estatisticamente iguais, foi selecionada como classificação representante, a classificação que apresentou a configuração mais simples (ie. no caso do k-NN, o menor número de casos vizinhos; no ms-NN, o conjunto com o menor valor da soma dos número de casos vizinhos em cada espaço e; para a árvore de decisão, o menor número de objetos por folha).

Para se ter uma ideia geral dos resultados de acurácia, foram gerados gráficos do tipo boxplot com os melhores resultados das classificações de cada classificador. É importante ressaltar que no teste pareado existe uma ordem das acurácias das classificações e essa ordem se perde neste tipo de gráfico. É possível que ao fazer uma comparação entre dois conjuntos pareados a partir de boxplot, tenha-se a impressão que eles sejam iguais. No entanto, o resultado de um classificador é sempre um pouco melhor que outro.

O índice de desempenho foi usado no final da análise para comparar quantas vezes um classificador é superior aos outros. Foram gerados gráficos para ilustrar o desempenho de cada classificador. Nesses gráficos o classificador que aparece sem dados é o que está sendo comparado com os outros classificadores. Pode-se considerar que um classificador é melhor que o outro quando as acurácias de um são maires que as acurácias de outro em mais de 50% das vezes. No entanto, neste trabalho um classificador é considerado estatisticamente superior à outro se as acurácias de um for 90% das vezes maior que as outras do outro.

5 EXPERIMENTOS DO ESTUDO DE CASO SCHISTO

Este capítulo está dividido em 4 Seções. Na primeira, é apresentada a metodologia dos experimentos. Na segunda e terceira, os resultados para o estudo de caso Schisto em nível municipal e local, respectivamente. Na última Seção são apresentadas algumas conclusões.

5.1 Metodologia dos experimentos

Neste estudo, cada caso é descrito por um vetor que contém sua classe, posição geográfica e valores para atributos selecionados. Para o estudo em escala municipal, os casos se referem aos municípios do Estado de Minas Gerais e têm polígonos como representação geométrica. No estudo em escala local, os casos se referem a localidades do Estado de Minas Gerais e são representados geometricamente por pontos.

Foram realizados dois tipos de experimentos em cada uma das duas escalas, um usando o espaço ATR e GEO e outro usando apenas espaço ATR. Como na escala municipal têm-se polígonos, quando o espaço geográfico foi usado, foram considerados no cômputo todos os casos de treinamento vizinhos por contiguidade como definidos na Equação 3.9. Para a escala local foram considerados, no espaço geográfico, os casos vizinhos mais próximos da localidade pela distância Euclidiana.

Os atributos dos dados em nível municipal e local descritos na Seção 4.1 foram inicialmente selecionados usando a correlação entre as variáveis observadas e os dados sobre a doença (Pv e Ip). Dos 62 atributos disponibilizados a nível municipal, 29 foram selecionados por possuírem correlação com a prevalência da esquistossomose superiores a 30%. Da mesma forma, foram selecionados 16 atributos dos 94 disponibilizados em nível local. Os atributos selecionados juntamente com as descrições foram divididos em conjuntos semanticamente distintos. Os atributos separados por conjuntos são apresentados nas Tabela 5.1 e 5.2 para o nível municipal e local, respectivamente.

Em nível municipal, os métodos k-NN e ms-NN foram testados usando combinações de 2, 3 e 4 conjuntos de atributos, conforme Tabela 5.3. Em nível local, esses métodos foram testados usando combinações de 2 e 3 conjuntos de atributos conforme Tabela 5.4. Vale ressaltar que as classificações usando o SVM e árvore de decisão foram geradas usando todos os conjuntos de atributos juntos.

O primeiro bloco da Tabela 5.3 contém as combinações de 2 conjuntos de atributos. O segundo bloco contém as combinações com 3 conjuntos de atributos. E o ter-

Tabela 5.1 - Atributos selecionados pela correlação com Pv – nível municipal.

Sigla	Descrição
	Atributos que representam a situação por domicílio (Sit)
SB	Média do número de domicílios Sem Banheiro
RenInf1sm	Porcentagem de chefe de família com renda inferior a um salário mínimo
Ren5e10sm	Porcentagem de chefe de família com renda entre cinco e dez salários mínimos
Ren10e15sm	Porcentagem de chefe de família com renda entre dez e quinze salários mínimos
RenSup15sm	Porcentagem de chefe de família com renda superior a quinze salários mínimos
EstInf1	Porcentagem de responsáveis por domicílios com menos de um ano de estudo
Est4e7	Porcentagem de responsáveis por domicílios com quatro a sete anos de estudo
Est8e10	Porcentagem de responsáveis por domicílios com oito a dez anos de estudo
Est11e15	Porcentagem de responsáveis por domicílios com onze a quinze anos de estudo
EstSup15	Porcentagem de responsáveis por domicílios com mais de quinze anos de estudo
ZRural	Porcentagem de domicílios da zona rural
	Atributos dos IDHs (Idh)
IDH_{91}	Índice de Desenvolvimento Humano do ano de 1991
IDH_{00}	Índice de Desenvolvimento Humano do ano de 2000
$IDHR_{91}$	Índice de Desenvolvimento Humano de Renda do ano de 1991
$IDHR_{00}$	Índice de Desenvolvimento Humano de Renda do ano de 2000
$IDHL_{91}$	Índice de Desenvolvimento Humano de Longevidade do ano de 1991
$IDHL_{00}$	Índice de Desenvolvimento Humano de Longevidade do ano de 2000
$IDHE_{91}$	Índice de Desenvolvimento Humano de Educação do ano de 1991
$IDHE_{00}$	Índice de Desenvolvimento Humano de Educação do ano de 2000
	Atributos derivados de sensoriamento remoto (Sr)
DEM	Média das elevações do sensor SRTM
Dec	Média da declividade derivado do sensor SRTM
EVI_I	Média do Índice de Vegetação Melhorado no Inverno
$NDVI_I$	Média do Índice de Vegetação da Diferença Normalizada no Inverno
NIR_I	Média da banda do infravermelho próximo do sensor MODIS no Inverno
VEG_I	Média da imagem-fração vegetação no Inverno
	Atributos representam o clima (Cli)
$PREC_I$	Média da Precipitação acumulada no Inverno
$Tmax_V$	Média da Temperatura máxima no Verão
$Tmin_I$	Média da Temperatura mínima no Inverno
$Tmin_V$	Média da Temperatura mínima no Verão

ceiro bloco contém as combinações com 4 conjuntos de atributos. Nessa Tabela, a vírgula é usada para distinguir os conjuntos de atributos usados em diferentes espaços. As classificações do k-NN (1 espaço) estão apresentadas na primeira coluna da Tabela 5.3. Na segunda, terceira e quarta colunas, estão os conjuntos de atributos usados para compor 2, 3 e 4 espaços ATR.

Em nível local, os métodos k-NN e ms-NN foram testados usando combinações de 2 e 3 conjuntos de atributos. As combinações com 2 conjuntos de atributos são apresentadas na primeira coluna da Tabela 5.4. Na segunda coluna da Tabela 5.4 são apresentadas as combinações dos demais conjuntos de atributos.

Tabela 5.2 - Atributos selecionados pela correlação com Ip – nível local.

Sigla	Descrição
	Atributos que representam a situação por domicílio (Sit)
EstRenpDom	Média do número de anos de estudo das pessoas responsáveis por domicílios particulares permanentes
NumMorDom	Média do número de moradores em domicílios particulares permanentes
DomSemAgua	Domicílios particulares permanentes com abastecimento de água de poço ou nascente na propriedade, não canalizada
DomEsgVala	Domicílios particulares permanentes com banheiro ou sanitário e esgotamento sanitário via vala
DomSemBanh	Domicílios particulares permanentes sem banheiro
NaoAlfab	Número de pessoas não alfabetizadas com 20 a 24 anos de idade
	Atributos derivados de sensoriamento remoto (Sr)
DEM	Média das elevações do sensor SRTM
$BLUE_I$	Banda Azul no Inverno
RED_I	Banda vermelho no Inverno
EVI_I	Índice de Vegetação Melhorado no Inverno
$NDVI_I$	Índice de Vegetação da Diferença Normalizada no Inverno
$Solo_I$	Imagem-fração Solo no Inverno
VEG_I	Imagem-fração Vegetação no Inverno
	Atributos representam o clima (Cli)
$Tmax_V$	Temperatura máxima no Verão
$Tmin_I$	Temperatura mínima no Inverno
$Tmin_V$	Temperatura mínima no Verão

Os conjuntos de atributos separados por vírgula, nas Tabelas 5.3 e 5.4, representam as combinações dos atributos usadas para gerar as classificações em cada espaço.

Um dos objetivos desse estudo é gerar mapas temáticos que apontem os municípios e localidades que, de acordo com o estudo, possuam fatores de risco favoráveis para a doença. Esses mapas podem, a critério da Secretaria de Saúde do Estado de Minas Gerais, ser usados como subsídio para o mapeamento e controle da esquistossomose nos municípios e localidades do Estado. Para esta Secretaria, é interessante o mapeamento de áreas com informação positiva. Para o estudo de caso em escala municipal dos 501 municípios com informação sobre a doença, tem-se apenas 40% (197 municípios) com informação positiva sobre a doença. Para o estudo de caso em escala local, as informação sobre a doença é positiva em todas as 1.216 localidades, dado pelo atributo Ip.

Neste contexto, têm-se três abordagens para os experimentos a nível municipal. Numa primeira abordagem são usadas apenas 197 amostras, referentes a 40% dos municípios que possuem informação sobre a doença. Na segunda, além dos municípios com informação positiva sobre a doença são levados em conta os municípios considerados indenes, totalizando 501 amostras (Figura 4.2(a)). Como na segunda

Tabela 5.3 - Combinações de atributos por espaço. Schisto em nível municipal.

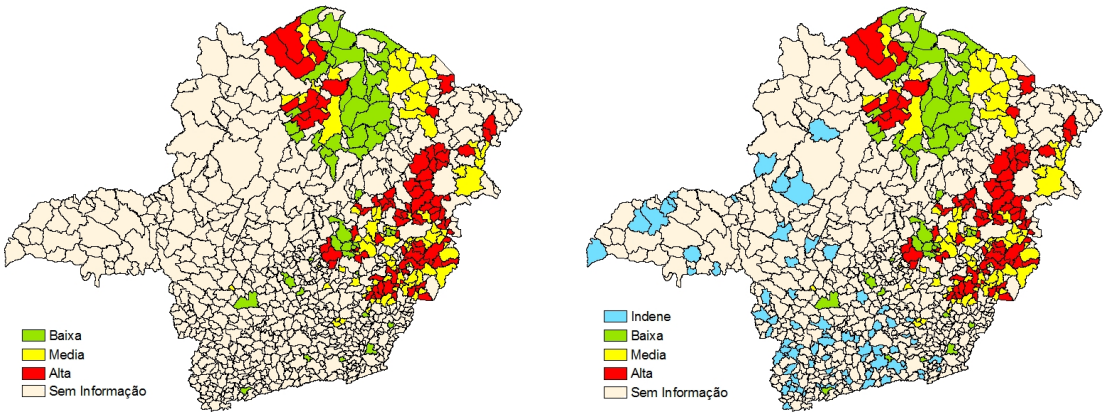
1 Espaço (k-NN)	2 Espaço	3 Espaço	4 Espaço
SitIdh	Sit,Idh		
SitSr	Sit,Sr		
SitCli	Sit,Cli		
IdhSr	Idh,Sr		
IdhCli	Idh,Cli		
SrCli	Sr,Cli		
SitIdhSr	Sit,IdhSr	Sit,Idh,Sr	
SitIdhCli	Sit,IdhCli	Sit,Idh,Cli	
SitSrCli	Sit,SrCli	Sit,Sr,Cli	
IdhSrCli	Idh,SitCli	Idh,Sr,Cli	
	Idh,SitSr		
	Idh,SrCli		
	Sr,SitCli		
	Sr,IdhCli		
	Cli,SitIdh		
	Cli,SitSr		
SitIdhSrCli	Sit,IdhSrCli	Sit,Idh,SrCli	Sit,Idh,Sr,Cli
	Idh,SitSrCli	Sit,Sr,IdhCli	
	Cli,SitIdhSr	Sit,Cli,IdhSr	
	Sr,SitIdhCli	Idh,Sr,SitCli	
		Idh,Cli,SitSr	
		Sr,Cli,SitIdh	

Tabela 5.4 - Combinações de atributos por espaço. Schisto em nível local.

1 Espaço (k-NN)	2 Espaço	3 Espaço
SitSr	Sit,Sr	
SitCli	Sit,Cli	
SrCli	Sr,Cli	
SitSrCli	Sit,SrCli	Sit,Sr,Cli
	Sr,SitCli	
	Cli,SitSr	

abordagem tem-se um número muito grande de amostras na classe indene (ver número de amostras por classe na Tabela 4.1), optou-se por usar apenas 78 amostras (o que equivale a 25% dos casos indenes). Na terceira abordagem, igualou-se o número de amostras da classe indene ao número de amostras da classe mais amostrada na primeira abordagem. Finalmente, na terceira abordagem são usadas no total 275 amostras, 25% das amostras da classe indene juntamente com as 197 amostras

com informação positiva da doença. A distribuição da prevalência da doença para a primeira e terceira abordagens é apresentada na Figura 5.1.



(a) Prevalência positiva em municípios para a primeira abordagem.

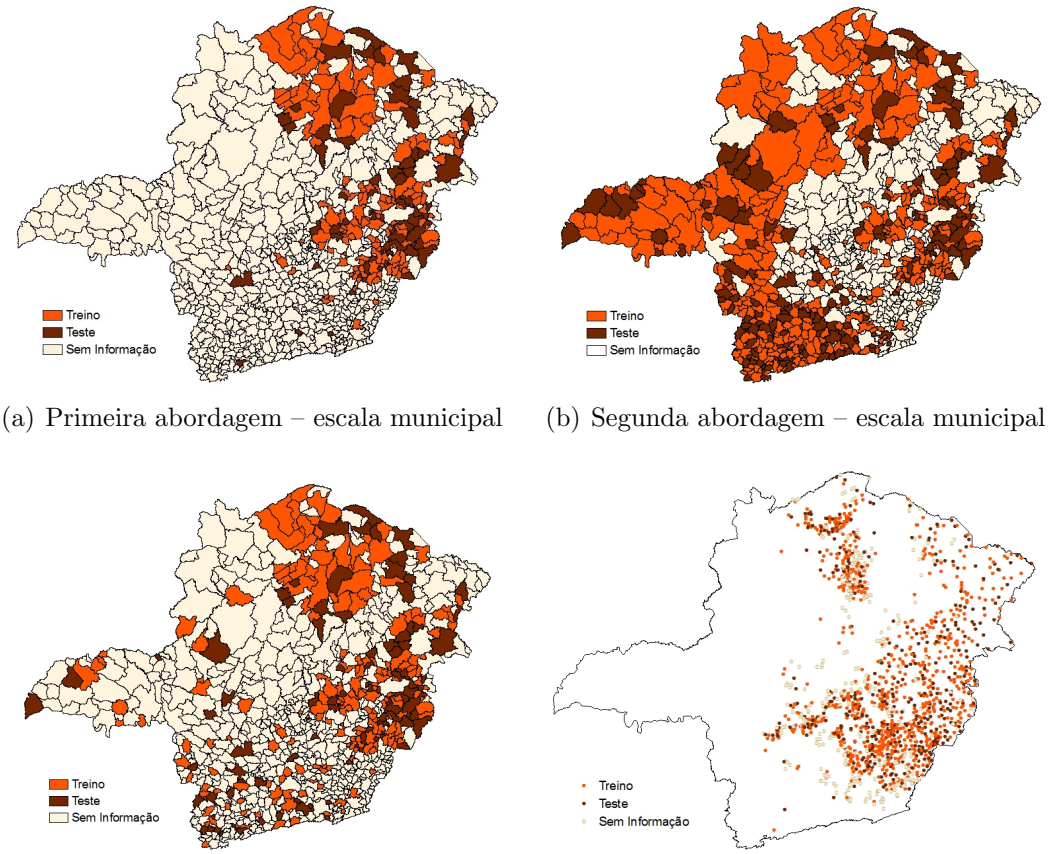
(b) Prevalência em municípios para a terceira abordagem.

Figura 5.1 - Distribuição da esquistossome em municípios do estado de Minas Gerais. Fonte: Secretaria de Saúde do Estado de Minas Gerais.

O número de casos é o que diferencia as três abordagens em nível municipal. Em nível local, tem-se apenas uma abordagem, na qual são usados os 1.220 casos que possuem Ip. Em todas as abordagens, em ambas escalas, as amostras foram estratificadas pela classe e foram separados aleatoriamente 2/3 delas como conjunto de treinamento e os 1/3 restante como conjunto de teste, como pode ser visto na Figura 5.2.

O conjunto de treinamento (T) de cada abordagem foi usado para classificar o conjunto de teste (P) e o conjunto de casos não rotulados (N). Foram realizadas várias classificações, variando os parâmetros para o método ms-NN proposto e para os métodos de comparação: k-NN, SVM e árvore de decisão. Para o k-NN foram testados de 1 a $\sqrt{tr}$ vizinhos, em que tr é o número de casos de treinamento (BISHOP, 2006); (WEBB, 2002). Para a árvore de decisão, foram realizados testes com 2 a 20 amostras mínimas por folha, gerando 19 classificações. Já para o SVM, foram usados grau polinomial de 1 a 3 e penalidades de 10^{-2} a 10^{3}), resultando também em 18 classificações.

Nos experimentos do modelo ms-NN em escala municipal, foram testados de 1 a $\sqrt{tr}$ vizinhos, usando dois tipos de distâncias (Euclidiana e Mahalanobis), em até 5 espaços. Nos experimentos com o espaço GEO, foram usados dois tipos de fun-



(a) Primeira abordagem – escala municipal (b) Segunda abordagem – escala municipal

(c) Terceira abordagem – escala municipal (d) Abordagem em escala local

Figura 5.2 - Distribuição das amostras em treinamento e teste.

ção de predominância, maioria e ponderada. Nos experimentos realizados somente com espaço ATR (até 4 espaços), foi usada a função de predominância maioria. Na Tabela 5.5 é possível verificar o número de classificações resultantes para cada abordagem.

Em escala local, os experimentos foram realizados usando 1 a 28 vizinhos em até 4 espaços. Para os experimentos com e sem espaço geográfico, foram testados dois tipos de distância (Euclidiana e Mahalanobis) e foi usada a função de predominância maioria.

Para cada método, ms-NN, k-NN, árvore de decisão e SVM, foi selecionada a parametrização que forneceu os melhores resultados, de acordo com a Seção 4.3. Estas classificações foram então comparadas pelo índice de desempenho.

Também foram gerados boxplot para todas as classificações, sendo que para o ms-NN foi considerado tanto o uso de somente o espaço ATR (ms-NN) quanto com inclusão

Tabela 5.5 - Número de classificações resultantes do modelo ms-NN para Schisto.

Abordagem	# casos	# vizinhos	# classificações sem GEO		
			2 Esp	3 Esp	4 Esp
3 classes	197	$1 \sim 11$	5.324	26.620	29.282*
4 classes	501	$1 \sim 18$	14.256	116.640	20.000*
4 classes (25% Ind)	275	$1 \sim 13$	7.436	43.940	20.000*
Localidades	1.220	$1 \sim 28$	9.408	2.000*	–

Abordagem	# casos	# vizinhos	# classificações com GEO			
			2 Esp	3Esp	4 Esp	5 Esp
3 classes	197	$1 \sim 11$	330	10.648	53.240	58.564*
4 classes	501	$1 \sim 18$	540	28.512	233.280	40.000*
4 classes (25% Ind)	275	$1 \sim 13$	390	14.872	87.880	40.000*
Localidades	1.220	$1 \sim 28$	6.272	263.424	20.000*	–

* foram usados de 1 a 10 vizinhos apenas

do espaço GEO (ms-NN+). Os melhores resultados estão destacados em cada gráfico. Para melhor visualização dos resultados, os boxplot do ms-NN, apresentados nesta seção, estão separados em blocos que representam o número de espaços. Cada classificação representante desses box plots está descrita pela notação A\B\C\D, onde:

- A se refere ao número de espaços;
- B se refere à configuração (conjunto de atributos e números de vizinhos, separados por vírgula), com a inclusão do símbolo "_+" quando se tem espaço geográfico;
- C se refere à função utilizada para se calcular os vizinhos, com "d_E" e "d_M" denotando as distâncias Euclidiana e Mahalanobis; e
- D se refere à função de predominância, usando "p_m" para maioria e "p_w" para ponderada pelo número de vizinhos no espaço.

As componentes acima podem ser omitidas em uma configuração desde que não prejudique o entendimento do texto.

Como resultado final, foram gerados 3 mapas com a melhor classificação para cada abordagem a nível municipal e 1 mapa, com a classificação a nível local. Esses mapas foram gerados usando a classificação que deu origem aos conjuntos para o estudo Monte Carlo.

5.2 Resultados em escala municipal

Nesta Seção são apresentados os resultados das classificações para cada abordagem a nível municipal. A abordagem com 3 classes é mostrada na Seção 5.2.1. Na Seção 5.2.2 são mostrados os resultados da abordagem com 4 classes, usando apenas 25% das amostras indenes. Na Seção 5.2.3 são mostrados os resultados da abordagem com 4 classes usando todos os casos.

A partir das classificações representantes, foram gerados gráficos do tipo boxplot para cada abordagem de experimentos. As classificações, de cada conjunto de configuração, a partir das quais foram selecionados as representantes está detalhado no Apêndice B, onde se encontram as classificações pelo SVM, árvore de decisão e as classificações pre-selecionadas por número de casos vizinhos do k-NN e ms-NN.

Para cada abordagem, numa primeira análise, foram usados todos os conjuntos de atributos para gerar o ms-NN com e sem espaço GEO. Nessa análise, os resultados foram comparados com os métodos SVM, árvore de decisão e k-NN. Numa segunda análise, foram usados 2 e 3 conjuntos de atributos para gerar o ms-NN com e sem espaço GEO. Nessa análise as classificações do ms-NN foram comparadas somente com o k-NN.

5.2.1 Abordagem com 3 classes

Nos experimentos apresentados nesta Seção foram usados somente os casos com informação positiva sobre a esquistossomose a nível municipal, descrito através de Pv que pode assumir os valores Baixa, Média e Alta.

As classificações representantes geradas do ms-NN usando somente espaço de atributos são apresentadas na Figura 5.3. A Figura 5.4 apresenta o boxplot com classificações representantes para o ms-NN com espaço GEO.

Para a abordagem com 3 classes, fica bem clara a diferença do uso da distância Euclidiana e Mahalanobis quando somente espaço ATR é utilizado (Figura 5.3). Neste caso, em média, a distância de Mahalanobis é mais acurada. Já quando se usa o espaço GEO (Figura 5.4), a distância de Mahalanobis é ou mais acurada ou possui acurácia estatisticamente igual à Euclidiana quando se usam 3, 4 e 5 espaços.

É interessante observar na Figura 5.4, que usando dois espaços com d_M e variando-se a função de predominância, os resultados são idênticos. O mesmo fato se repete com 3 e 4 espaços usando os dois tipos de distância e função de predominância

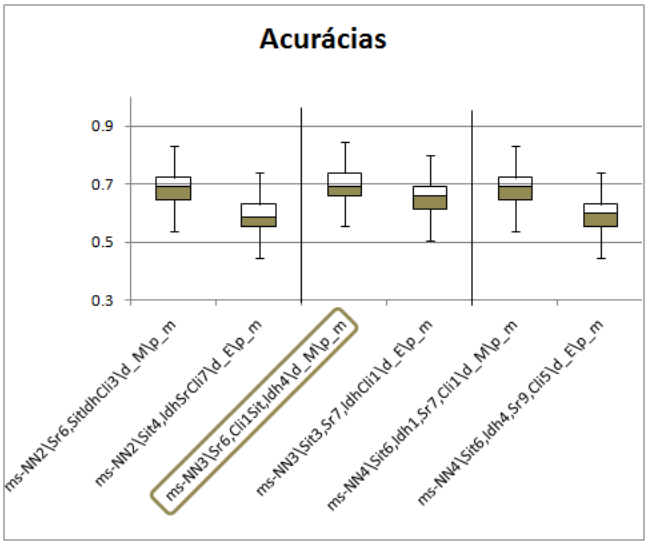


Figura 5.3 - ms-NN com até 4 espaços ATR, usando todos os conjuntos de atributos com a abordagem com 3 classes para Schisto por município.

ponderada.

No boxplot da Figuras 5.3 e 5.4, as acurácias das classificações ficaram entre $0,4$ e $0,84$. De cada um dos boxplots apresentados, foi selecionada uma classificação. Essas classificações são mostradas novamente na Figura 5.5(a), onde além das duas classificações representantes do ms-NN, têm-se as classificações representantes do SVM, árvore de decisão e k-NN.

A partir da Figura 5.5(a), é possível perceber o ganho do ms-NN em relação aos classificadores da literatura. Utilizando-se ou não o espaço GEO, o ms-NN obteve melhor resultado. Enquanto as acurácias das classificações da literatura ficaram entre $0,40$ e $0,78$, as acurácias dos ms-NN selecionados ficaram acima de $0,55$ e chegaram a quase $0,90$. Dentre os classificadores da literatura, o classificador por árvore de decisão é o mais acurado, apesar de a mediana possuir valor igual à mediana do SVM. A acurácia mínima do ms-NN é igual à acurácia da mediana do k-NN, que apresentou o pior resultado ($0,55$ de mediana). A média e o desvio padrão das acurácias dessas classificações são apresentadas na Tabela 5.6.

Na Figura 5.5(b), é possível verificar que as acurácias das classificações do ms-NN com GEO são 60% das vezes maiores que as classificações do ms-NN sem GEO. No entanto, essas classificações são estatisticamente iguais de acordo com o índice de desempenho de 90% adotado. Porém, classificação usando o ms-NN com GEO obteve melhores resultados quando comparada às obtidas com outros métodos da

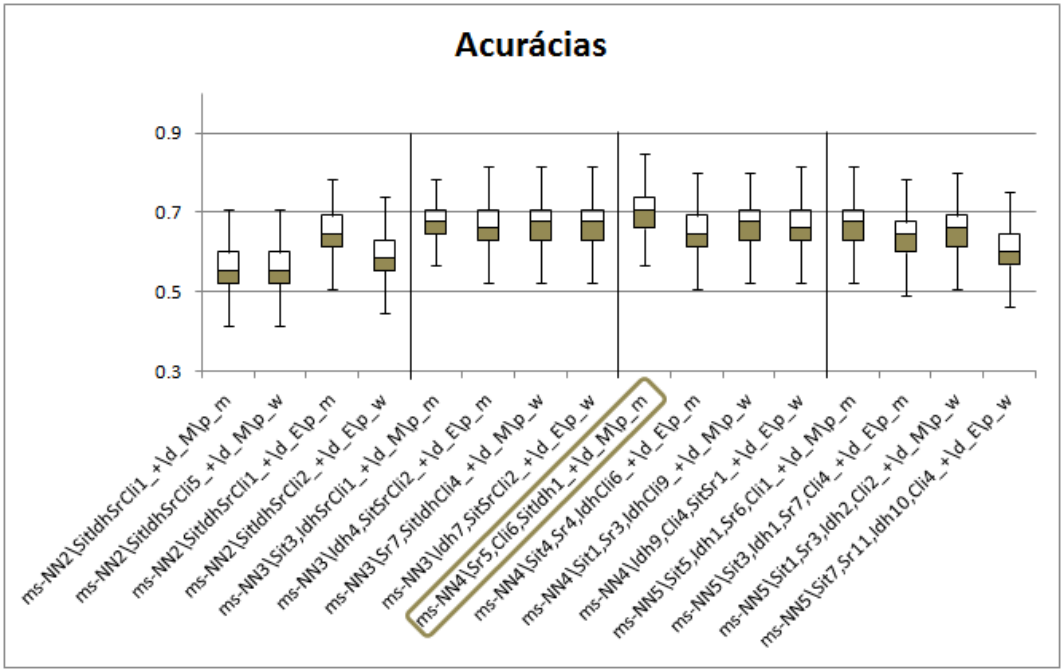
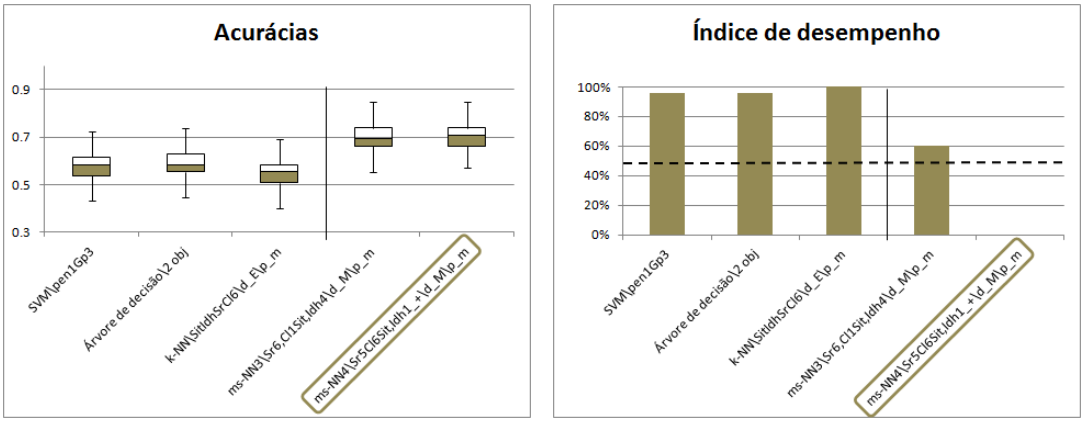


Figura 5.4 - ms-NN com até 4 espaços ATR e 1 espaço Geo, usando todos os conjuntos de atributos com a abordagem com 3 classes, para Schisto por município.



(a) (b)

Figura 5.5 - Comparação dos melhores resultados do ms-NN com os métodos da literatura, quando se usam todos os conjuntos de atributos com a abordagem com 3 classes, para Schisto por município.

literatura abordados aqui. Essa melhora pode ser visualizada na Figura 5.5(b), onde pode-se perceber que, de acordo com o índice de desempenho, a classificação do ms-NN com GEO é estatisticamente superior às classificações da literatura em mais de 95% das vezes.

Tabela 5.6 - Média das acurácias das melhores classificações do ms-NN e métodos da literatura com a abordagem com 3 classes para Schisto por município.

Métodos	média	desvio padrão
SVM	$0,58$	$0,06$
Arv	$0,59$	$0,06$
k-NN	$0,55$	$0,06$
ms-NN3	$0,69$	$0,05$
ms-NN4+	$0,70$	$0,05$

Neste contexto, é possível dizer que o ms-NN4+ é estatisticamente superior aos métodos da literatura abordados aqui e é igual ao ms-NN3 quando se usam todos os conjuntos de atributos e 3 classes. Então, para este experimento e de acordo com a metodologia usada, o uso do espaço GEO não melhora significativamente a acurácia das classificações. No entanto, as matrizes de confusão geradas (Tabelas 5.7 a 5.11) usando a classificação que deu origem aos conjuntos para o estudo Monte Carlo demostra que a classificação pelo ms-NN+ é mais acurada que todos os outros métodos.

Tabela 5.7 - Matriz de confusão do método árvore de decisão para a abordagem com 3 classes para Schisto por município (Ac = 0,58).

Referência	Baixa	Média	Alta
Baixa (15)	10	4	1
Media (24)	6	17	1
Alta (26)	3	12	11

Tabela 5.8 - Matriz de confusão do método SVM para a abordagem com 3 classes para Schisto por município (Ac = 0,57).

Referência	Baixa	Média	Alta
Baixa (15)	10	2	3
Media (24)	5	13	6
Alta (26)	2	10	14

Neste estudo de caso, as classes são ordenadas, por isso alguns erros são considerados inaceitáveis. Um exemplo desse tipo de erro é classificar um município que originalmente é da classe Baixa como Alta, ou vice-versa. Neste trabalho esse tipo de erro

Tabela 5.9 - Matriz de confusão do método k-NN para a abordagem com 3 classes para Schisto por município (Ac = 0,55).

Referência	Baixa	Média	Alta
Baixa (15)	9	3	3
Media (24)	3	7	14
Alta (26)	1	5	20

Tabela 5.10 - Matriz de confusão do método ms-NN para a abordagem com 3 classes para Schisto por município (Ac = 0,60).

Referência	Baixa	Média	Alta
Baixa (15)	8	6	1
Media (24)	3	10	11
Alta (26)	1	4	21

Tabela 5.11 - Matriz de confusão do método ms-NN+ para a abordagem com 3 classes para Schisto por município (Ac = 0,71).

Referência	Baixa	Média	Alta
Baixa (15)	11	4	0
Media (24)	2	14	8
Alta (26)	1	4	21

é considerado ruim. A concepção desse trabalho é que o mapa com a classificação possa ser usado como subsídio para alocações de recursos para os municípios que mais precisam. Com esses erros, municípios poderiam não receber recursos necessários ou os recursos poderiam ser alocados indevidamente. Nas matrizes de confusão apresentadas, são poucos os erros considerados ruins. O maior número de municípios classificados com esses erros, foi na classificação pelo SVM, que obteve no total 5 erros. O ms-NN+ obteve o melhor resultado, com apenas 1 município classificado erroneamente.

Em relação à acurácia dessas classificações, o ms-NN+ obteve uma melhora significativa em relação aos todos os outros métodos. O melhor resultado de acurácia (Ac) dentre os modelos da literatura foi para a classificação por árvore de decisão que obteve $0,58$. O ms-NN obteve $0,60$ enquanto o ms-NN+ obteve $0,71$ de acurácia.

A Figura 5.6 apresenta o mapa com a classificações do ms-NN4+. Esse mapa foi gerado usando a classificação que deu origem aos conjuntos para o estudo Monte

Carlo.

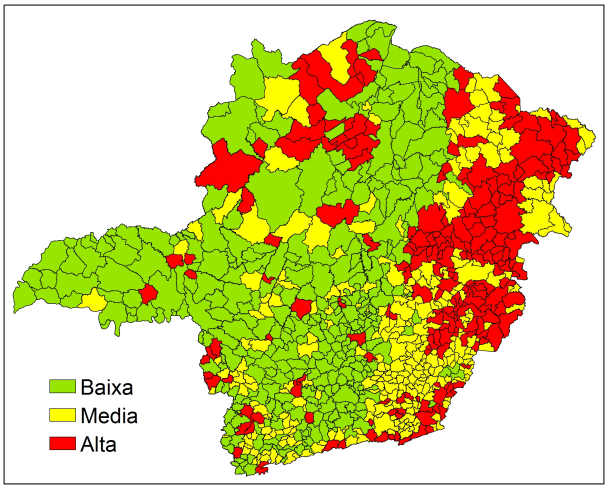


Figura 5.6 - Classificação selecionada para abordagem com 3 classes.

Visualmente, a classificação resultante é bastante condizente com os dados Pv originais (Figura 4.2(a)). A maior parte dos municípios classificadas como de prevalência Alta estão na região noroeste e leste do Estado. Entretanto, alguns municípios foram classificados como de prevalência Alta embora estejam na área indene.

Outra análise feita para este estudo de caso foi classificar 2 e 3 conjuntos de atributos usando o k-NN e o ms-NN com e sem GEO. Os resultados das acurácias são ilustrados nas Figuras 5.7 e 5.8. De acordo com os boxplot, pode-se perceber que as acurácias do ms-NN com e sem GEO na maioria das vezes é melhor que os resultados do k-NN.

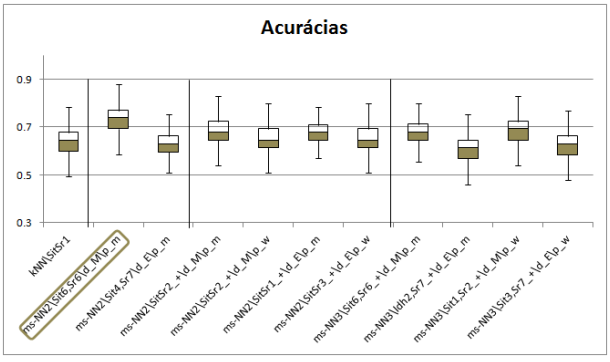


Figura 5.7 - Comparação dos resultados do ms-NN com o k-NN, quando se usam 2 conjuntos de atributos com a abordagem com 4 classes usando 25% dos casos indene, para Schisto por município.

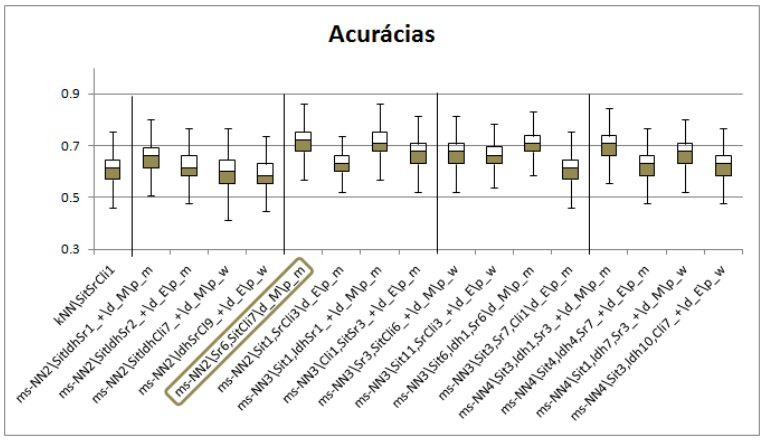


Figura 5.8 - Comparação dos resultados do ms-NN com o k-NN, quando se usam 3 conjuntos de atributos com a abordagem com 4 classes usando 25% dos casos indene, para Schisto por município.

Quando 2 ou 3 conjuntos de atributos são usados, as classificações com maior média foram geradas usando o ms-NN com 2 espaços ATR, distância de Mahalanobis e o voto da maioria como função de predominância. De acordo com o índice de desempenho usado, as acurácias do ms-NN2 são estatisticamente superiores às do k-NN, usando 2 ou 3 conjuntos de atributos. As médias e o desvio padrão das acurácias dessas classificações e das classificações usando o k-NN são apresentadas nas Tabelas 5.12 e 5.13.

Tabela 5.12 - Média das acurácias da melhor classificação do ms-NN e k-NN com a abordagem com 3 classes usando 2 conjuntos de atributos para Schisto por município.

Métodos	média	desvio padrão
k-NN	0,64	0,06
ms-NN2\d_M\p_m	0,73	0,05

Tabela 5.13 - Média das acurácias da melhor classificação do ms-NN e k-NN com a abordagem com 3 classes usando 3 conjuntos de atributos para Schisto por município.

Métodos	média	desvio padrão
k-NN	0,61	0,06
ms-NN2\d_M\p_m	0,72	0,05

5.2.2 Abordagem com 4 classes usando 25% dos casos indenes

As informações sobre a esquistossomose em nível municipal foram disponibilizadas para 501 municípios do Estado de Minas Gerais. Desses, apenas 197 possuem informação positiva sobre a doença e 304 municípios são indenes. Como a classe indene é maior que as outras classes juntas, optou-se por criar um experimento usando, para essa classe, valor igual ao da classe mais amostrada, no caso a classe Alta. O número de casos na classe indene, neste experimento, corresponde a 25% do total casos indenes.

Os resultados com as classificações representantes geradas do ms-NN usando somente espaço de atributos são apresentados na Figura 5.9. Nesta abordagem, o uso de diferentes distâncias não é conclusivo. Os valores das acurácias do ms-NN ficaram entre $0,50$ e $0,81$ nos boxplots. A classificação com 4 espaços ATR usando distância de Mahalanobis foi a mais acurada.

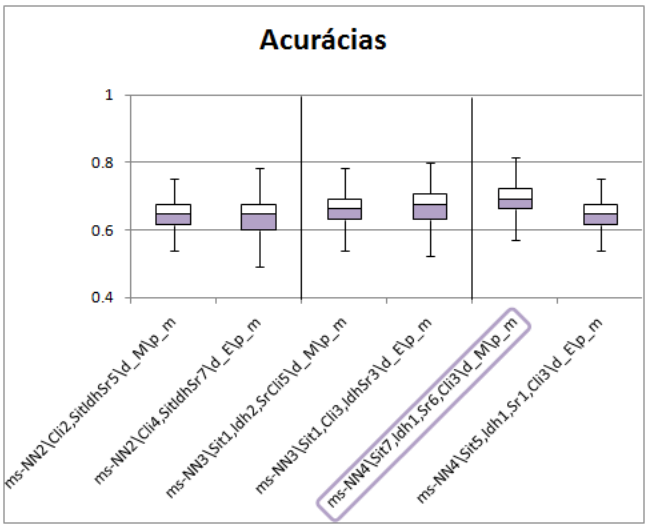


Figura 5.9 - ms-NN com até 4 espaços ATR, usando todos os conjuntos de atributos com a abordagem com 4 classes usando 25% dos casos indene, para Schisto por município.

Usando o espaço GEO, as acurácias das classificações do ms-NN alcançaram entre $0,57$ e $0,86$ de acurácia, conforme Figura 5.10. Já as acurácias das classificações do ms-NN sem GEO ficaram entre $0,49$ e $0,81$ (Figura 5.9). Isso significa que em todos os casos testados, o uso do espaço geográfico aumenta a acurácia do classificador.

As classificações do ms-NN com espaço GEO usando 3 espaços (Figura 5.10) são

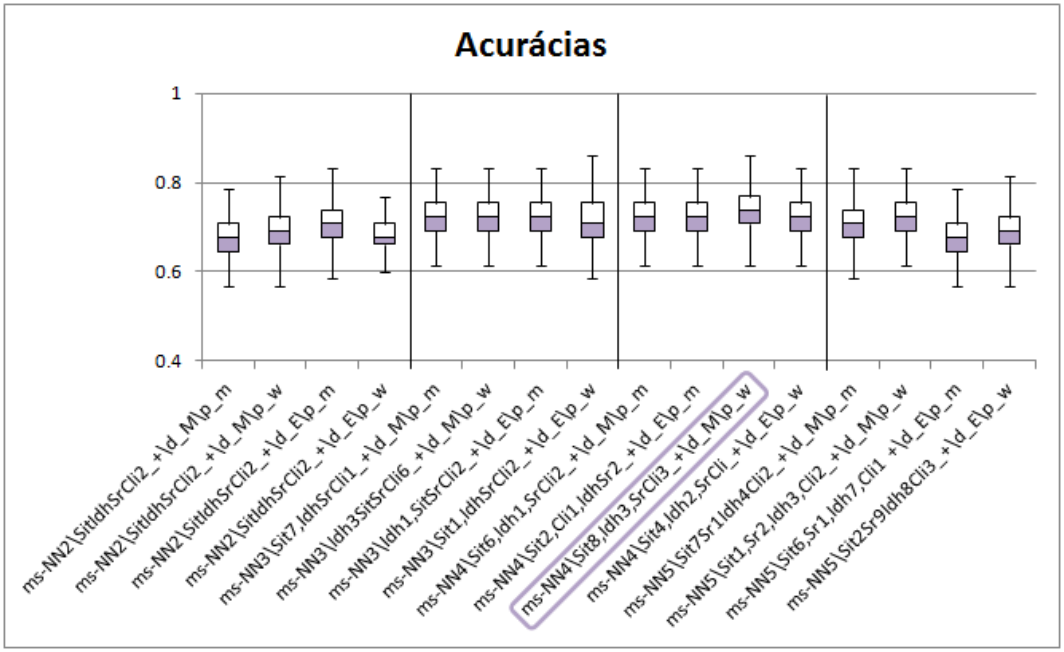


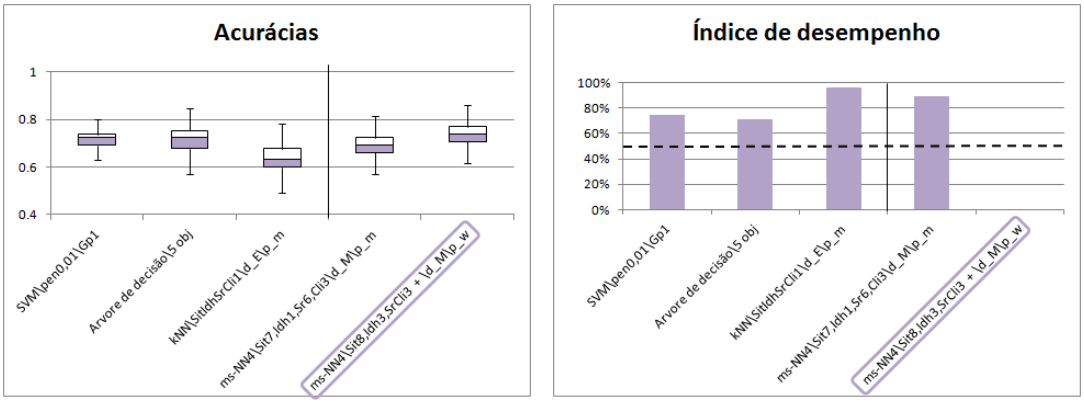
Figura 5.10 - ms-NN com até 4 espaços ATR e 1 espaço Geo, usando todos os conjuntos de atributos com a abordagem com 4 classes usando 25% dos casos indene, para Schisto por município.

idênticas quando se usa a função de predominância maioria (p_m) para ambas as distâncias e ponderada (p_w) para a de Mahalanobis (d_M). O mesmo fato acontece quando se usam 4 espaços com p_m para ambas distâncias.

Das classificações do ms-NN sem GEO apresentadas nas Figura 5.9, a classificação com 4 espaços usando d_M foi selecionada por apresentar a maior média de acurácia. Da mesma forma, a classificação com 4 espaços usando d_M e p_w foi selecionada dentre as classificações do ms-NN com GEO apresentadas na Figura 5.10. Essas classificações, juntamente com as classificações dos métodos da literatura, são apresentadas na Figura 5.11(a).

O SVM e árvore de decisão possuem a mesma mediana (Figura 5.11(b)), no entanto o SVM possui maior média e menor variância. Sendo assim, dentre os métodos da literatura, o SVM é o mais acurado. As acurácias do ms-NN sem GEO é maior que o k-NN, mas menor que o SVM e árvore de decisão. Já a média e o máximo das acurácias do ms-NN com GEO são superiores às acurácias dos outros métodos. A Tabela 5.14 apresenta a médias, o máximo e o desvio-padrão de cada classificação apresentada na Figura 5.11(a).

As acurácias das classificações pelo ms-NN GEO é maior em mais de 65% das vezes,



(a) (b)

Figura 5.11 - Comparação dos melhores resultados do ms-NN com os métodos da literatura, quando se usam todos os conjuntos de atributos com a abordagem com 4 classes usando 25% dos casos indene, para Schisto por município.

Tabela 5.14 - Média das acurácias das melhores classificações do ms-NN e métodos da literatura com a abordagem com 4 classes usando 25% dos casos indene, para Schisto por município.

Métodos	média	desvio padrão
SVM	0, 72	0, 03
Arv	0, 72	0, 05
k-NN	0, 64	0, 05
ms-NN4	0, 69	0, 05
ms-NN4+	0, 74	0, 05

se comparada com as acurácias das classificações pelo SVM e árvore de decisão. Então, de acordo com o índice de desempenho usado neste trabalho (90%), estas classificações são estatisticamente iguais. Porém, em relação ao k-NN, o ms-NN com GEO é estatisticamente superior (96% das vezes melhor). O ms-NN com GEO também é estatisticamente superior ao ms-NN sem GEO. Sendo assim, de maneira geral, a classificação do ms-NN com GEO, usando todos os conjuntos de atributos, é igual ou melhor em relação aos outros métodos.

Além disso, a acurácia da classificação pelo ms-NN4+ usada para gerar os conjuntos para o estudo Monte Carlo é superior as acurácias dos todos os outros métodos (Tabela 5.15). As informações relevantes das matrizes de confusão dessas classificações foram resumidas na Tabela 5.16, onde são mostrados os erros que são considerados ruins (i.e., ser da classe Indene ou Baixa e ser classificado como Alta e vice-versa).

Com exceção da classificação por árvore de decisão, em que 1 município da classe indene foi classificado como classe alta, nenhum município da classe indene foi classificado como alta. Neste contexto, a classificação por árvore de decisão obteve também o maior número de municípios classificados erroneamente (8 erros), enquanto o ms-NN obteve o menor número (1 erro).

Tabela 5.15 - Acurácias das classificações que originaram os conjuntos para o estudo Monte Carlo para a abordagem com 4 classes, usando 25% das amostras indenes, para Schisto por município.

Método	Acurácia
Árvore de decisão	0,64
SVM	0,50
k-NN	0,61
ms-NN	0,71
ms-NN+	0,72

Tabela 5.16 - Resumo das matriz de confusão com os erros considerados ruins para a abordagem com 4 classes, usando 25% das amostras indenes, para Schisto por município.

Método	Número de municípios de uma classe classificados como outra			
	Indene como Alta	Baixa como Alta	Alta como Indene	Alta como Baixa
Árvore de decisão	1	2	3	2
SVM	0	0	2	0
k-NN	0	2	0	2
ms-NN	0	1	0	0
ms-NN+	0	2	0	1

A Figura 5.12 apresenta o mapa com a classificações do ms-NN+. Esse mapa foi gerado usando a classificação que deu origem aos conjuntos para o estudo Monte Carlo. Visualmente essa classificação é semelhante aos dados apresentados na Figura 4.2(a).

As acurácias das classificações do ms-NN quando se usam 2 e 3 conjuntos, na maioria das vezes, apresentou valores mais altos quando comparadas as classificações do k-NN. De acordo com a Figura 5.13, usando 2 conjuntos de atributos, somente uma classificação do ms-NN apresentou pior resultado de acurácia. Quando se usa 3 conjuntos de atributos somente duas das classificações apresentaram resultados de acurácia inferiores aos do k-NN, conforme Figura 5.14.

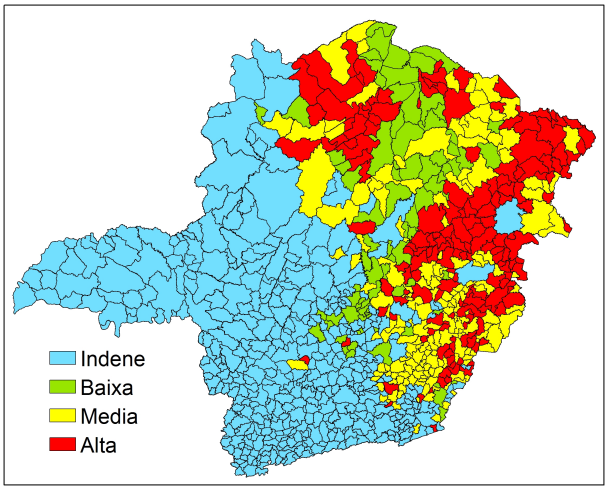


Figura 5.12 - Classificação selecionada para abordagem com 4 classes, usando 25% dos casos indene.

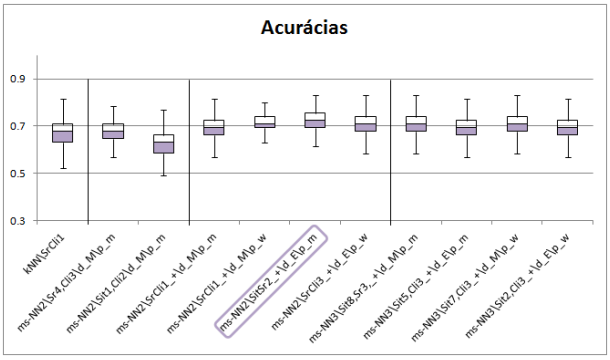


Figura 5.13 - Comparação dos resultados do ms-NN com o k-NN, quando se usam 2 conjuntos de atributos com a abordagem com 4 classes usando 25% dos casos indene, para Schisto por município.

A classificação selecionada como representante do ms-NN, quando são usados 2 conjuntos de atributos, foi gerada usando 2 espaços, um ATR com distância Euclidiana e um GEO usando o voto da maioria como função de predominância. Nas classificações usando 3 conjuntos de atributos foi selecionada a classificação com 3 espaços, dois ATR com distância de Mahalanobis e um GEO usando a função de predominância ponderada pela número de vizinhos do espaço. Essas classificações são superiores às classificações do k-NN, de acordo com o índice de desempenho usado. As Tabelas 5.17 e 5.18 apresentam a média e o desvio padrão das acurácias das classificações do k-NN e das classificações selecionadas do ms-NN.

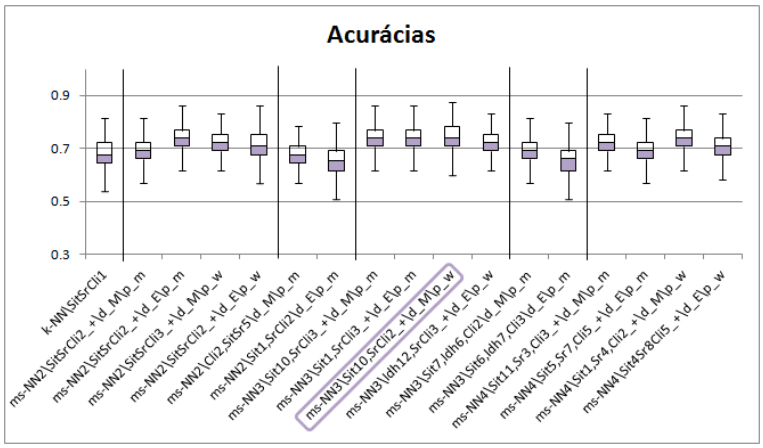


Figura 5.14 - Comparação dos resultados do ms-NN com o k-NN, quando se usam 3 conjuntos de atributos com a abordagem com 4 classes usando 25% dos casos indene, para Schisto por município.

Tabela 5.17 - Média das acurácias da melhor classificação do ms-NN e k-NN com a abordagem com 4 classes usando 25% dos casos indene, e 2 conjuntos de atributos para Schisto por município.

Métodos	média	desvio padrão
k-NN	$0,67$	$0,05$
ms-NN2_+\d_E\p_m	$0,72$	$0,05$

Tabela 5.18 - Média das acurácias da melhor classificação do ms-NN e k-NN com a abordagem com 4 classes usando 25% dos casos indene, e 3 conjuntos de atributos para Schisto por município.

Métodos	média	desvio padrão
k-NN	$0,68$	$0,05$
ms-NN3+\d_M\p_w	$0,75$	$0,05$

5.2.3 Abordagem com 4 classes usando todos os casos indenes

Neste experimento, são usados todos os casos com informação sobre a esquistossomose disponibilizados. A Figura 5.15 apresenta as classificações representantes do ms-NN usando somente espaços ATR. Já na Figura 5.16, são apresentadas as classificações representantes do ms-NN quando usa-se o espaço GEO.

As classificações com 4 espaços ATR são estatisticamente iguais. As outras classificações são estatisticamente diferentes, embora sejam visualmente parecidas. Nesta abordagem, a diferença entre o uso da distância Euclidiana e Mahalanobis também

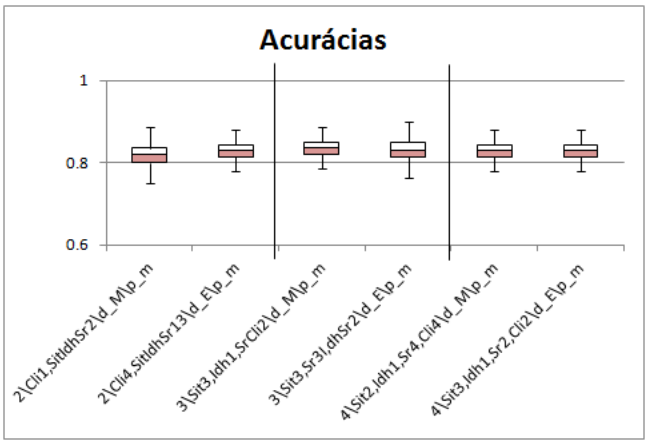


Figura 5.15 - ms-NN com até 4 espaços ATR, usando todos os conjuntos de atributos para a abordagem com 4 classes usando todos os casos indene, para Schisto por município.

não é conclusiva.

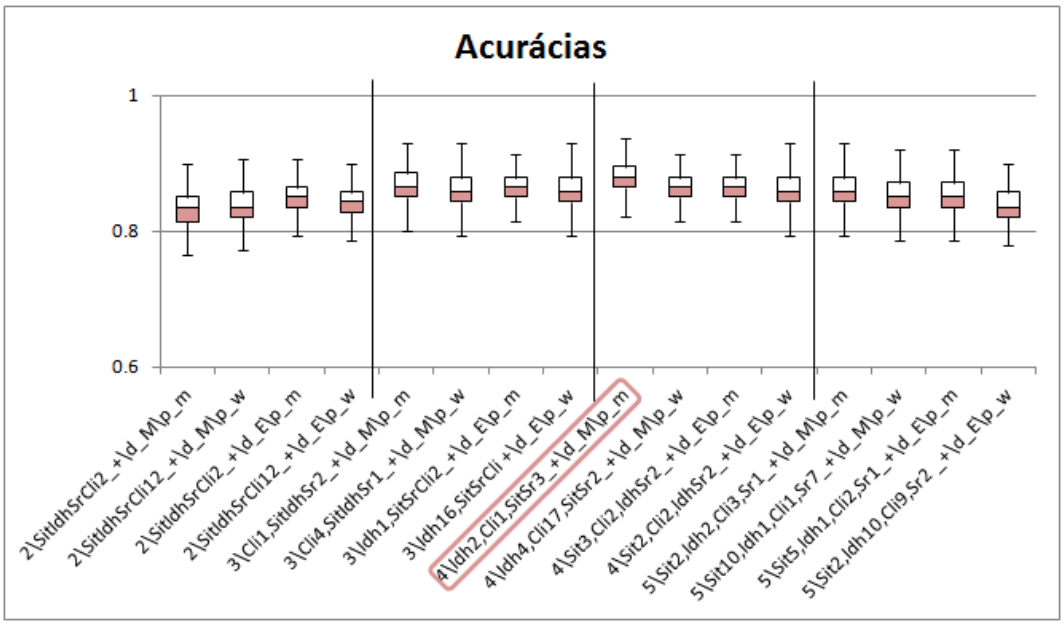


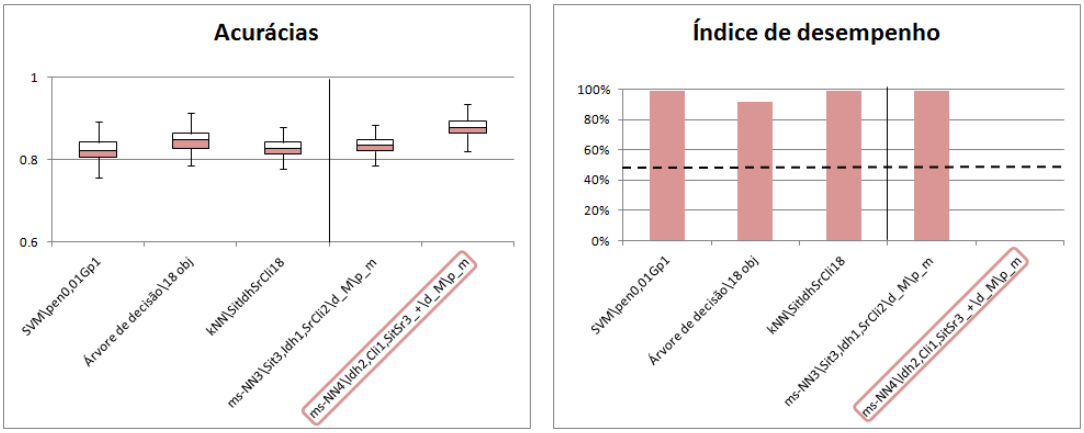
Figura 5.16 - ms-NN com até 4 espaços ATR e 1 espaço Geo, usando todos os conjuntos de atributos para a abordagem com 4 classes usando todos os casos indene, para Schisto por município.

As classificações pelo ms-NN usando apenas espaço ATR obtiveram entre $0,75$ e $0,90$ de acurácia (Figura 5.15). Já as acurácias das classificações do ms-NN com espaço GEO (Figura 5.16) ficaram entre $0,76$ e $0,94$. As medianas das acurácias do

ms-NN sem espaço GEO ficaram entre $0,82$ e $0,84$ e do ms-NN com GEO ficaram entre $0,84$ e $0,87$. Esses resultados comprovam que a acurácia aumenta quando o espaço GEO é usado.

Usando o ms-NN com GEO, o melhor resultado foi obtido pela classificação que usa 4 espaços com d_M e p_m. Os resultados das classificações usando 3 espaços e p_w para ambas distâncias são idênticos, o que pode ser observado na Figura 5.16. O mesmo acontece com as classificações em 4 espaços usando d_M com p_w e usando d_E com p_m. Também são idênticos os resultados das classificações usando 4 d_E e usando p_m com 5 espaços, d_M e p_m. Por fim, as classificações com 5 com d_M\p_w e com d_E\p_m são iguais, de acordo com o boxplot.

As classificações usando 3 espaços ATR e usando 4 espaços, sendo um deles GEO, ambas com d_M e p_m, foram selecionadas para serem comparadas com as classificações da literatura. A Figura 5.17(a) apresenta o boxplot com essas classificações.



(a) (b)

Figura 5.17 - Comparação dos melhores resultados do ms-NN com os métodos da literatura, quando usam-se todos os conjuntos de atributos para a abordagem com 4 classes com todos os casos indene, para Schisto por município.

Na Figura 5.17(a), pode-se perceber o ganho no uso do ms-NN com GEO. De acordo com o boxplot, as acurácias dos modelos da literatura apresentam valores entre $0,75$ e $0,88$ enquanto as acurácias do ms-NN com GEO, valores entre $0,82$ e $0,93$. A menor acurácia do ms-NN com GEO é semelhante à mediana da classificação pelo SVM e pelo k-NN. Dentre os modelos da literatura, a árvore de decisão foi o método mais acurado. As acurácias da árvore de decisão também são superiores às acurácias

do ms-NN sem GEO. A média e o desvio padrão das acurácias de cada método, apresentado na Figura 5.17(a), estão na Tabela 5.19.

Tabela 5.19 - Média das acurácias das melhores classificações do ms-NN e métodos da literatura, quando usam-se todos os conjuntos de atributos para a abordagem com 4 classes com todos os casos indene, para Schisto por município.

Métodos	média	desvio padrão
SVM	$0,82$	$0,02$
Arv	$0,85$	$0,02$
k-NN	$0,82$	$0,02$
ms-NN4	$0,83$	$0,02$
ms-NN4+	$0,88$	$0,02$

O ms-NN4+ é superior à árvore de decisão em mais de 90% das vezes e superior aos outros métodos em 99% das vezes, de acordo com a Figura 5.17(b). Sendo assim, de acordo com o índice de desempenho, a classificação do ms-NN com GEO é estatisticamente superior aos outros métodos abordados.

A superioridade da classificação em relação aos métodos da literatura, abordados neste trabalho, também pode ser observada na Tabela 5.20. Nesta tabela são apresentadas as acurácias das classificações usadas para gerar os conjuntos para o estudo Monte Carlo de cada método. Um resumo das matrizes de confusão dessas classificações com os erros considerados ruins, para este estudo de caso, é mostrado na Tabela 5.21.

Tabela 5.20 - Acurácias das classificações que originaram os conjuntos para o estudo Monte Carlo para a abordagem com 4 classes, usando todas as amostras da classe Indene, para Schisto por município.

Método	Acurácia
Árvore de decisão	0,75
SVM	0,64
k-NN	0,80
ms-NN	0,80
ms-NN+	0,85

A classificação pelo SVM obteve o maior número de municípios da classe Alta classificado como Baixa (Tabela 5.21). É interessante observar também, que a classificação

Tabela 5.21 - Resumo das matrizes de confusão com os erros considerados ruins para a abordagem com 4 classes, usando todas as amostras indenes, para Schisto por município.

Método	Número de municípios de uma classe classificados como outra			
	Indene como Alta	Baixa como Alta	Alta como Indene	Alta como Baixa
Árvore de decisão	0	1	6	2
SVM	3	0	2	23
k-NN	0	1	2	0
ms-NN	0	6	2	0
ms-NN+	0	2	1	1

pelo SVM foi a única que obteve municípios da classe Indene classificado como Alta e não obteve municípios da classe Baixa classificado como Alta. O k-NN obteve o menor número de municípios classificados erroneamente (3 erros), seguido pelo ms-NN+ (4 erros), ms-NN (8 erros) e árvore de decisão (9 erros).

A Figura 5.18 apresenta o mapa com a classificações do ms-NN+. Esse mapa foi gerado usando a classificação que deu origem aos conjuntos para o estudo Monte Carlo.

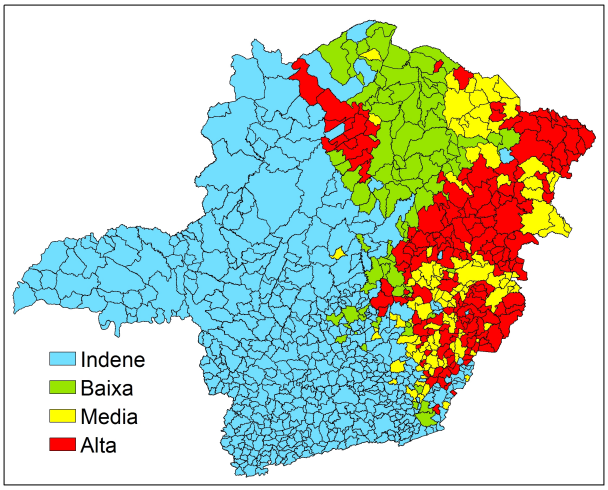


Figura 5.18 - Classificação selecionada para abordagem com 4 classes, usando todos os casos indene.

A adição do espaço GEO no ms-NN aumenta a acurácia quando se usam 2 e 3 conjuntos de atributos. De acordo com a Figura 5.19, usando 2 conjuntos de atributos, a acurácia do ms-NN é igual ou maior que as acurácias das classificações do k-NN. Na Figura 5.20, nota-se que em somente em duas classificações, quando são usados 3

conjuntos de atributos, as acurácias do ms-NN sem GEO são inferiores às do k-NN (ms-NN\d_E usando 2 ou 3 espaços). A diferença entre o k-NN e o ms-NN aumenta, quando é adicionado o espaço GEO, usando-se 2 ou 3 conjuntos de atributos.

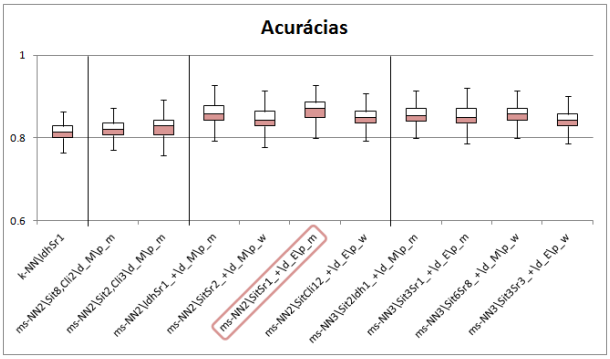


Figura 5.19 - Comparação dos resultados do ms-NN com o k-NN, quando se usam 2 conjuntos de atributos com a abordagem com 3 classes, para Schisto por município.

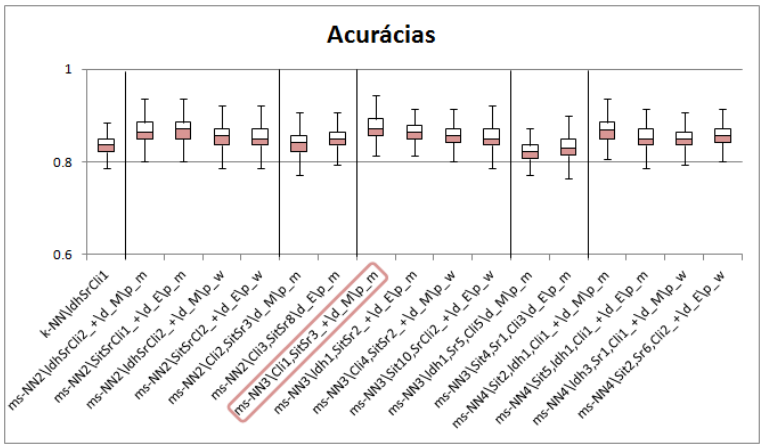


Figura 5.20 - Comparação dos resultados do ms-NN com o k-NN, quando se usam 3 conjuntos de atributos com a abordagem com 3 classes, para Schisto por município.

Das classificações do ms-NN, foi selecionada a que obteve o melhor resultado, quando se usam 2 e 3 conjuntos de atributos, para ser comparada com a do k-NN. A média da acurácia dessas classificações são apresentadas nas Tabelas 5.22 e 5.23. Nos dois casos, a acurácia do ms-NN+ é superior, de acordo com o índice de desempenho, à acurácia do k-NN.

Tabela 5.22 - Média das acurácias da melhor classificação do ms-NN e k-NN com a abordagem com 3 classes usando 2 conjuntos de atributos para Schisto por município.

Métodos	média	desvio padrão
k-NN	$0,67$	$0,05$
ms-NN2_+\d_E\p_m	$0,87$	$0,02$

Tabela 5.23 - Média das acurácias da melhor classificação do ms-NN e k-NN com a abordagem com 4 classes usando todos os casos indene, com 3 conjuntos de atributos para Schisto por município.

Métodos	média	desvio padrão
k-NN	$0,83$	0.02
ms-NN3+\d_M\p_m	$0,88$	$0,02$

5.3 Resultados em escala local

Nos experimentos apresentados nesta Seção foram usados todos os dados sobre a esquistossomose a nível local, descritos através de Ip, que pode assumir os valores Baixa, Média e Alta.

Nesta Seção, são apresentados os gráficos do tipo boxplot com as classificações representantes do o ms-NN com e sem GEO, usando todos os conjuntos de atributos. Os melhores resultados são apresentados nos boxplot, juntamente com as classificações representantes dos métodos da literatura usados aqui. As classificações de cada conjunto de configurações, a partir das quais foram selecionadas as representantes, estão detalhadas no Apêndice C, onde encontram-se as classificações pelo SVM, árvore de decisão e as classificações pre-selecionadas por número de casos vizinhos do k-NN e ms-NN.

Também é apresentada uma segunda análise, em que foram usados 2 conjuntos de atributos para gerar o ms-NN com e sem espaço GEO. Nessa análise as classificações do ms-NN foram comparadas somente com o k-NN.

As classificações representantes geradas do ms-NN usando somente espaço ATR são apresentadas na Figura 5.21. A Figura 5.22 apresenta o boxplot com classificações representantes para o ms-NN com espaço GEO. Nessas Figuras, percebe-se que a nível local as melhores classificações são as que usam distância Euclidiana.

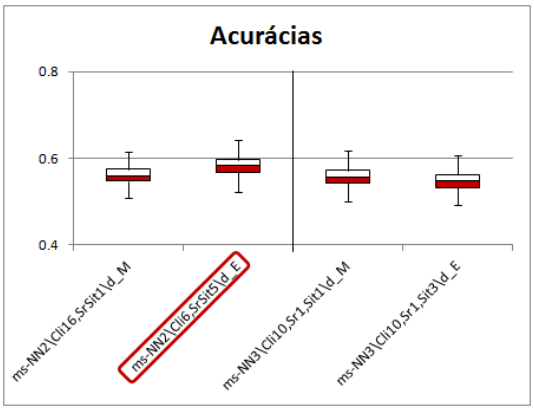


Figura 5.21 - ms-NN com até 3 espaços ATR, usando todos os conjuntos de atributos, para Schisto por localidade.

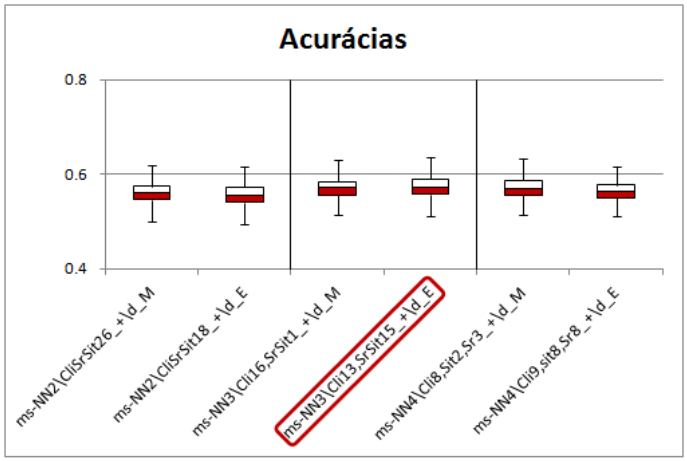
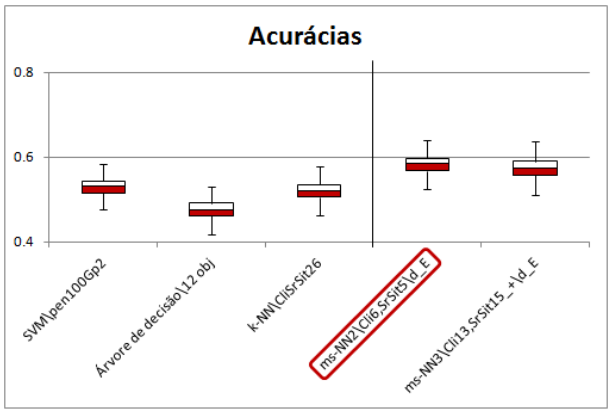


Figura 5.22 - ms-NN com até 3 espaços ATR e 1 espaço GEO, usando todos os conjuntos de atributos, para Schisto por localidade.

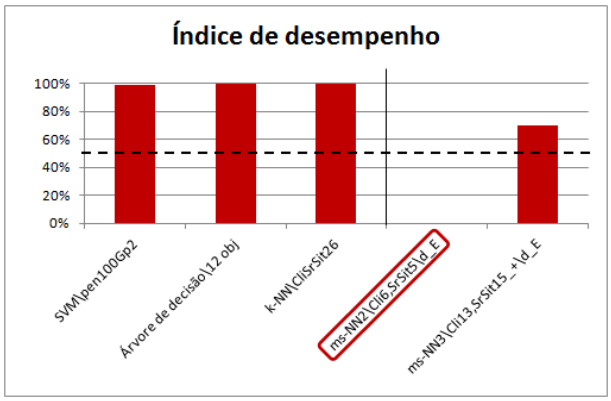
É interessante observar que as classificações usando 2 espaços são parecidas na Figura 5.22. O mesmo fato se repete com as classificações com 3 espaços e com a classificação com 4 espaços usando a distância da Mahalanobis. No entanto, elas não são estaticamente iguais.

As classificações que estão destacadas nas Figuras 5.21 e 5.22 são mostradas novamente na Figura 5.5(a). Nesta Figura, além das duas classificações representantes do ms-NN, têm-se as classificações representantes do SVM, árvore de decisão e k-NN.

A partir da Figura 5.23(a), é possível perceber o ganho do ms-NN em relação aos classificadores da literatura. Utilizando-se ou não o espaço GEO, o ms-NN obteve melhor resultado. Enquanto as acurácias das classificações da literatura ficaram entre



(a)



(b)

Figura 5.23 - Comparação dos melhores resultados do ms-NN com os métodos da literatura, quando se usam todos os conjuntos de atributos, para Schisto por localidade.

0, 41 e 0, 58, as acurácias dos ms-NN selecionados ficaram acima de 0, 51 e chegaram a 0, 64. Dentre os classificadores da literatura, o classificador SVM é o mais acurado, apesar de a mediana possuir valor semelhante à mediana do k-NN. A acurácia mínima do ms-NN é semelhante à acurácia da mediana do SVM e do k-NN. Neste experimento, a árvore de decisão obteve o pior resultado. Os resultados foram alcançados usando o ms-NN sem GEO. A média e o desvio padrão das acurácias dessas classificações são apresentadas na Tabela 5.24.

Na Figura 5.23(b), é possível verificar que as acurácias das classificações do ms-NN sem GEO são estatisticamente iguais às da classificação do ms-NN com GEO, de acordo com índice de desempenho adotado de 90%. A classificação usando o ms-NN sem GEO obtive o melhor resultado quando comparada às classificações obtidas com outros métodos da literatura. Essa melhora pode ser visualizada na Figura 5.23(b), onde pode-se perceber que a classificação do ms-NN sem GEO é

Tabela 5.24 - Média das acurácias das melhores classificações do ms-NN e métodos da literatura para Schisto por localidade.

Métodos	média	desvio padrão
SVM	$0,53$	$0,02$
Árvore de decisão	$0,47$	$0,02$
k-NN	$0,52$	$0,02$
ms-NN2	$0,58$	$0,02$
ms-NN3+	$0,57$	$0,02$

superior às classificações da literatura abordadas aqui em mais de 99% das vezes.

Neste contexto, é possível dizer que o ms-NN2 é estatisticamente superior aos métodos da literatura abordados aqui e é igual ao ms-NN3+ quando se usam todos os conjuntos de atributos. No entanto, comparando as acurácias das classificações que deram origem aos conjuntos para o estudo Monte Carlo (Tabela 5.25), tem-se que a classificação usando o ms-NN+ é a mais acurada. Quanto aos erros considerados ruins, o ms-NN+ e o k-NN possuem os menores números de localidades classificadas erroneamente (Tabela 5.26).

Tabela 5.25 - Acurácias das melhores classificações do ms-NN e métodos da literatura que deram origem ao conjuntos para o estudo Monte Carlo Schisto por localidade.

Métodos	Acurácias
SVM	$0,54$
Árvore de decisão	$0,50$
k-NN	$0,52$
ms-NN2	$0,57$
ms-NN3+	$0,60$

Usando todos os atributos, a classificação do ms-NN3+ que deu origem aos conjuntos para o teste Monte Carlo é apresentada na Figura 5.24. Assim como a nível municipal, a classificação a nível local é condizente com os dados originais.

As acurácias das classificações do k-NN e ms-NN usando 2 conjuntos de atributos podem ser visualizadas na Figura 5.25. De acordo com o ínidice de desempenho, as acurácias da classificação ms-NN2\d_M é 85% da vezes maior que a classificação pelo k-NN. A média das acurácias do k-NN e do ms-NN2 pode ser observada na Tabela 5.27.

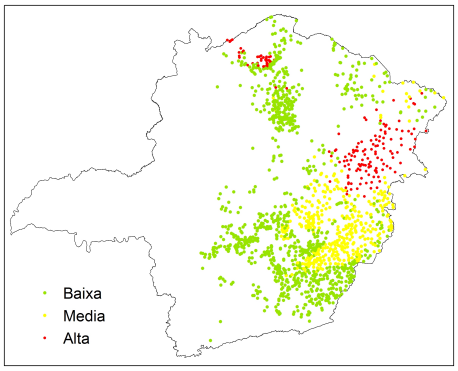


Figura 5.24 - Classificação selecionada (ms-NN3+) para Schisto por localidade.

Tabela 5.26 - Resumo das matriz de confusão com os erros considerados ruins para Schisto por localidade.

Método	Número de municípios de uma classe classificados como outra	
	Baixa como Alta	Alta como Baixa
Árvore de decisão	9	15
SVM	9	15
k-NN	8	11
ms-NN2	10	15
ms-NN3+	7	12

Tabela 5.27 - Média das acurácias da melhor classificação do ms-NN e k-NN usando 2 conjuntos de atributos para Schisto por localidade.

Métodos	média	desvio padrão
k-NN	$0,54$	$0,02$
ms-NN2\d_M	$0,57$	$0,02$

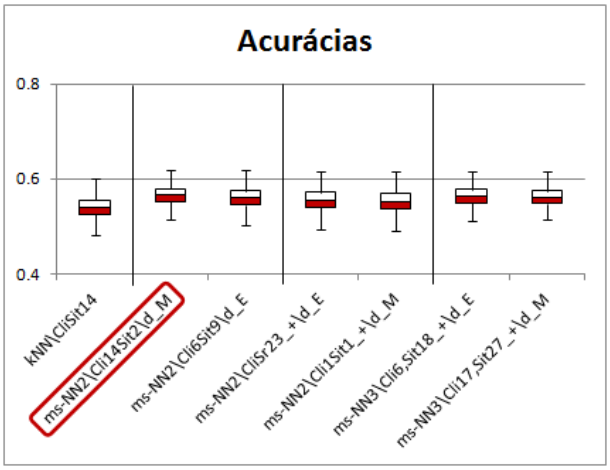


Figura 5.25 - Comparação dos melhores resultados do ms-NN com o k-NN, quando se usam 2 conjuntos de atributos, para Schisto por localidade.

5.4 Conclusões

Neste estudo de caso foram realizados 4 experimentos, 3 em nível municipal e 1 em nível local. Em todos eles, o ms-NN foi testado com e sem espaço GEO e comparados com o SVM, árvore de decisão e k-NN. Em nível municipal, o espaço GEO foi usado com distância por contiguidade e vizinhança variável. Neste caso, foram testadas as funções de predominância maioria e ponderada pelo número de vizinhos do espaço. Em nível local, o espaço GEO foi usado com distância Euclidiana e vizinhança fixa. Nesta abordagem, foi usada somente a função de predominância maioria.

De modo geral, o ms-NN com e sem espaço GEO é superior ao k-NN quando se usam 2, 3 ou 4 conjuntos de atributos. Em todas as abordagens em nível municipal e na abordagem em nível local, os resultados obtidos pelo ms-NN são superiores aos resultados dos métodos da literatura abordados aqui. Em relação aos erros considerados ruins, o ms-NN obteve o menor número de erros. Além disso, na maioria das vezes, a acurácia do ms-NN com GEO é superior ao ms-NN sem GEO.

As acurácias das classificações em nível municipal foram maiores que as acurácias em nível local. O mesmo aconteceu em Guimarães (2010). As acurácias mais altas foram obtidas para as abordagem com 4 classes, sendo mais alta as acurácias quando foram usados todas as amostras da classe indene.

Foram gerados 3 mapas com as classificações para cada abordagem em nível municipal e 1 mapa com a abordagem em nível local. Em todos os casos, a classificação é condizente com os dados originais.

6 EXPERIMENTOS DO ESTUDO DE CASO TAPAJÓS

6.1 Descrição dos experimentos

Um dos objetivos desse estudo de caso é classificar o uso e cobertura de solo usando dados de SR e derivados. Para este estudo de caso são usados no total 428 segmentos rotulados apresentados Tabela 4.3. Dessas amostras, foram selecionadas 2/3 como conjunto de treinamento (T). A partir do conjunto T, foi feita a classificação dos polígonos da segmentação. Para validar a classificação, foram realizadas duas abordagens para o conjunto de teste P. Na primeira abordagem, foi feito o teste por segmento (P_{seg}), em que foram usados todos os polígonos de teste. Para a segunda abordagem, o teste foi feito por pixel (P_{px}). Neste teste foram selecionados 150 pixels de cada classe que receberam o rótulo dos segmentos rotulados. O número de casos de teste para cada abordagem pode ser visto na Tabela 6.4. Neste estudo de caso, o número de amostras de treinamento (285) é o mesmo para todas as abordagens. Para este estudo de caso, foi feito um teste inicial usando 3-folders. Como os resultados eram muito semelhantes e o número de classificações do ms-NN era muito grande, optou-se por usar apenas um conjunto de treinamento e teste.

Tabela 6.1 - Número de casos de teste do modelo ms-NN para Tapajós

Abordagem	# de casos de teste	tipo de teste
Primeira (P_{seg})	143	segmentos
Segunda (P_{px})	900	pixels

Neste estudo de caso, foram usados 4 conjuntos de atributos:

- derivados sensor TM (Tm):
  - B_3;
  - B_4;
  - B_5.
- derivados do sensor PALSAR (Rd):
  - HH;
  - HV.
- MLME (Mm):

  - Veg;
  - Solo;
  - Somb.

- textura (Tx) dos atributos do sensor TM:
  - Entropia;
  - Homogeneidade;
  - Dissimilaridade.

As descrições desses atributos são apresentados na Tabela 6.2. Os métodos k-NN e ms-NN foram testados usando combinações de 2, 3 e 4 dos conjuntos de atributos, conforme Tabela 6.3. As classificações usando os métodos da literatura (SVM, árvore de decisão e K-NN) foram geradas usando todos os conjuntos de atributos juntos. Nas classificações usando o ms-NN, os conjuntos de atributos foram divididos em espaços conforme Tabela 6.3. Neste caso todos os atributos foram usados numa mesma classificação do ms-NN.

Tabela 6.2 - Atributos – estudo de caso Tapajós.

Sigla	Sensor	Descrição
B_3	TM	Média dos pixels, por segmentos, da banda 3 da imagem TM
B_4	TM	Média dos pixels, por segmentos, da banda 4 da imagem TM
B_5	TM	Média dos pixels, por segmentos, da banda 5 da imagem TM
HH	PALSAR	Média dos pixels, por segmentos, da polarização HH da imagem PALSAR
HV	PALSAR	Média dos pixels, por segmentos, da polarização HH da imagem PALSAR
Veg	–	Média dos pixels, por segmentos, da imagem-fração vegetação
Solo	–	Média dos pixels, por segmentos, da imagem-fração solo
Somb	–	Média dos pixels, por segmentos, da imagem-fração sombra
$entr_3$	–	Entropia para os segmentos da banda 3 da imagem TM
$entr_4$	–	Entropia para os segmentos da banda 4 da imagem TM
$entr_5$	–	Entropia para os segmentos da banda 5 da imagem TM
$homog_3$	–	Homogeneidade para os segmentos da banda 5 da imagem TM
$homog_4$	–	Homogeneidade para os segmentos da banda 5 da imagem TM
$homog_5$	–	Homogeneidade para os segmentos da banda 5 da imagem TM
$dissim_3$	–	Dissimilaridade para os segmentos da banda 5 da imagem TM
$dissim_4$	–	Dissimilaridade para os segmentos da banda 5 da imagem TM
$dissim_5$	–	Dissimilaridade para os segmentos da banda 5 da imagem TM

Na Tabela 6.3, a vírgula é usada para distinguir os conjuntos de atributos usados em diferentes espaços. As classificações dos métodos da literatura estão apresentadas

Tabela 6.3 - Combinações de atributos por espaço - Tapajós

Métodos Literatura	2 Esp	3 Esp	4 Esp
TmRdMmTx	Tm,RdMmTx	Tm,Tx,RdMm	Tm,Rd,Mm,Tx
	Rd,TmMmTx	Tm,Rd,MmTx	
	Mm,TmRdTx	Tm,Mm,RdTx	
	Tx,TmRdMm	Rd,Mm,TmTx	
		Rd,Tx,TmMm	
		Mm,Tx,TmRd	

na primeira coluna da Tabela 6.3. Na segunda, terceira e quarta colunas, estão os conjuntos de atributos usados para compor 2, 3 e 4 espaços ATR.

A Tabela 6.4 apresenta o número total de classificações para cada método. Para árvore de decisão, SVM e k-NN, foram usados os parâmetros definidos na Seção 4.3, sendo que para o k-NN foram usados até 16 casos vizinhos. Para cada abordagem, do total de classificações dos métodos da literatura, foi selecionada uma classificação representante.

Para o ms-NN, foram testados de 1 a 16 vizinhos, em até 5 espaços, com dois tipos de distâncias (Euclidiana e Mahalanobis) e foi usado o voto da maioria como função de predominância. Nos experimentos com o espaço GEO, foram usados dois tipos de função de predominância, maioria e ponderada. Nos experimentos realizados somente com espaço ATR (até 4 espaços), foi usada a função de predominância maioria. Nas classificações geradas usando 4 espaços ATR com e sem espaço GEO, por uma questão computacional, foram usados de 1 a 10 vizinhos por espaço. Em síntese, no método ms-NN, além da variação do número de vizinhos, para cada conjunto de atributos da Tabela 6.3, foram feitas 4 classificações variando a distância e a função de predominância. totalizando 427.264 classificações.

Nos experimentos do modelo ms-NN em escala municipal, foram testados de 1 a $\sqrt{tr}$ vizinhos, usando dois tipos de distâncias (Euclidiana e Mahalanobis), em até 5 espaços. Nos experimentos com o espaço GEO, foram usados dois tipos de função de predominância, maioria e ponderada. Nos experimentos realizados somente com espaço ATR (até 4 espaços), foi usada a função de predominância maioria. Na Tabela 5.5 é possível verificar o número de classificações resultantes para cada abordagem.

Para cada distância usada no ms-NN, foram pre-selecionadas as melhores classificações. Essas classificações representam as melhores configurações de números de

Tabela 6.4 - Número de classificações para cada método - Tapajós.

Método			# classificações
Árvore de decisão			19
SVM			18
k-NN			16
ms-NN	2 Espaços	1 ATR e 1 GEO	64
		2 ATR	2.048
	3 Espaços	2 ATR e 1 GEO	4.096
		3 ATR	49.152
	4 Espaços	3 ATR e 1 GEO	98.304
		4 ATR	2.000
	5 Espaços	4 ATR e 1 GEO	40.000

vizinhos em cada espaço. A escolha dessas classificações foi feita de acordo com a Seção 4.3.

As médias das acurácias das classificações pre-selecionadas, para cada abordagem de teste, estão apresentadas em gráficos no Apêndice D. Também, no Apêndice D estão apresentados as médias das acurácias das classificações dos métodos SVM, árvore de decisão e k-NN.

Para cada abordagem de teste foram escolhidas seis classificações representantes do ms-NN, duas para cada número de espaço. Essas classificações foram escolhidas para serem comparadas com as classificações representantes dos métodos da literatura.

A classificação representante, selecionada como a melhor dentre as classificações do ms-NN, foi comparada com as outras classificações representantes usando o indice de desempenho. Nesses gráficos, a classificação que aparece sem dados é a classificação que está sendo comparada com as outras classificações.

Para melhor visualização dos resultados, os boxplot do ms-NN, apresentados nesta Seção, estão separados em blocos que representam o número de espaços. Cada classificação representante desses boxplots está descrita pela notação A\B\C, onde:

- A se refere ao número de espaços;
- B se refere à configuração (conjunto de atributos e números de vizinhos, separados por vírgula); e
- C se refere à função utilizada para se calcular os vizinhos, com “d_E” e

"d_M" denotando as distâncias Euclidiana e Mahalanobis.

As componentes acima podem ser omitidas em uma configuração desde que não prejudique o entendimento do texto.

6.2 Resultados e análises

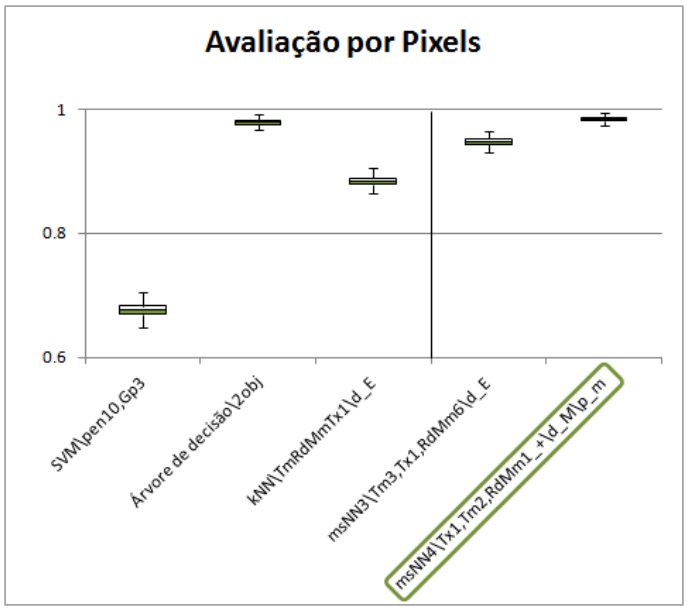
A partir das classificações representantes (configurações em destaque em cada gráfico apresentado no Apêndice D) foi gerado um gráfico do tipo boxplot para cada abordagem de teste. Esses gráficos podem ser vistos nas Figura 6.1. Em destaque nos gráficos, estão as classificações que apresentaram o melhor resultado em cada abordagem de teste.

A partir da Figura 6.1, percebe-se que, de modo geral, o teste por segmento é mais acurado que o teste por pixel. No entanto, a menor variância foi encontrada no teste por pixel. De todas as classificações representantes, o SVM é o classificador que obteve o pior resultado para as duas abordagens de teste. Dentre os classificadores da literatura, o classificador por árvore de decisão é o mais acurado no teste por pixel e segmento. É interessante que, nos dois tipos de teste, as classificações usando o ms-NN apresentaram melhores resultados, quando comparados aos obtidos usando-se SVM ou k-NN. Por fim, o ms-NN com GEO aparece com o melhor classificador. As médias e os desvios-padrão das acurácias das classificações são apresentadas na Tabela 6.6.

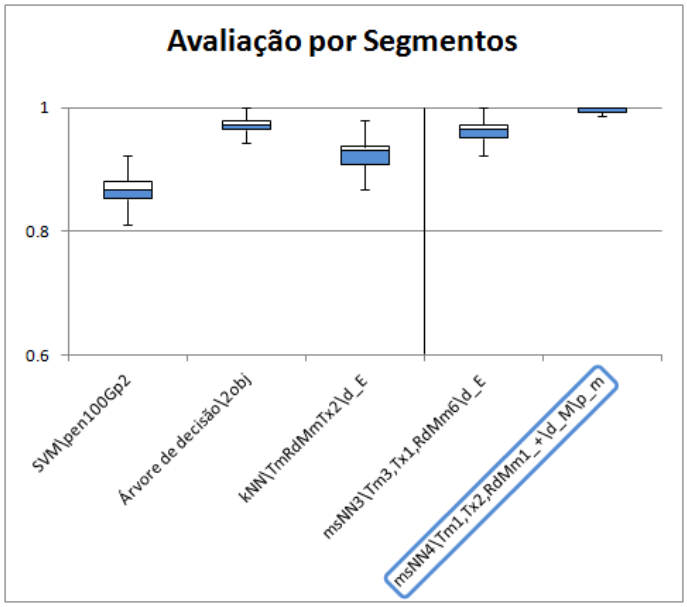
Tabela 6.5 - Média das acurácias das classificações selecionadas como representantes para as duas abordagens de teste.

Método	Pixel		Segmento	
	média	desvio padrão	média	desvio padrão
SVM	$0,68$	$0,01$	$0,87$	$0,02$
Árvore de decisão	$0,97$	$0,01$	$0,97$	$0,01$
k-NN	$0,88$	$0,01$	$0,93$	$0,02$
ms-NN3\d_E	-	-	$0,96$	$0,02$
ms-NN3\d_M	$0,95$	$0,01$	-	-
ms-NN4+\d_M\p_m	$0,98$	$< 0,01$	-	-
ms-NN4+\d_M\p_m	-	-	$0,99$	$0,01$

De acordo com o índice de desempenho apresentado na Figura 6.2(a), as acurácias do ms-NN+, na avaliação teste por pixel, são superior às acurácias das classifica-



(a)

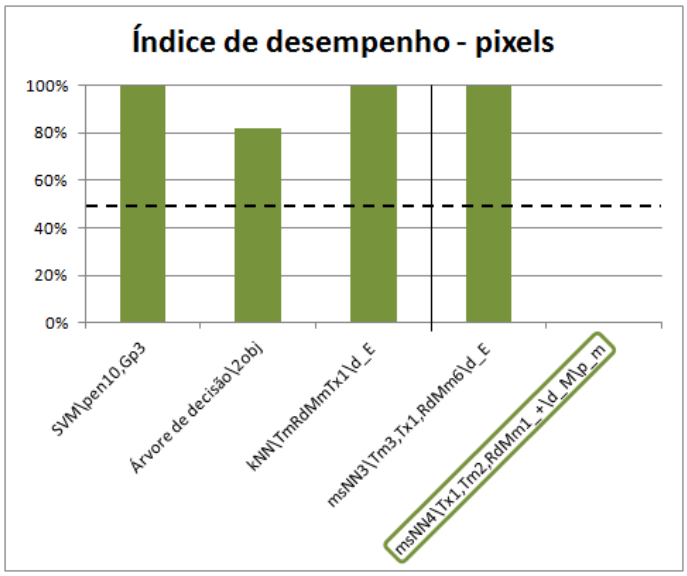


(b)

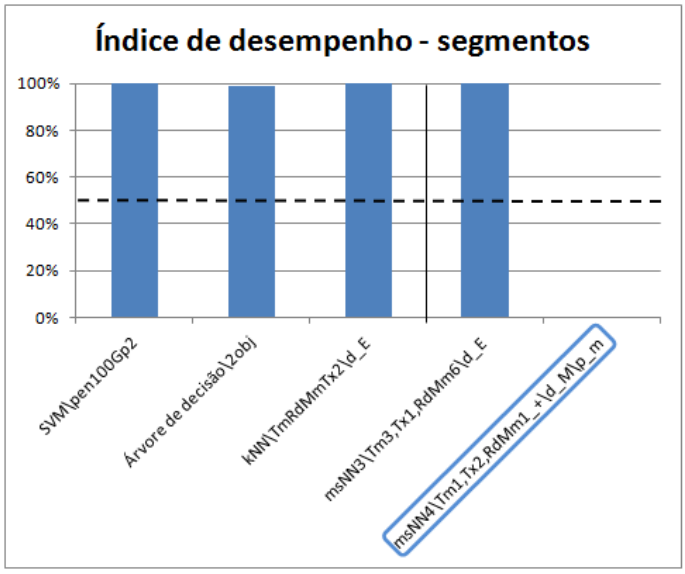
Figura 6.1 - Acurácias das classificações representantes.

ções usando SVM, k-NN e ms-NN. Em relação à árvore de decisão, as acurácias do ms-NN+ são mais de 100% das vezes maiores. Na avaliação por segmento (Figura 6.2(b)), as acurácias do ms-NN+ são superiores a todos outros métodos abordados aqui. As imagens classificadas usando cada um dos métodos são apresentadas na Figura 6.3.

Na Figura 6.3, é possível perceber que as classificações são similares. Nesta Figura,



(a)



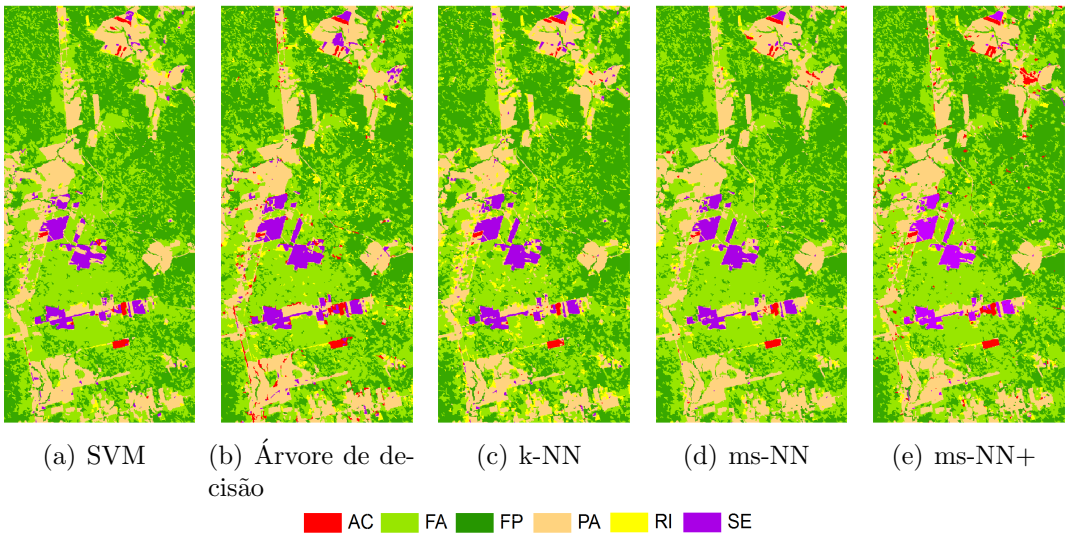
(b)

Figura 6.2 - Índice de desempenho para o teste por pixel e segmento.

na parte sudoeste das imagens, existem áreas que são pasto (PA) e/ou rodovia e foram classificadas como AC principalmente pela árvore de decisão (Figura 6.3(b)). O ms-NN (Figura 6.3(d)) foi o único método que classificou corretamente. Nota-se também, que existem vários segmentos de RI distribuídos na imagem da classificação por árvore de decisão e pelo k-NN em lugares que deveriam ser FP ou FA. Esses segmentos não aparecem na classificação do ms-NN mostrada na Figura 6.3(d) e do SVM apresentadas Figura 6.3(a).

Tabela 6.6 - Acurácias das classificações selecionadas como representantes usadas para gerar os conjuntos do estudo Monte Carlo, para abordagens de teste por pixel.

Método	Acurácia
SVM	0,88
Árvore de decisão	0,91
k-NN	0,89
ms-NN3\d_M	0,95



(a) SVM (b) Árvore de decisão (c) k-NN (d) ms-NN (e) ms-NN+

Figura 6.3 - Imagens classificadas usando melhores classificações da abordagem de teste por pixel.

As matrizes de confusão dessas classificações podem ser vistas nas Tabelas 6.7 a 6.10, onde o k-NN (Tabela 6.9) é o único que confunde a classe PA com outras classes e a classe RI com a classe FP. Nessas Tabelas, também pode-se perceber que o ms-NN consegue distinguir melhor a classe RI da classe FA em relação aos outros métodos. Além disso o ms-NN possui o menor número de erros de classificação da classe FA com a classe FP e vice-versa.

Tabela 6.7 - Matriz de confusão para o SVM.

referência	FP	RI	SE	AC	FA	PA
FP (1096)	1016	0	0	0	80	0
RI (317)	0	109	0	0	169	39
SE (459)	0	0	297	0	0	162
AC (285)	0	0	0	263	0	22
FA (1225)	111	0	0	0	1114	0
PA (1431)	0	0	0	0	0	1431

Tabela 6.8 - Matriz de confusão para a árvore de decisão.

referência	FP	RI	SE	AC	FA	PA
FP (1096)	997	0	0	0	99	0
RI (317)	0	177	0	0	140	0
SE (459)	0	0	459	0	0	0
AC (285)	0	0	0	221	0	64
FA (1225)	111	0	0	0	1114	0
PA (1431)	0	0	0	0	0	1431

Tabela 6.9 - Matriz de confusão para o k-NN.

referência	FP	RI	SE	AC	FA	PA
FP (1096)	1069	0	0	0	27	0
RI (317)	63	148	0	0	106	0
SE (459)	0	0	459	0	0	0
AC (285)	0	0	0	270	0	15
FA (1225)	147	35	0	0	977	66
PA (1431)	0	36	53	0	0	1342

Tabela 6.10 - Matriz de confusão para o ms-NN.

referência	FP	RI	SE	AC	FA	PA
FP (1096)	1012	0	0	0	84	0
RI (317)	0	235	0	0	82	0
SE (459)	0	0	459	0	0	0
AC (285)	0	0	0	270	0	15
FA (1225)	50	0	0	0	1175	0
PA (1431)	0	0	0	0	0	1431

Tabela 6.11 - Matriz de confusão para o ms-NN+.

referência	FP	RI	SE	AC	FA	PA
FP (1096)	966	0	0	0	130	0
RI (317)	0	177	0	0	140	0
SE (459)	0	0	459	0	0	0
AC (285)	0	0	0	270	0	15
FA (1225)	50	0	0	0	1175	0
PA (1431)	0	0	0	0	0	1431

6.3 Conclusões

Neste estudo de caso, o ms-NN foi testado usando espaços ATR e GEO e como critério de predominância foram usados a maioria e ponderação. Foram realizadas duas abordagens de teste, por pixel e por segmento. Em ambas as abordagens, as acurácias do ms-NN+ obtiveram os melhores resultados quando comparados com os métodos abordados aqui. É interessante que as acurácias do ms-NN sem GEO foram superiores às acurácias do SVM e do k-NN.

Foram geradas 5 imagens classificadas, uma para cada método para a abordagem por pixel. A partir dessas imagens, pode-se perceber que, embora no estudo Monte Carlo a classificação por árvore de decisão alcance a maior média de acurácia, a classificação pelo ms-NN possui resultados melhores, porque gerou menos confusão entre as classes.

7 ms-NN COM RELAÇÕES DIFUSAS

Neste capítulo, é abordado o uso de relações difusas como base para classificação. Inicialmente são descritos conceitos básicos da Teoria dos Conjuntos Difusos, para então apresentar duas abordagens para classificação baseadas nesta Teoria. Na primeira abordagem usando relações difusas parametrizadas, não foi utilizado o conceito de ms-NN, mas uma aplicação desta abordagem para dados do Schisto mostra que seu uso em ms-NN é promissor. A segunda abordagem, de relações difusas obtidas a partir de partições difusas, foi aplicada com o ms-NN no estudo de caso do Tapajós.

7.1 Conceitos básicos da Teoria dos Conjuntos Difusos

A Teoria dos Conjuntos Difusos foi introduzida por Lotfi Zadeh em 1965 (ZADEH, 1965) para modelar matematicamente conceitos vagos do discurso humano (veja também Dubois e Prade (1988)). Na teoria clássica dos conjuntos, pode-se modelar a pertinência de um elemento a um conjunto qualquer com valores 0 ou 1, ou seja, o elemento pertence ou não ao conjunto. Um conjunto difuso, por outro lado, considera a possibilidade de pertinência parcial, utilizando o intervalo $[0, 1]$ ao invés do conjunto $\{0, 1\}$. Todo conjunto clássico, aqui chamado de nítido, é um caso particular de conjunto difuso.

Um conjunto difuso A, em um universo U, é descrito através de uma função de pertinência $A(x) : \Omega \rightarrow [0, 1]$, que mapeia os valores do domínio no intervalo dos reais em $[0, 1]$. A função de pertinência $A(x)$ indica o grau de compatibilidade entre x e o conceito expresso por A:

- $A(x) = 1$, x é completamente compatível com A;
- $A(x) = 0$, x é completamente incompatível com A;
- $0 < A(x) < 1$, x é parcialmente compatível com A, com grau $A(x)$;

O conjunto difuso A é dito ser normalizado quando $\exists w \in \Omega, A(w) = 1$.

Seja um conjunto difuso A definido em U, com função de pertinência $A(x) : U \rightarrow [0, 1]$. O núcleo de A é definido como

$$Nu(A) = \{x \in U / A(x) = 1\} = A_1 \tag{7.1}$$

e seu suporte como

$$Su(A) = \{x \in U / A(x) > 0\} = \lim_{\alpha \to 0} A_\alpha. \tag{7.2}$$

Na maior parte das aplicações, um conjunto difuso normalizado A pode ser representado utilizando a notação $< a, b, c, d >$, onde $Su(A) = [a, d]$ e $Nu(A) = [b, c]$ e as funções entre a e b e d e c são estritamente monotônicas (crescente e decrescente, respectivamente). Quando a função de pertinência é linear por partes e $Nu(A) > 1$, o conjunto difuso é chamado de trapezoidal. Um conjunto difuso triangular é tal que $Nu(A) = 1$ e pode ser denotado simplificadamente por $< a, b, d >=< a, b, b, d >$.

A conjunção e a disjunção de conjuntos difusos são obtidas utilizando-se os operadores chamados de T-normas e T-conormas, respectivamente. Uma T-norma $\top : [0, 1] \times [0, 1] \to [0, 1]$ satisfaz as seguintes propriedades $x, y, z, w \in [0, 1]$, $\top$:

- **Comutatividade:** $x \top y = y \top x$
- **Associatividade:** $(x \top y) \top z = x \top (y \top z)$
- **Monotonicidade:** $x \top w \leq y \top z$, se $x \leq y, w \leq z$
- **Elemento neutro = 1:** $x \top 1 = x$

Uma T-conorma $\perp : [0, 1] \times [0, 1] \to [0, 1]$ obedece a comutatividade, associatividade e monotonicidade, mas tem 0 como elemento neutro, i.e. $x \perp 0 = x$. As principais T-normas são o mínimo e o produto e a principal T-conorma é o máximo.

Uma coleção de conjuntos difusos $\mathbf{A}= \{A_1, ..., A_n\}$ em X é chamada de partição difusa.

Uma relação difusa nada mais é que um conjunto difuso definido em um domínio multidimensional Ω, i.e., $R : \Omega \to [0, 1]$. Uma relação difusa pode ser usada para medir o quanto dois elementos de Ω são similares (ou próximos). Em particular, neste caso, $S(w, w_0) = 1$ significa que w e w_0 são indistinguíveis, enquanto que $S(w, w_0) = 0$ significa que w e w_0 não têm nada em comum. Também se pode entender $d_S(w, w_0) = 1 - S(w, w_0)$ como uma relação de dissimilaridade entre w e w_0.

Propriedades normalmente requeridas para estas relações são reflexividade ($\forall w \in \Omega, S(w, w) = 1$) e simetria ($\forall w \in \Omega, S(w, w_0) = S(w_0, w)$), sendo neste caso

chamadas de Relações de Proximidade. Elas são usualmente chamadas de Relações de Similaridade quando obedecem T-transitividade dada por $\forall w, w', w_0 \in \Omega, S(w, w') \geq \top(S(w, w_0), S(w_0, w'))$. A relação S é dita ser separável quando $\forall w, w_0 \in \Omega, S(w, w_0) = 1$ sse $w = w_0$. Neste trabalho, serão consideradas relações de proximidade, que não são necessariamente T-transitivas nem separáveis. Por abuso de linguagem, podemos usar o termo "similar" para descrever a relação entre dois elementos quaisquer de Ω de uma forma geral.

Na Seção a seguir, são apresentadas duas abordagens para classificação utilizando relações difusas.

7.2 Classificação com relações difusas parametrizadas

Em Armengol et al. (2005) foi proposta uma abordagem para Raciocínio Baseado em Casos (CBR, do inglês *Case Based Reasoning*) utilizando relações difusas para verificar a similaridade entre dois casos quaisquer.

Uma base de casos em CBR é composta de casos do tipo $c = (a, \omega)$, onde a é a descrição do caso, com valores para um conjunto de atributos, e ω sua classe associada. A atribuição de uma classe para um novo caso c_0 depende da semelhança entre a_0 e as descrições dos casos na base. Na abordagem proposta em Armengol et al. (2005), essa semelhança é calculada como uma média ponderada das funções de similaridade existentes para cada atributo. Neste trabalho, os vetores de peso são calculados de maneira a minimizar o erro de classificação dos casos já contidos na base. Uma vez calculada a similaridade por atributo, o novo caso é classificado usando uma função de agregação (média simples).

A abordagem Armengol et al. (2005) foi utilizada em (MARTINS-BEDÊ et al., 2009), para classificar a prevalência da esquistossomose em municípios do estado de Minas Gerais. Os dados de prevalência da esquistossomose usados foram os 197 municípios com informação positiva sobre a doença (usados na Seção 5.2.1). Neste trabalho, além das variáveis descritas na Seção 4.1, foram usadas variáveis de caracterização de vizinhança que medem a disparidade entre municípios vizinhos com relação às variáveis de renda, educação, saneamento, acesso à água, etc, descritas em Martins (2009).

Em Martins-Bedê et al. (2009), foi selecionado um conjunto menor de variáveis, de acordo com testes usando regressão linear múltipla (MARTINS, 2009). Os atributos selecionados foram aqueles que apresentaram alta correlação com a prevalência da

doença e baixa correlação com os outros atributos. Foram utilizadas duas abordagens principais, uma global e uma regional. Na abordagem global, um único modelo de regressão foi gerado e utilizado para estimar o risco de doença para todo o Estado. Já na abordagem regional, o Estado foi dividido em quatro regiões homogêneas e um modelo de regressão linear foi criado para cada uma delas.

O número de atributos utilizados nos experimentos variaram. Na abordagem global, foram utilizadas cinco variáveis. Na abordagem regional, foram utilizadas duas variáveis para a região R1 , 5 para a região R2, 4 para a região R3 e 3 para a região R4 (ver detalhes em (MARTINS, 2009). Em ambas as abordagens, aproximadamente 2/3 das amostras foram usadas como conjunto de treinamento, e o 1/3 restante como o conjunto de teste. O algoritmo SKATER (ASSUNÇÃO et al., 2006) foi utilizado para obter as regiões homogêneas no modelo regional. Este algoritmo cria áreas de tal forma que as áreas vizinhas com características semelhantes pertençam à mesma região.

Neste trabalho, foi usada uma família parametrizada de relações difusas para cada atributo a, definida por

$$S_{\lambda_a}(x, y) = \max(0, 1 - \frac{\mid x - y \mid}{\lambda_a \cdot l(a)})),$$

onde $\lambda_a > 0$ e $l(a)$ é o tamanho do domínio do atributo a. Nos experimentos, o valor do parâmetro λ para cada atributo na descrição de um caso foi calculado como uma porcentagem do seu intervalo de variação.

Denota-se a seguir uma relação de proximidade para um atributo de descrição do caso como V_{λ_a}, $\lambda_a \in]0, 1]$. A notação de um conjunto de tais relações pode ser sintetizada como $V_{(\lambda_1,...,\lambda_n)}$, o que significa que V_{λ_1} é aplicada ao atributo a_1, V_{λ_2} a a_2, e assim por diante.

A solução de um caso, descrita pela variável *classe*, foi definida no domínio $Cl = \{L, M, H\}$, como prevalência Baixa, Média e Alta, respectivamente. A relação de proximidade da variável *classe* é dada por T_λ, definida como $T_\lambda(w, w) = 1$ and $T_\lambda(w, w') = T\_\lambda(w', w)$, para todo $w, w' \in Cl$, $T_\lambda(H, M) = T_\lambda(M, L) = \lambda$ e $T_\lambda(H, L) = 0$, para todo $\lambda \in]0, 1]$.

A Tabela 7.1 traz os melhores resultados obtidos a partir de experimentos feitos com os dados, para modelos de regressão empregados em (MARTINS, 2009) e a Tabela 7.2,

pela abordagem de CBR difusa descrita acima. Na Tabela 7.2, ao lado da acurácia da abordagem CBR difusa, indicamos as relações de similaridade usadas para cada atributo.

Tabela 7.1 - Acurácia da classificação, para as abordagens global e regional, usando modelos de regressão.

Região	Modelo Regional	Modelo Global
R1	0,56	0,50
R2	0,51	0,40
R3	0,72	0,48
R4	0,76	0,59

Tabela 7.2 - Acurácia da classificação, para as abordagens global e regional, usando CBR difusa.

Região	Modelo Regional	Modelo Global
R1	0,56 $(V_{(0,3;0,4)}; T_{0,5})$	0,56 $(V_{(0,2;0,2)}; T_0)$
R2	0,56 $(V_{(0,2;0,2;0,2;0,2;0,2)}; T_{0,5})$	0,49 $(V_{(0,1;0,1;0,1;0,1;0,1)}; T_0)$
R3	0,62 $(V_{(0,2;0,4;0,3;0,3)}; T_0)$	0,71 $(V_{(0,2;0,2;0,2;0,2)}; T_0$
R4	0,38 $(V_{(0,2;0,2;0,2)}; T_{0,5})$	0,65 $(V_{(0,1;0,1;0,1)}; T_{0,5})$

A abordagem CBR difusa foi aplicada usando vários conjuntos de parâmetros para as relações de similaridade. O máximo e a média aritmética foram utilizados como para agregar as similaridades, tendo a média aritmética apresentado os melhores resultados, descritos na Tabela 7.2. Observe que para a região R1, obtivemos a mesma precisão (0,56), utilizando relações $(V_{(0,3,0,4)}, T_{0,5})$ e $(V_{(0,1,0,1)}, T_0)$ para a abordagem de aprendizagem regional.

Os resultados obtidos com a abordagem CBR difusa são comparáveis aos obtidos com os modelos de regressão e, num caso, a abordagem CBR difusa é melhor do que a regressão (região R_2). É interessante notar que na abordagem CBR difusa, ao contrário do que aconteceu com os modelos de regressão, a abordagem de aprendizagem global tem, muitas vezes, um desempenho melhor do que a regional. Além disso, a abordagem CBR difusa global obteve invariavelmente resultados melhores do que a regressão na abordagem global.

7.3 Classificação com relações difusas compatíveis com ordem

São apresentadas abaixo uma nova abordagem para classificação, que utiliza uma relação difusa criada a partir de uma partição difusa (SANDRI; MARTINS-BEDÊ, 2014).

7.3.1 Proposta de função de classificação difusa

As definições de $\text{OCFR}_{\preceq}$, $\text{CFP}_{\preceq}$, 2-Ruspini $\text{CFP}_{\preceq}$ e S^{+} apresentadas abaixo foram propostas originalmente em Sandri e Martins-Bedê (2014).

Uma relação difusa binária $S : \Omega^2 \to [0, 1]$ é dita ser *compatível com a ordem total* $(\Omega, \preceq)$ ($\text{OCFR}_{\preceq}$ ou OCFR), quando obedece as seguintes propriedades:

- $\forall x, y, z \in \Omega,\ S(x, x) = 1$ (*reflexidade*)
- $\forall x, y, z \in \Omega,\ S(x, y) = S(y, x)$ (*simetria*)
- $\forall x, y, z \in \Omega$, se $x \preceq y \preceq z$, então $S(x, z) \leq \min(S(x, y), S(y, z))$ (*compatibilidade com ordem total* $(\Omega, \preceq)$, ou $\preceq$-*compatibilidade*).

Seja $(\Omega, \preceq)$ uma ordem total e $\mathbf{A} = \{A_1, ..., A_n\}$ uma partição difusa (uma coleção de conjuntos difusos) $\mathbf{A}$ em Ω é dita ser uma *partição difusa convexa com respeito à ordem total* $(\Omega, \preceq)$ ($\text{CFP}_{\preceq}$ ou CFP), quando obedece as seguintes propriedades:

a) $\forall A_i \in \mathbf{A}, \exists x \in \Omega, A_i(x) = 1$ (*normalização*),

b) $\forall x, y, z \in \Omega, \forall A_i \in \mathbf{A}$, se $x \preceq y \preceq z$ então
$A_i(y) \geq \min(A_i(x), A_i(z))$ (*convexidade*),

c) $\forall x \in \Omega, \exists A_i \in \mathbf{A}, A_i(x) > 0$ (*cobertura de domínio*),

d) $\forall A_i, A_j \in \mathbf{A}, \text{se} i \neq j$ então $ncleo(A_i) \cap ncleo(A_j) = \emptyset$
(*interseção vazia entre núcleos*).

Em particular, uma CFP $\mathbf{A}$ é chamada de *2-Ruspini* $CFP_{\preceq}$ quando ela é aditiva ($\forall x \in \Omega,\ \sum_i \mu_{A_i}(x) = 1$) e cada elemento em X tem pertinência positiva a no máximo 2 conjuntos difusos em $\mathbf{A}$. Como consequência da definição, os conjuntos difusos numa 2-Ruspini CFP $\mathbf{A}$ podem somente ser conjuntos nítidos, trapezoidais ou triangulares.

Seja $\mathbf{A}$ uma CPF. Sejam as relações difusas $S^*(x, y)$ e $S_L(x, y)$:

$$\forall x, y \in X, S^*(x, y) = \sup_i \min(\mu_{A_i}(x), \mu_{A_i}(y))$$

$$\forall x, y \in X, S_L(x, y) = \inf_i \ \ 1- \mid \mu_{A_i}(x) - \mu_{A_i}(y) \mid$$

A relação S_+ abaixo é uma OCFR$_\preceq$ quando **A** é uma 2-Ruspini CFP$_\preceq$.

$$\forall x, y \in X, S^+(x, y) = \begin{cases} 0, \text{if } S^*(x, y) = 0 \\ S_L(x, y), \text{senão} \end{cases} \tag{7.3}$$

Seja **A** uma 2-Ruspini CFP. A função f^+ abaixo pode ser utilizada em tarefas de classificação por ser métrica (distância) quando todos os conjuntos que a compõem tem como núcleo um ponto (por exemplo, conjuntos difusos triangulares) e uma pseudométrica em caso contrário (p.ex. conjuntos nítidos ou trapezoidais) (SANDRI et al., 2014).

$$\forall x, y \in X, f^+(x, y) = 1 - S^+(x, y) \tag{7.4}$$

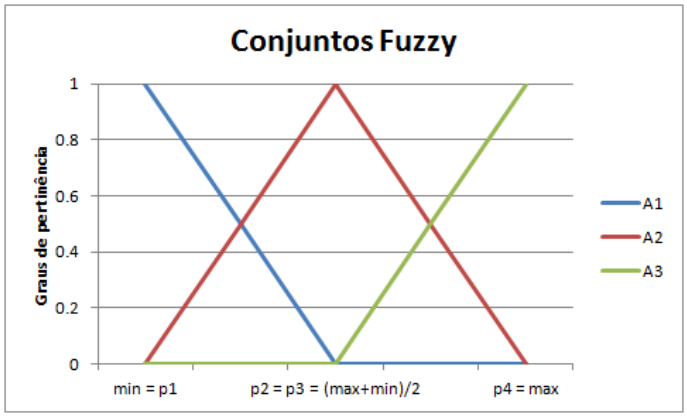
7.3.2 Uso da função f^+ no caso de estudo do Tapajós

Neste trabalho, foram geradas classificações para o k-NN e o ms-NN usando como distância a função f^+. Essa função foi calculada para os conjuntos difusos triangular e trapezoidal, apresentados na Figura **??**.

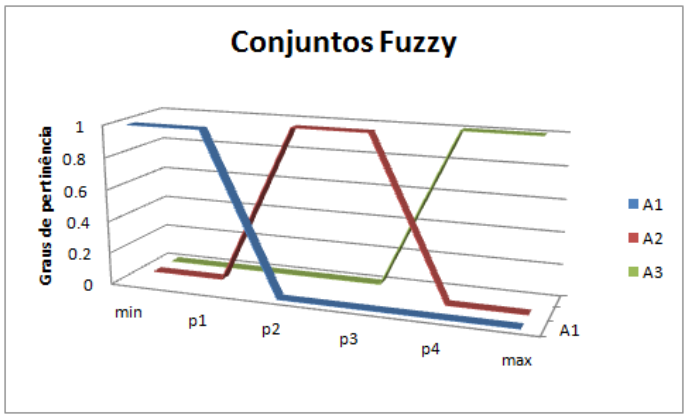
Os conjuntos difusos da Figura 7.1 foram gerados para cada atributo usado. Os valores limites (min e max da Figura 7.1) usados para gerar os conjuntos difusos foram os valores mínimo e máximo, somados 20%, no intervalo de variação de cada atributo . Nos conjuntos trapezoidal (Figura 7.1(b)) os valores p1 , p2, p3 e p4 são os percentis $0,2$, $0,4$, $0,6$ e $0,8$ do intervalo de valores de cada atributo. Nos conjuntos triangular (Figura7.1(a)) p1 = min, p2 = p3 = (min-max)/2 e p4 = max.

Nesta Seção foi adotada a notação d_Ftz quando foi usada a pseudométrica com conjuntos trapezoidal e d_Ftg quando foi usada a distância com conjuntos triangular. E também foram feitos testes usando duas abordagens, por pixel e por segmento.

Nas Figuras 7.2 e 7.3, são apresentadas, para cada abordagem de teste, as acurácias das classificações usando a função f^+ para o k-NN e para o ms-NN. Nessa Figura,



(a) Triangular



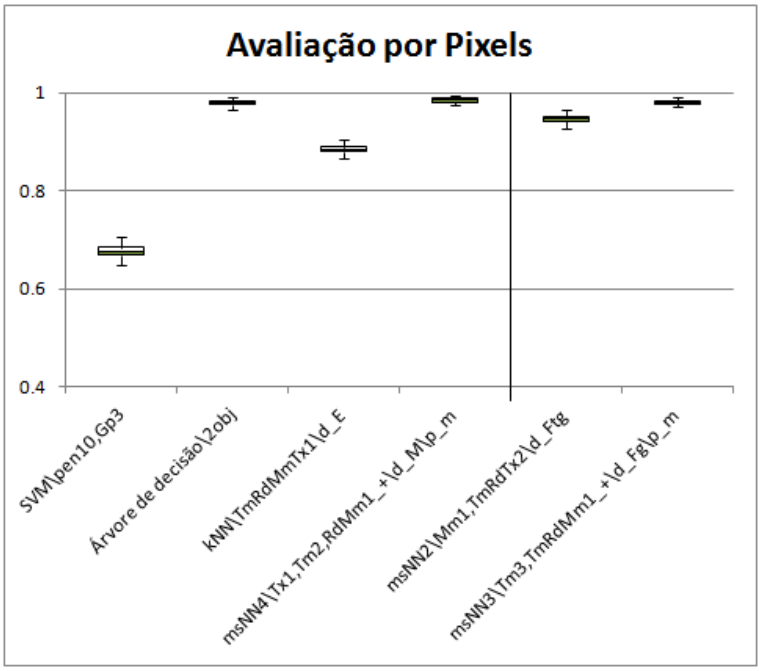
(b) Trapezoidal

Figura 7.1 - Gráfico com os conjuntos difuso (A1, A2 e A3).

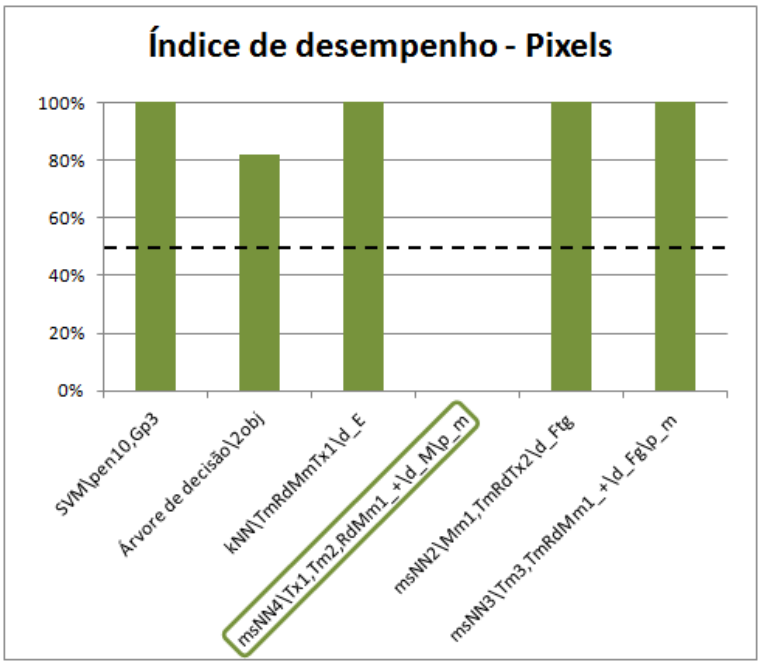
nota-se que a d_Ftg é mais acurada que a d_Ftz e que a abordagem por segmento é mais acurada que a por pixel.

Na abordagem por pixel, o melhor resultado é a classificação que usa 2 espaços ATR. Já na abordagem por segmento, o melhor resultado é a classificação que usa 3 espaços ATR. As médias dessas classificações são apresentadas na Tabela 7.3 juntamente com as médias das classificações do métodos da literatura e da melhor classificação do ms-NN apresentadas anteriormente na Tabela 6.6.

De acordo com as médias das acurácias das classificações (Tabela 7.3), as classificações do ms-NN+ usando conjuntos difusos são idênticas às classificações ms-NN+ apresentadas na Seção 6.

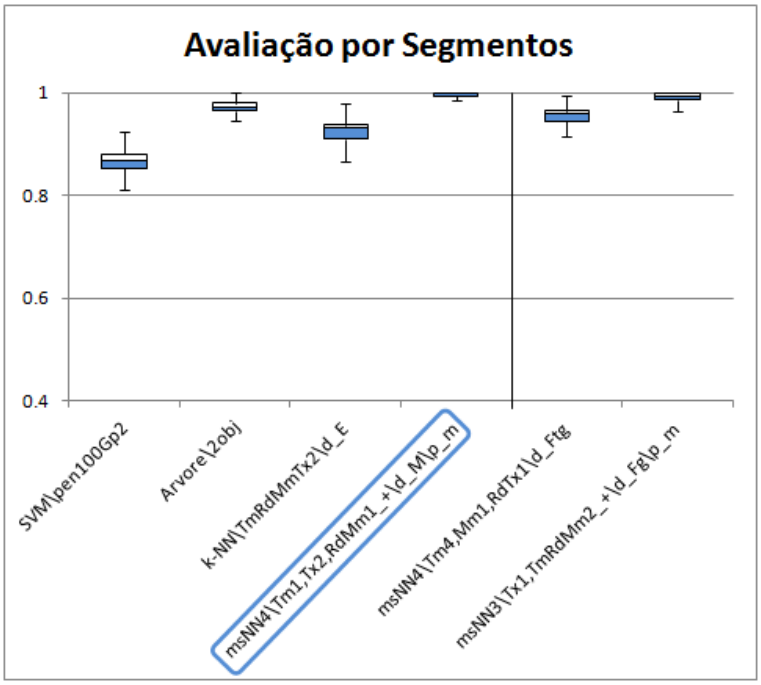


(a)

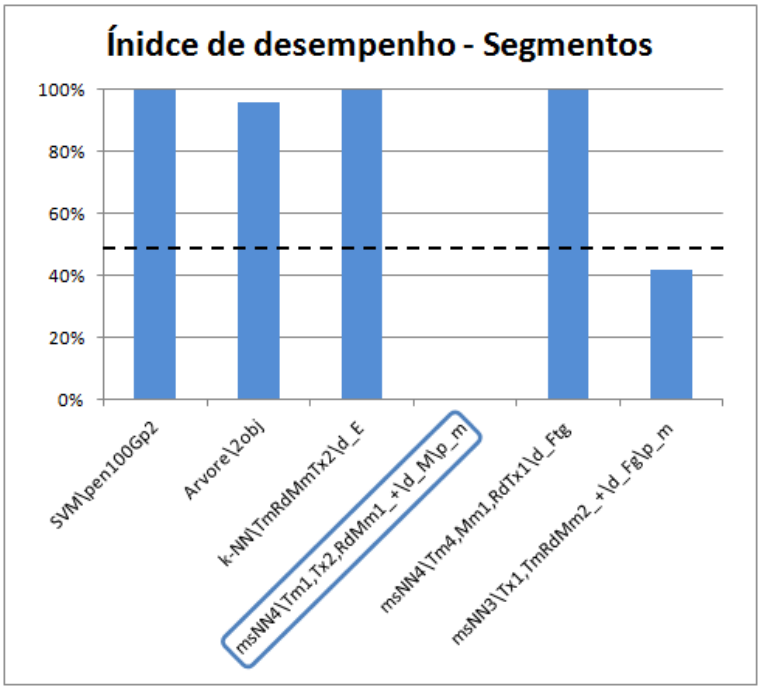


(b)

Figura 7.2 - Acurácias das classifações representantes (a) e índice de desempenho (b), para a abordagem de teste por Pixel



(a)



(b)

Figura 7.3 - Acurácias das classifações representantes (a) e índice de desempenho (b), para a abordagem de teste por Segmento

7.4 Conclusões

Neste capítulo, foi descrito o uso de relações difusas como base para classificação. As duas abordagens apresentadas, uma com relações difusas parametrizadas e a outra

Tabela 7.3 - Média das acurácias das classificações selecionadas como representantes usando todos os conjuntos de atributos e conjuntos difusos trapezoidal e triangular.

Método	Pixel		Segmento	
	média	desvio padrão	média	desvio padrão
SVM	$0,68$	$0,01$	$0,87$	$0,02$
Árvore de decisão	$0,97$	$0,01$	$0,97$	$0,01$
kNN	$0,88$	$0,01$	$0,93$	$0,02$
ms-NN4+\d_M\p_m	$0,98$	$< 0,01$	$0,99$	$0,01$
msNN3\d_Ftz\p_m	-	-	$0,94$	$0,01$
msNN3+\d_Ftg\p_m	$0,98$	$0,04$	$0,99$	$0,01$

com relações difusas obtidas a partir de partições difusas, são promissoras para o uso em ms-NN. A vantagem da primeira abordagem é a possibilidade de otimizar o parâmetro da relação e a da segunda, é a facilidade de se obter os conjuntos difusos com um especialista.

Em trabalhos futuros, pretende-se utilizar a primeira abordagem em ms-NN; em relação à segunda abordagem pretende-se testar o uso de agrupamento (clustering) para a obtenção da partição difusa. Aqui também, quando usado o ms-NN, a ponderação pode ocorrer entre espaços, como foi formalizado na Seção 3.2. A ponderação dentro do espaço pode ser feita uma ponderação por atributo, de tal forma que atributos mais importantes poderão ter peso maior. Além disso, pode-se usar a ponderação por atributo e para cada caso, como proposto em Armengol et al. (2005).

8 CONCLUSÕES E TRABALHOS FUTUROS

Neste trabalho, é proposto uma abordagem de cunho geral para classificação, chamado de ms-NN, que estende o tradicional método k-NN. A abordagem proposta utiliza vários espaços de atributos, ao invés de um único como o k-NN, sendo possível utilizar diferentes funções de distância e tipos de vizinhança em cada espaço. Os espaços podem ser de dois tipos: de atributos ou geográfico. As vizinhanças podem ser as usuais (k vizinhos mais próximos), assim como por contiguidade, útil em aplicações que envolvam segmentação, e por raio de influência, podendo também ser fixas ou variáveis. As distâncias podem ser a usuais, tais como Euclidiana e de Mahalanobis, como outras mais adequadas em uma dada aplicação. Além disso, outras funções de predominância podem ser utilizadas que simplesmente a de maioria simples.

Para validar a abordagem proposta, são apresentados aqui dois estudos de caso. Um deles em GeoSaúde, para classificação de prevalência de esquistossomose no Estado de Minas Gerais. Outro para classificação de padrões de uso e cobertura do solo, em uma área da região do Tapajós.

A complexidade do k-NN é quadrática em relação ao número de casos de treinamento, mas existem muitos estudos na literatura que visam melhorar esta complexidade. O ms-NN tem complexidade também quadrática, e os métodos para redução e complexidade criados para o k-NN podem ser trivialmente estendidos para o ms-NN.

Muitas das ferramentas oferecidas pelo arcabouço teórico não foram ainda aplicadas na prática. O uso de um tipo de distância que seja mais adequado para o tipo de dado e a flexibilidade de usar outro critério de predominância, são exemplos disso. O tipo de distância poderia ser escolhido de acordo com a estatística e origem do dado. As características das distâncias também podem ser usadas para auxiliar a escolha do tipo de distância para cada tipo de dado. Também de acordo com o tipo de dado, poderia ser dado um peso a cada espaço. Esse peso poderia ser dado pelo usuário ou poderia ser ajustado a partir de um pré treinamento usando amostras de validação.

A motivação para separar os dados em vários espaços é principalmente buscar evitar o fenômeno de Hughes. Outra justificativa de separar o dados multi-fonte em diferentes espaços tem relação com a questão física do problema e não com a probabilidade estatística dos dados. Cada espaço tem uma coerência da física dos dados, que são relevantes tanto em GeoSaúde quanto em LUCC. Em ambos os casos, tem-se dados

multi-fonte que podem variar no tempo. Então, este estudo também é um subsídio no ponto de vista prático, já que pode-se usar dados multi-fonte separadamente em um mesmo classificador. Outra vantagem do ms-NN é que o resultado da classificação pode ser melhorado à medida que novos dados são disponibilizados.

Neste trabalho, em cada classificação foi usada uma mesma função de distância em todos os espaços do tipo ATR. Quando foram utilizados somente espaços ATR, foi aplicado o critério de predominância maioria. Quando foram utilizados espaços ATR aliado a espaço GEO com distância por contiguidade, foi aplicada a a ponderação pelo número de vizinhos de um mesmo espaço (i.e., todos os espaços possuem o mesmo peso, que é dividido entre o número de vizinhos do espaço). Como trabalho futuro, pretende-se usar dados multi-fonte em espaços com distâncias distintas e associar diferentes tipos de ponderação.

Neste trabalho foi feito um teste exaustivo para chegar ao melhor resultado do ms-NN. Tendo em vista que não é em toda aplicação que se pode testar todos os parâmetros, pretende-se, como trabalho futuro, usar algoritmos de aprendizado para encontrar tais parâmetros.

Na aplicação Schisto, foram feitas duas análises, a nível municipal e local. Nesta aplicação, foram usados dois tipos de espaço, ATR e GEO. No espaço ATR foi usada vizinhança fixa com as distâncias Euclidiana e de Mahalanobis. No espaço GEO, foram usados dois tipos de vizinhança: i) fixa, quando se usa distância Euclidiana (em nível local) e ii) variável, quando se usa distância por contiguidade (em nível municipal). Em trabalhos futuros pretende-se verificar o impacto de usar as classes indene e baixa como sendo uma única classe.

Na aplicação Tapajós, além das distâncias Euclidiana e de Mahalanobis, foram usadas distâncias baseadas em relações difusas, que demonstraram ser promissoras para o ms-NN. A inclusão do espaço GEO tornou o ms-NN mais eficaz que a árvore de decisão. Em trabalhos futuros pretende-se utilizar relações difusas parametrizadas que já demonstraram resultados satisfatórios para uma aplicação com os dados do Schisto. Além disso, pretende-se aplicar os dados de Tapajós usando pixels como representação geométrica no ms-NN.

De modo geral, para as aplicações usadas neste trabalho, o método ms-NN proposto obteve melhores resultados quando comparados ao k-NN. O ms-NN também se mostrou eficaz quando comparados com SVM e árvore de decisão para o estudo de caso Schisto, principalmente ao adicionar o espaço geográfico na classificação.

Os parâmetros do ms-NN foram testados exaustivamente. Como o número de classificações é muito grande, em geral optou-se por fazer todos os experimentos usando apenas um conjunto de treinamento e teste. Como trabalho futuro pretende-se aplicar n-folders para todos os experimentos.

Os trabalhos da tese devem continuar para ambas as aplicações, utilizando diferentes combinações de espaços, tipos de distâncias e vizinhança para aumentar a eficiência do método. Além disso, outras distâncias deverão ser adicionadas ao protótipo.

REFERÊNCIAS BIBLIOGRÁFICAS

AGUIAR, A.; CÂMARA, G.; CARTAXO, R. Modeling spatial relations by generalized proximity matrices. In: BRAZILIAN SYMPOSIUM ON GEOINFORMATICS (GEOINFO2003), 5., 2003, Campos do Jordão, SP, Brazil. **Proceedings...** São José dos Campos: INPE, 2003. 1

AMARAL, R. S.; TAUTIL, P. L.; LIMA, D. D.; ENGELS, D. An analysis of the impact of the schistosomiasis control programme in Brazil. **Mem Inst Oswaldo Cruz**, v. 101, p. 79–85, 2006. 22

ARMENGOL, E.; ESTEVA, F.; GODO, L.; TORRA, V. On learning similarity relations in fuzzy case-based reasoning. In: **Lecture Notes in Computer Science**. [S.l.]: Springer Berlin Heidelberg, 2005. v. 3135, p. 14–32. ISBN 978-3-540-23990-1. 75, 83

ASSUNÇÃO, R. M.; NEVES, M. C.; CÂMARA, G.; FREITAS, C. C. Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees,. **International Journal of Geographical Information Science**, v. 20, p. 797–811, 2006. 76

BAO, Y.; ISHII, N.; DU, X. Improved heterogeneous distance functions. **Journal of Artificial Intelligence Research**, v. 6, n. 1, p. 1 – 34, 1997. 3

_____. Combining multiple k-nearest neighbor classifiers using different distance functions. In: YANG, Z.; YIN, H.; EVERSON, R. (Ed.). **Intelligent Data Engineering and Automated Learning**. Springer, 2004. (Lecture Notes in Computer Science, v. 3177), p. 634–641. ISBN 978-3-540-22881-3. Disponível em: <http://dx.doi.org/10.1007/978-3-540-28651-6_93>. 3

BAY, S. D. Nearest neighbor classification from multiple feature subsets. **Intelligent Data Analysis**, v. 3, n. 3, p. 191 – 209, 1999. ISSN 1088-467X. 2

BISHOP, C. M. **Pattern recognition and machine learning**. New York, NY: Springer, 2006. (Information Science and Statistics). ISBN 0387310732. 7, 8, 9, 11, 29, 35

CÂMARA, G.; SOUZA, R. C. M.; FREITAS, U. M.; GARRIDO, J. Spring: Integrating remote sensing and GIS by object oriented data modelling. **Computers & Graphics**, v. 20, n. 3, p. 395–403, 1996. 25

CÂMARA, G.; VINHAS, L.; FERREIRA, K.; QUEIROZ, G.; SOUZA, R.; MONTEIRO, A.; CARVALHO, M.; CASANOVA, M.; FREITAS, U. Terralib: An open source gis library for large-scale environmental and socio-economic applications. In: HALL, G.; LEAHY, M. (Ed.). **Open source approaches in spatial data handling**. Springer Berlin Heidelberg, 2008, (Advances in Geographic Information Science, v. 2). p. 247–270. ISBN 978-3-540-74830-4. Disponível em: <http://dx.doi.org/10.1007/978-3-540-74831-1_12>. 95

CARVALHO, O. d. S.; ROCHA, R. S.; MASSARA, C. L.; KATZ, N. Expansão da esquistossomose mansoni em Minas Gerais. **Memórias do Instituto Oswaldo Cruz**, scielo, v. 82, p. 295 – 298, 00 1987. ISSN 0074-0276. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0074-02761987000800056&nrm=iso>. 22

CARVALHO, O. S.; DUTRA, L. V.; MOURA, A. C. M.; FREITAS, C. d. C.; AMARAL, R. S.; DRUMMOND, S. C.; FREITAS, C. R.; SCHOLTE, R. G. C.; GUIMARÃES, R. J. d. P. Souza e; MELO, G. d. R.; CORREIA, V. R. d. M.; GUERRA, M. Desenvolvimento de um sistema de informações para o estudo, planejamento e controle da esquistossomose no estado de minas gerais. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 12. (SBSR), 16-21 abr. 2005, Goiânia. **Anais...** São José dos Campos: INPE, 2005. p. 2083–2086. ISBN 85-17-00018-8. 22

CASANOVA, M. A.; CÂMARA, G.; DAVIS, C. A.; VINHAS, L.; QUEIROZ, G. R. **Bancos de dados geograficos**. 1. ed. Curitiba,PR: Mundogeo, 2005. 506 p. 12, 13

CRAMMER, K.; SINGER, Y. On the algorithmic implementation of multi-class kernel-based vector machines. **Journal of Machine Learning Research**, p. 265–292, 2001. 28

DUBOIS, D.; PRADE, H. **Possibility Theory**: Qualitative and quantitative aspects. [S.l.]: Plenum Press, New-York, 1988. 169226 p. 73

ESCADA, M. I. S.; AMARAL, S.; RENNÓ, C. D.; PINHEIRO, T. F. **Levantamento do uso e cobertura da terra e da rede de infra-estrutura no distrito florestal da br-163**. 2009. 52 p. Disponível em: <http://urlib.net/8JMKD3MGP8W/357DD7L>. Acesso em: 22 jan. 2013. 25

FIX, E.; HODGES, J. **Discriminatory analysis, nonparametric discrimination: Consistency properties**. Randolph Field, Texas: [s.n.], 1951. 1

FONSECA, J. M. M. R. D. **Indução de árvores de decisão**. 151 p. Dissertação (Mestrado) — Universidade Nova de Lisboa, Lisboa, 1994. 6

GUIMARÃES, P. R. B. **Métodos quantitativos estatísticos**. Paris: IESDE Brasil S.A, 2008. 245 p. 30

GUIMARÃES, R. J. P. S. **Ferramentas de geoprocessamento para o estudo e controle da esquistossomose no Estado de Minas Gerais**. 172 p. Tese (Doutorado em Biomedicina) — Santa Casa de Belo Horizonte, São José dos Campos, 2010. 22, 24, 61

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. The weka data mining software: An update. **SIGKDD Explorations**, v. 11, p. 338–353, 2009. 29

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). **Canais**: Estados. 2013. Disponível em: <http://www.ibge.gov.br/>. Acesso em: 07 dez. 2013. 21

INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS / DIVISÃO DE PROCESSAMENTO DE IMAGENS. **TerraView**. São José dos Campos - SP, Brasil, 2010. Disponível em: <http://www.dpi.inpe.br/terraview/>. Acesso em: nov. 2013. 95

_____. **Tutorial do TerraView**. São José dos Campos - SP, Brasil, 2010. Disponível em: <http://www.dpi.inpe.br/terraview/php/docs.php?body=Tutorial_i>. Acesso em: nov. 2013. 95, 97

KORTING, T. S. **GeoDMA: a toolbox integrating data mining with object-based and multi-temporal analysis of satellite remotely sensed imagery**. 119 p. Tese (Doutorado) — Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, 2012-08-20 2012. Disponível em: <http://urlib.net/sid.inpe.br/mtc-m19/2012/07.31.18.22>. Acesso em: 02 dez. 2013. 27

MARTINS-BEDÊ, F. d. T.; GODO, L.; SANDRI, S. A.; DUTRA, L. V.; FREITAS, C. d. C.; CARVALHO, O. S.; GUIMARÃES, R. J. P. S.; AMARAL, R. S. Classification of schistosomiasis prevalence using fuzzy case-based reasoning. **Lecture Notes in Computer Science**, v. 5517, p. 1053–1060, 2009. Disponível em: <http://urlib.net/sid.inpe.br/mtc-m1980/2009/12.14.16.52>. 75

MARTINS, F. d. T. **Mapeamento do risco da esquistossomose no Estado de Minas Gerais, usando dados ambientais e sociais**. 144 p. Dissertação (Mestrado) — Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2008-02-26 2009. Disponível em: <http://urlib.net/sid.inpe.br/mtc-m17{\spacefactor\@m}80/2008/02.07.13.17>. Acesso em: 03 dez. 2013. 23, 75, 76

MEDEIROS, I. P. d.; FILHO, C. A. P. d. C.; ERTHAL, G. J.; DUTRA, L. V. Classificação de imagens pelo método de árvore de decisão oblíqua. In: EPIPHANIO, J. C. N.; GALVÃO, L. S. (Ed.). **Anais...** São José dos Campos: Instituto Nacional de Pesquisas Espaciais (INPE), 2011. p. 4255–4262. ISBN 978-85-17-00056-0 (Internet) and 978-85-17-00057-7 (DVD). Disponível em: <http://urlib.net/dpi.inpe.br/marte/2011/06.27.14.17>. 6

MICHIE, D.; SPIEGELHALTER, D. J. **Machine Learning, Neural and Statistical Classification**. [S.l.]: Prentice Hall, 1994. 289 p. 7

MITCHELL, M. **An Introduction to Genetic Algorithms**. Cambridge, MA: MIT Press, 1996. 221 p. ISBN 9780262631853. 3

PEREIRA, L. d. O. **Avaliação de métodos de integração de imagens ópticas e de Radar para a classificação do uso e cobertura da terra na Região Amazônica**. 270 p. Dissertação (Mestrado) — Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, 2012-08-27 2012. Disponível em: <http://urlib.net/sid.inpe.br/mtc-m19/2012/08.30.12.50>. Acesso em: 06 nov. 2012. 26

PONZONI, F. J.; SHIMABUKURO, Y. E. **Sensoriamento remoto no estudo da vegetação**. 1. ed. São José dos Campos, SP: Parêntese, 2009. 127 p. ISBN 9788560507023. 25

REIS, M. S. **Geração das amostras, processamento das Imagens ALOS e TM, segmentação e geração do MLME (comunicação pessoal)**. 2013. 25

SANDRI, S.; MARTINS-BEDÊ, F.; DUTRA, L. Using a fuzzy based pseudometric in classification. aceito. In: INTERNATIONAL CONFERENCE ON INFORMATION PROCESSING AND MANAGEMENT OF UNCERTAINTY IN KNOWLEDGE-BASED SYSTEMS (IPMU 2014), 15., Montpellier,França. **Proceedings...** [S.l.], 2014. 1, 79

SANDRI, S.; MARTINS-BEDÊ, F. T. A method for deriving order compatible fuzzy relations from convex fuzzy partitions. **Fuzzy Sets and Systems**, 2014.

ISSN 0165-0114. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0165011413004855>. 1, 2, 78

SHRIVASTAVA, S. K.; MEWADA, P. Aco based feature subset selection for multiple k-nearest neighbor classifiers. **International Journal on Computer Science and Engineering**, v. 3, n. 5, p. 1831 – 1838, 2011. 2, 3

VAPNIK, V.; GOLOWICH, S. E.; SMOLA, A. Support vector method for function approximation, regression estimation, and signal processing. In: **Advances in neural information processing systems 9**. [S.l.]: MIT Press, 1996. p. 281–287. 5

WANG, L.-J.; WANG, X.-L.; CHEN, Q.-C. Ga-based feature subset clustering for combination of multiple nearest neighbors classifiers. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING AND CYBERNETICS. **Proceedings...** [S.l.], 2005. v. 5, p. 2982–2987. 2, 3

WEBB, A. R. **Statistical pattern recognition**. 1. ed. Chichester: John Wiley & Sons, 2002. 1, 7, 8, 9, 11, 29, 35

WORLD HEALTH ORGANIZATION (WHO). **The control of schistosomiasis**. 1985. 113 p. 22

YAMADA, T.; YAMASHITA, K.; ISHII, N.; IWATA, K. Text classification by combining different distance functions with weights. In: INTERNATIONAL CONFERENCE ON SOFTWARE ENGINEERING, ARTIFICIAL INTELLIGENCE, NETWORKING, AND PARALLEL/DISTRIBUTED COMPUTING, 17., 2006. **Proceedings...** Madison, WI: IEEE, 2006. 2

ZADEH, L. Fuzzy sets. **Information and Control**, v. 8, p. 338–353, 1965. 73

APÊNDICE A - PLUG-IN IMPLEMENTADO

Este Apêndice apresenta a descrição das funcionalidades do software implementado, sua interface e uma breve explicação de como deve ser usado.

O Algoritmo 3.2 foi implementado na linguagem C++, usando a biblioteca TerraLib (CÂMARA et al., 2008), como um plug-in do aplicativo TerraView, denominado *ms-NN classification*. Esta biblioteca, de código aberto, facilita a manipulação de dados espaço-temporais e dá suporte ao desenvolvimento de aplicações que usam bancos de dados geográficos. TerraView (INPE/DPI, 2010a) é um aplicativo construído para exemplificar o uso da biblioteca TerraLib, que possui uma interface para consulta e manipulação dos dados geográficos e permite a adição de módulos de extensão (plug-in).

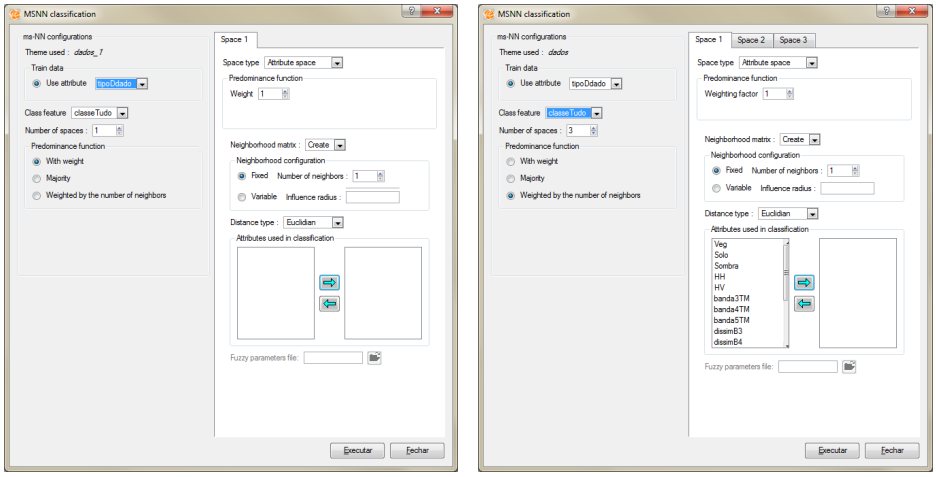
Para usar este plug-in [1], deve ser criado um banco de dados (BD) no TerraView (mais informações sobre como usar o TerraView pode ser obtida em INPE/DPI (2010b)). O BD deve conter todos os atributos que se deseja usar, um campo com os rótulos originais e um campo que contenha a informação sobre o tipo de dado. Os atributos podem ser do tipo inteiro, texto ou real. O campo com o rótulo deve ser do tipo texto. O tipo de dado define quais são os casos de treinamento e teste. O campo tipo de dado deve ser do tipo inteiro. Os casos com valores 1 serão os casos de treinamento e com 2, os de teste.

A Figura A.1(a) apresenta a tela principal do plug-in implementado. Em *Train data* deve ser selecionado o campo da tabela do BD que contém a informação do tipo de atributo. No campo *Class feature* deve ser selecionado o campo do BD que contém os rótulos para classificação. Em *Predominance function* deve ser selecionada a função de predominância que se deseja usar (ver Seção 3.3) dentre as opções:

- *With weight*: habilita o campo *Predominance function - Weight* do lado direito da tela, dentro da(s) aba(s) *Space*. Nessa opção pode ser selecionado um valor de peso para cada espaço e todos os casos vizinhos dentro do mesmo espaço terão esse peso. Neste contexto, os votos dos casos em diferentes espaços poderão ter pesos diferenciados no cômputo final.
- *Majority*: desabilita o campo *Predominance function* do lado direito da tela, dentro da(s) aba(s) *Space*. Com essa opção será usado o voto da maioria dos casos vizinhos para classificar o caso.

[1] O plug-in é disponibilizado para usuários externos ao INPE, mediante solicitação

- *Weighted by the number of neighbors*: habilita o campo *Predominance function - Weighting factor* do lado direito da tela, dentro da(s) aba(s) *Space*, conforme Figura A.1(b). Nesta opção, o peso de cada caso vizinho em um mesmo espaço será o valor do campo *Weighting factor* dividido pelo número de vizinhos do espaço.



(a) (b)

Figura A.1 - Telas do plug-in ms-NN classification: tela principal e escolha dos pesos

Em *Number of spaces* deve ser colocado o número de espaços que se deseja usar. Neste campo pode ser usado qualquer valor maior ou igual a 1. Ao inserir um valor maior que 1 neste campo, será apresentado ao lado direito da tela o mesmo número de abas. Essas abas representam os espaços, conforme Figura A.1(b). Em cada aba, deve-se escolher um tipo de espaço, conforme Figura entre as opções *Attribute space* e *Geographic space*. Também em cada aba deve-se escolher a configuração (tipo) de vizinhança em *Neighborhood configuration*. A opções são *Fixed* e *Variable*. Na primeira deve ser colocado o número de vizinhos a ser considerado no campo *Number of neighbors*. A configuração de vizinhança está disponível para ambos os tipos de espaço.

Para a opção *Attribute space* estão implementados 4 tipos de distância (Euclidiana, Mahalanobis, hamming e fuzzy) para o espaço ATR conforme Figura A.3(a). A distância de hamming pode ser usada com atributos do tipo inteiro ou texto. Já as distâncias Euclidiana, Mahalanobis e fuzzy, podem ser usadas com atributos do tipo real. No caso do uso da distância fuzzy, deve-se usar um arquivo com os parâmetros do conjunto fuzzy. Esse arquivo deve conter 6 parametros de fuzzyficação, conforme

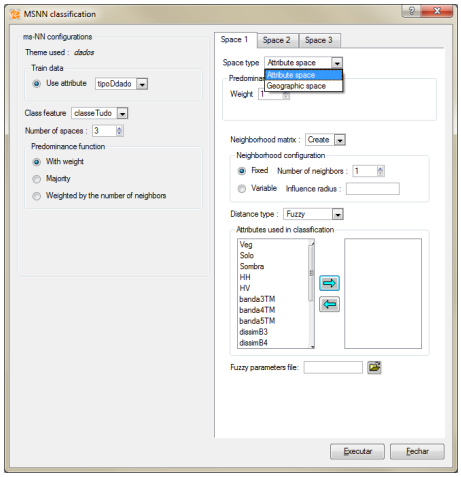
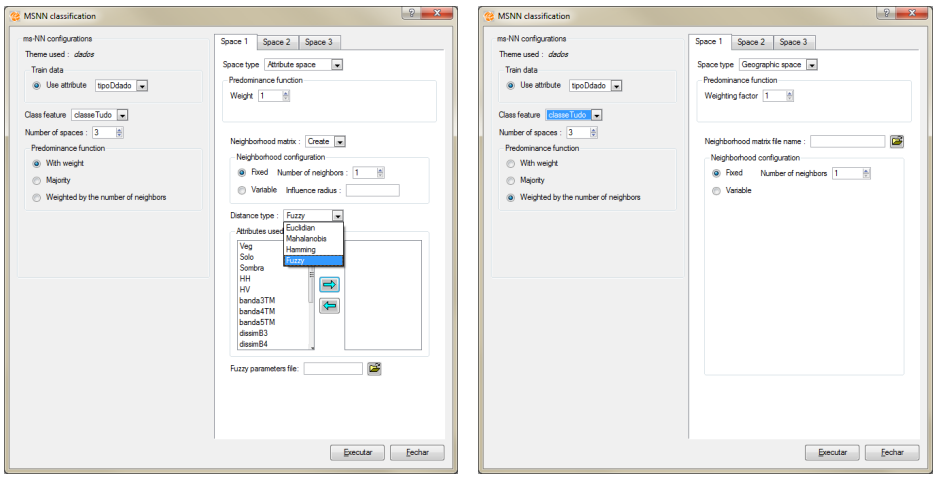


Figura A.2 - Opção para escolha do tipo de espaço.

Figura 7.1 (ver Seção 7).



(a) (b)

Figura A.3 - Telas para configurar o tipo de distância em cada espaço, seja de atributos ou geográfico.

Para usar o espaço geográfico, primeiramente deve-se obter, no próprio TerraView, uma matriz de proximidade. Em INPE/DPI (2010b) há uma explicação passo a passo de como criar essa matriz. A matriz de proximidade é usada como matriz de vizinhança para o espaço geográfico. Essa matriz deve ser aberta em *Neighborhood matrix file name*, como mostra a Figura A.3(b).

Após a execução do plug-in, uma nova coluna será adicionada ao banco de dados

com a classificação atribuída a cada polígono pelo algoritmo.

APÊNDICE B - GRÁFICOS SCHISTO

Neste Apêndice são apresentados os gráficos para o estudo de caso Schisto por município com:

- as classificações resultantes dos métodos SVM e árvore de decisão;
- as classificações pre-selecionadas dos métodos k-NN e ms-NN para todas as configurações de conjuntos de atributos.

Cada gráfico possui uma ou mais classificações destacadas. As classificações destacadas são as classificações selecionadas como representantes do método.

Nos gráficos com as classificações resultantes dos métodos SVM e árvore de decisão estão destacadas as classificações que foram selecionadas como representante de cada método. Os gráficos apresentam as médias de acurácia para cada configuração do classificador usada. No eixo horizontal dos gráficos do SVM estão as configurações dos parâmetros usados neste método: a penalidade (Pen) e o grau do polinômio (Gp). Para a árvore de decisão no eixo horizontal estão o número de objeto mínimo por folha, que foi usado como parâmetro de poda.

As classificações com as acurácias do SVM e árvore de decisão para o estudo de caso Schisto a nível municipal estão apresentadas nas Figuras B.1, B.7 e B.13.

As classificações pre-selecionadas usando o mesmo número de combinações de conjuntos de atributos estão separadas por blocos nos gráficos do k-NN e do ms-NN. Nos gráficos do k-NN, o eixo horizontal contém a mesma configuração de parâmetros do texto.

Os gráficos com as classificações pre-selecionadas usando o k-NN são apresentados nas Figuras B.2, B.8 e B.14.

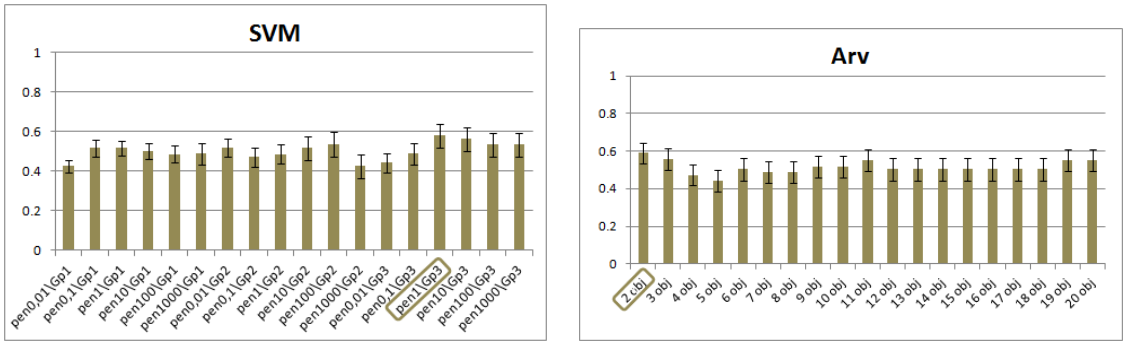
Os gráficos do ms-NN a nível municipal para os três conjuntos de dados estão listados a seguir:

- para a abordagem com 3 classes: Baixa, Média e Alta (sem indene):
  - Figura B.3 para 2 espaços;
  - Figura B.4 para 3 espaços;
  - Figura B.5 para 4 espaços;

  - Figura B.6 para 5 espaços.

- para a abordagem com 4 classes: Baixa, Média, Alta e 25% dos casos classe Indene (com 25% indene):
  - Figura B.9 para 2 espaços ATR,
  - Figura B.10 para 3 espaços;
  - Figura B.11 para 4 espaços;
  - Figura B.12 para 5 espaços.
- para a abordagem com 4 classes: Indene, Baixa, Média e Alta (com indene):
  - Figura B.15 para 2 espaços;
  - Figura B.16 para 3 espaços;
  - Figura B.17 para 4 espaços;
  - Figura B.18 para 5 espaços.

Nas Figuras com as classificações do ms-NN tem-se gráficos para a distância euclidiana (d_E) e para a distância de mahalanobis (d_M) usando as funções de pertinências maioria (p_m) e ponderada pelo número de vizinhos de um mesmo espaço (p_w)



(a) (b)

Figura B.1 - Média das acurácias para o método SVM e Árvore de decisão para abordagem com 3 classes.

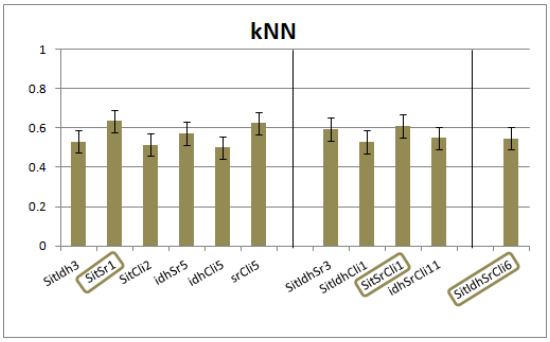


Figura B.2 - Média das acurácias para o método k-NN para abordagem com 3 classes

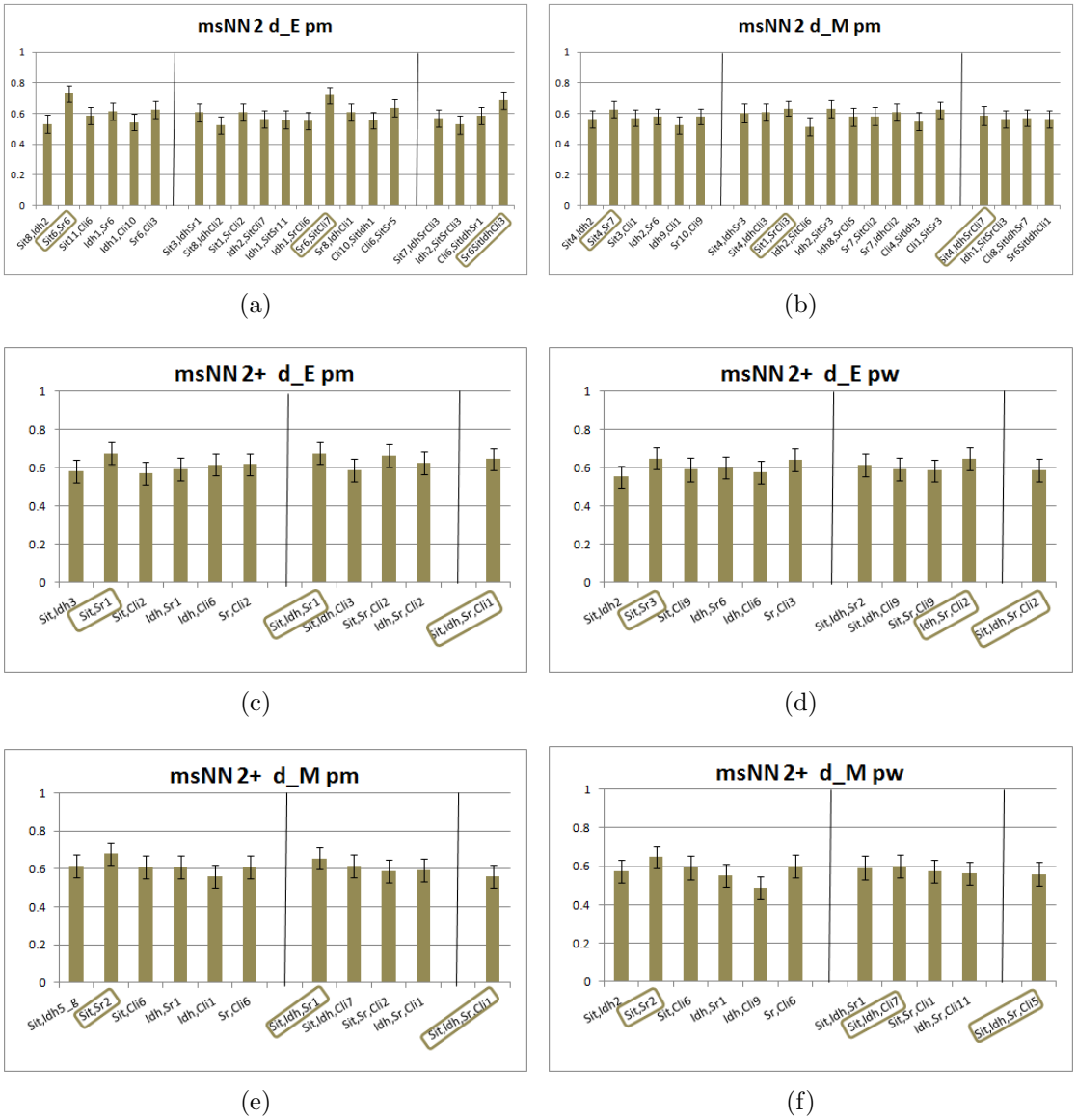
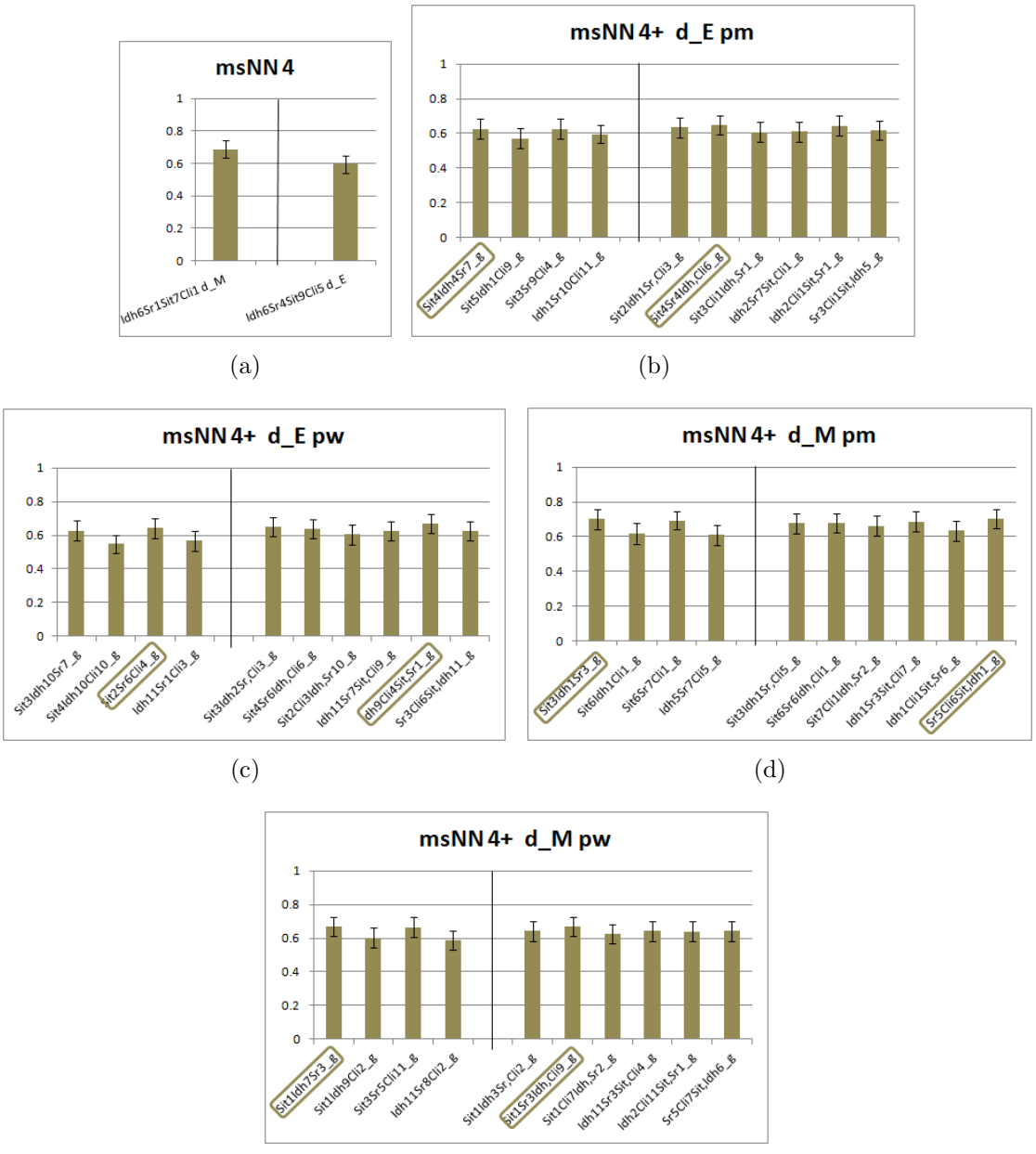


Figura B.3 - Média das acurácias para o método ms-NN para 2 espaços: 2 ATR (a) e (b); 1 ATR e 1 GEO (c), (d), (e) e (f) para abordagem com 3 classes.

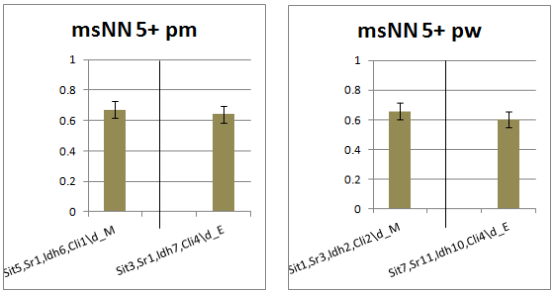
(a) (b) (c) (d) (e) (f)

Figura B.4 - Média das acurácias para o método ms-NN para 3 espaços: 3 ATR (a) e (b); 2 ATR e 1 GEO (c), (d), (e) e (f) para abordagem com 3 classes.



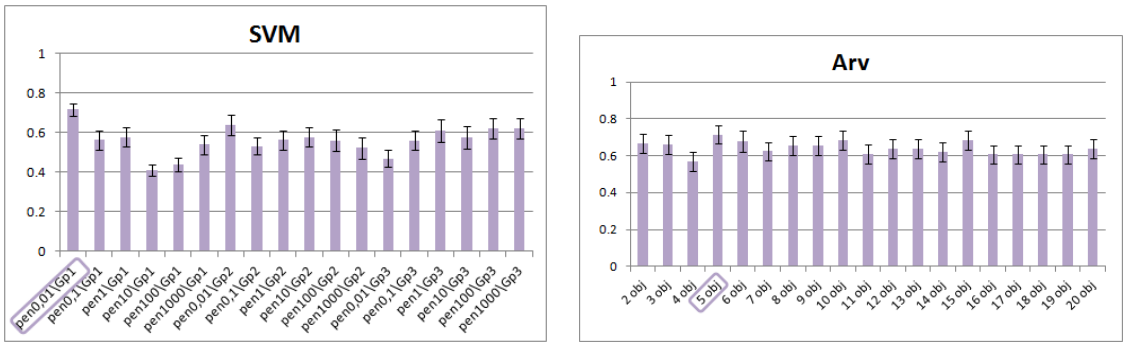
(a) (b) (c) (d) (e)

Figura B.5 - Média das acurácias para o método ms-NN para 4 espaços: 4 ATR (a); 3 ATR e 1 GEO (b), (c) e (d) para abordagem com 3 classes.



(a) (b)

Figura B.6 - Média das acurácias para o método ms-NN para 5 espaços, 4 ATR e 1 GEO para abordagem com 3 classes.



(a) (b)

Figura B.7 - Média das acurácias para o método SVM e Árvore de decisão para abordagem com 4 classes, usando 25% das amostras indenes.

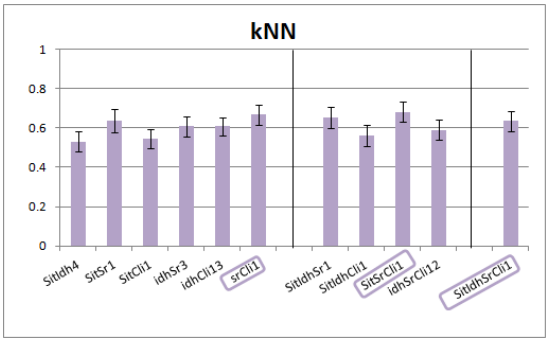
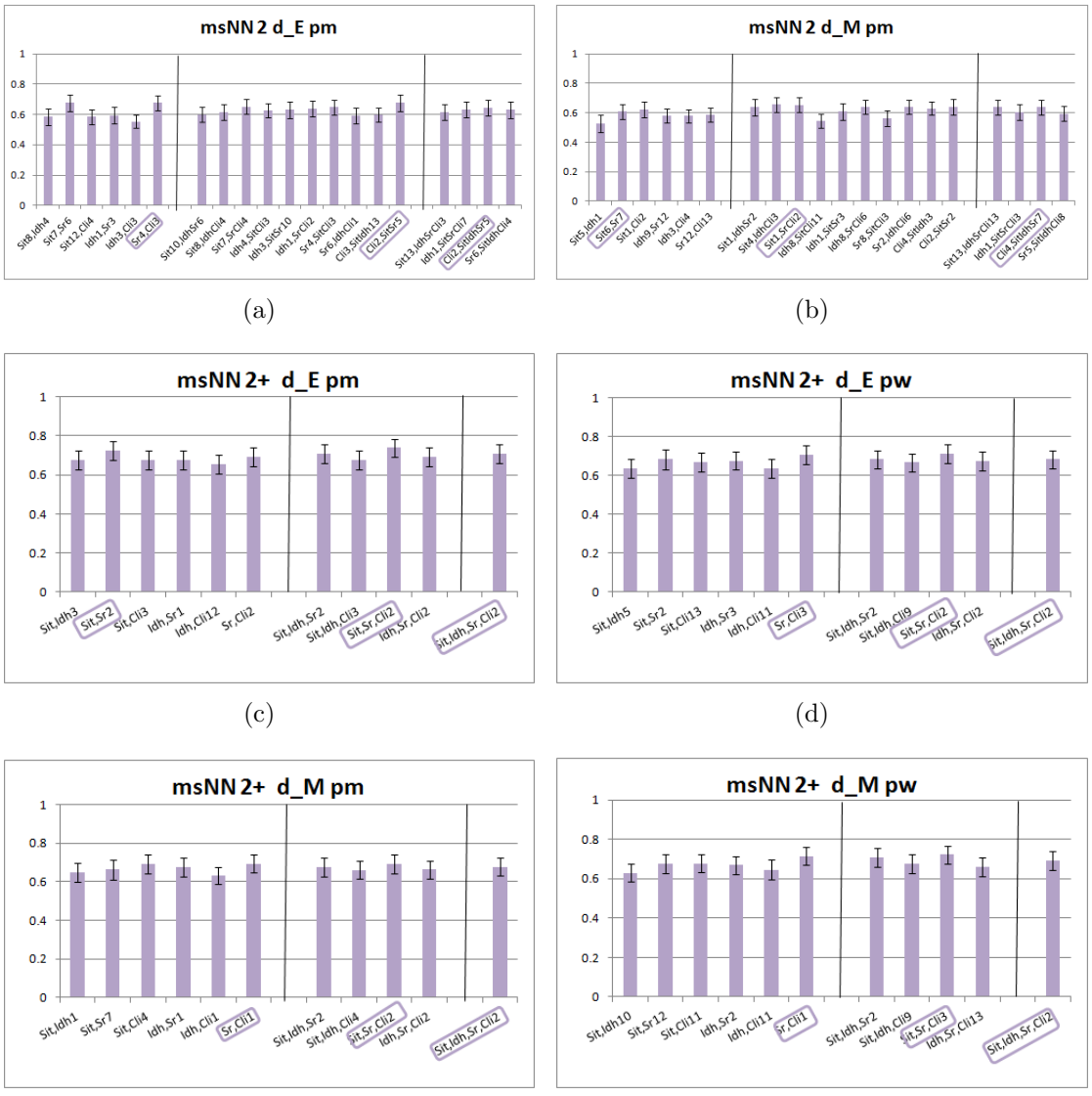


Figura B.8 - Média das acurácias para o método k-NN para abordagem com 4 classes, usando 25% das amostras indenes.



(a) (b) (c) (d) (e) (f)

Figura B.9 - Média das acurácias para o método ms-NN para 2 espaços: 2 ATR (a) e (b); 1 ATR e 1 GEO (c), (d), (e) e (f) para abordagem com 4 classes, usando 25% das amostras indenes.

(a) (b) (c) (d) (e) (f)

Figura B.10 - Média das acurácias para o método ms-NN para 3 espaços: 3 ATR (a) e (b); 2 ATR e 1 GEO (c), (d), (e) e (f) para abordagem com 4 classes, usando 25% das amostras indenes.

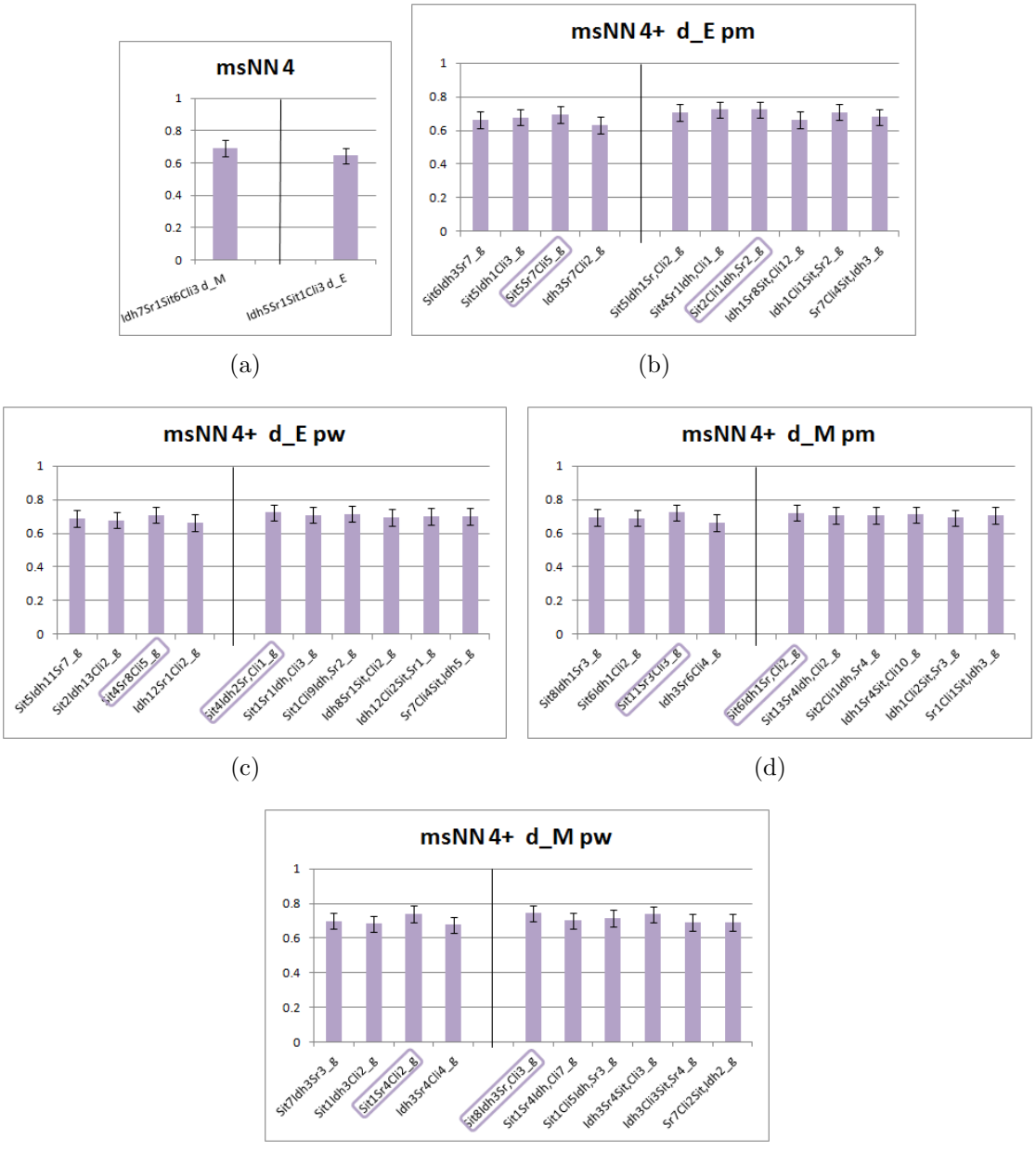
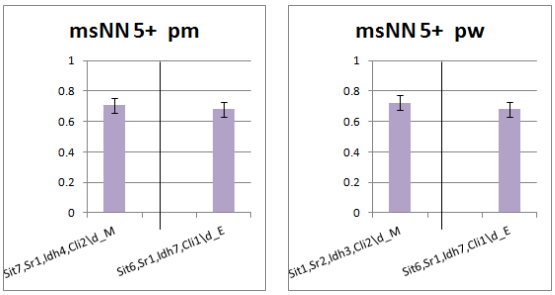
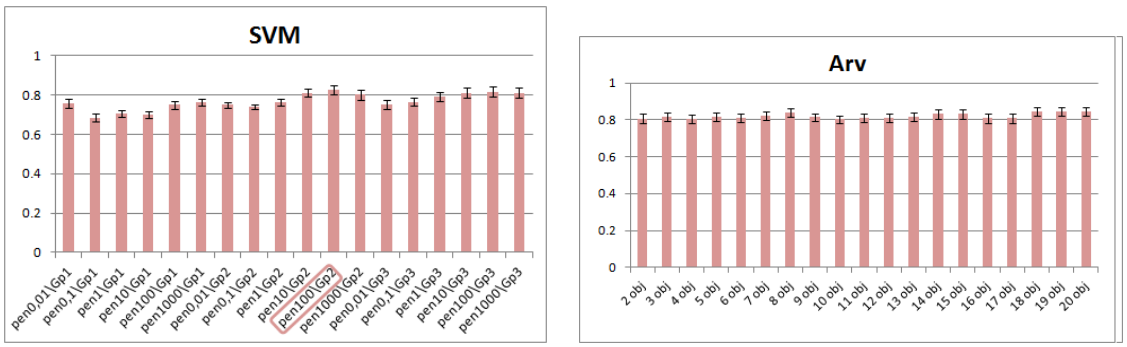


Figura B.11 - Média das acurácias para o método ms-NN para 4 espaços: 4 ATR (a); 3 ATR e 1 GEO (b), (c) e (d) para abordagem com 4 classes, usando 25% das amostras indenes.



(a) (b)

Figura B.12 - Média das acurácias para o método ms-NN para 5 espaços, 4 ATR e 1 GEO para abordagem com 4 classes, usando 25% das amostras indenes.



(a) (b)

Figura B.13 - Média das acurácias para o método SVM e Árvore de decisão para abordagem com 4 classes, usando todas as amostras indenes.

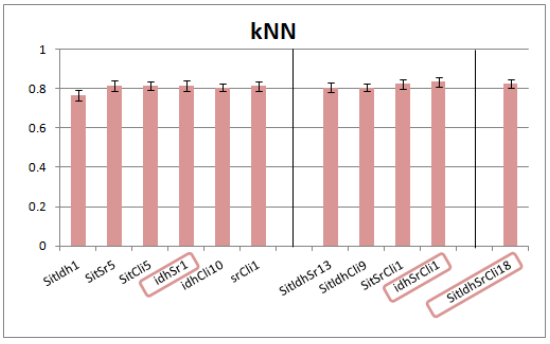
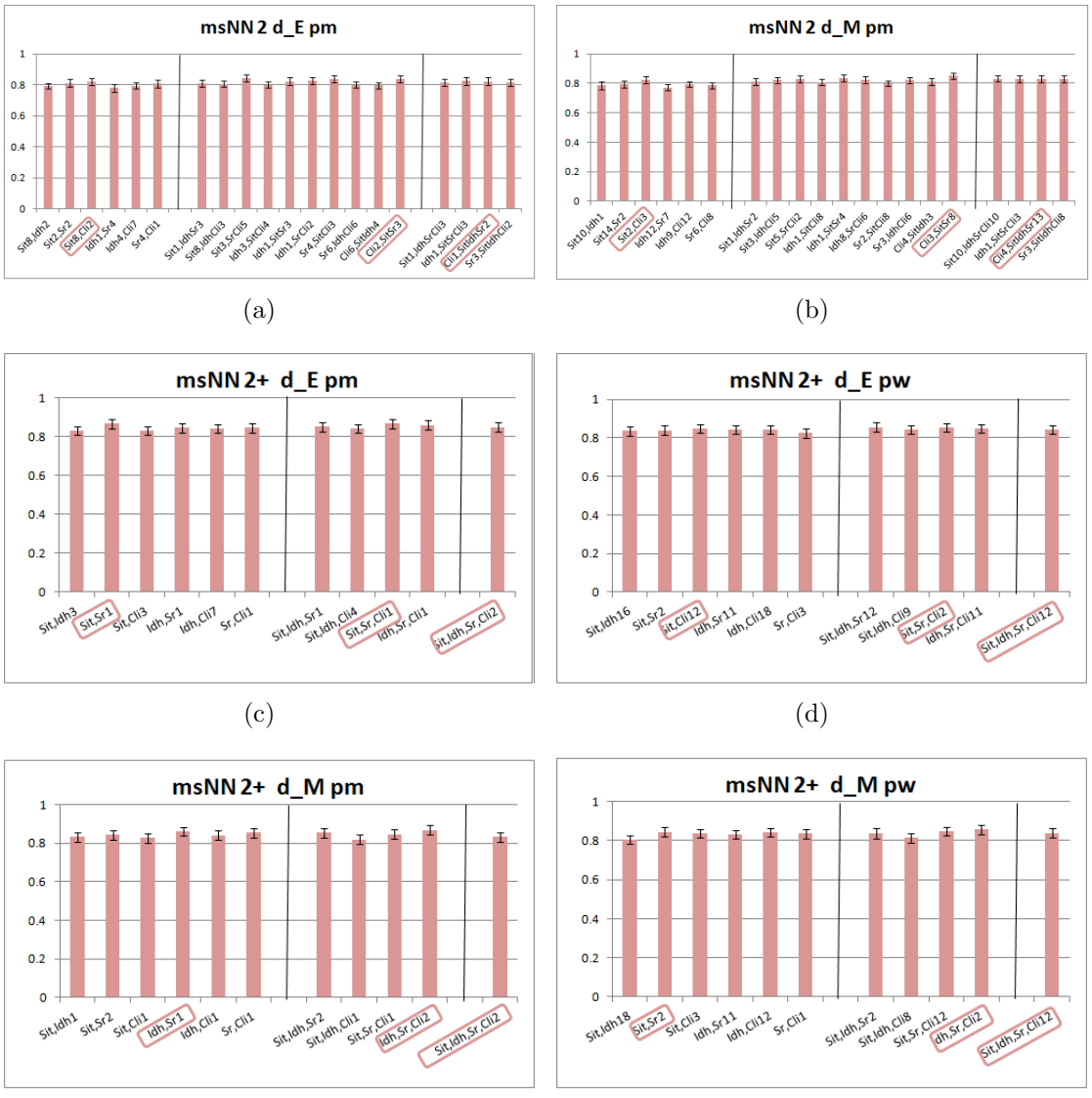
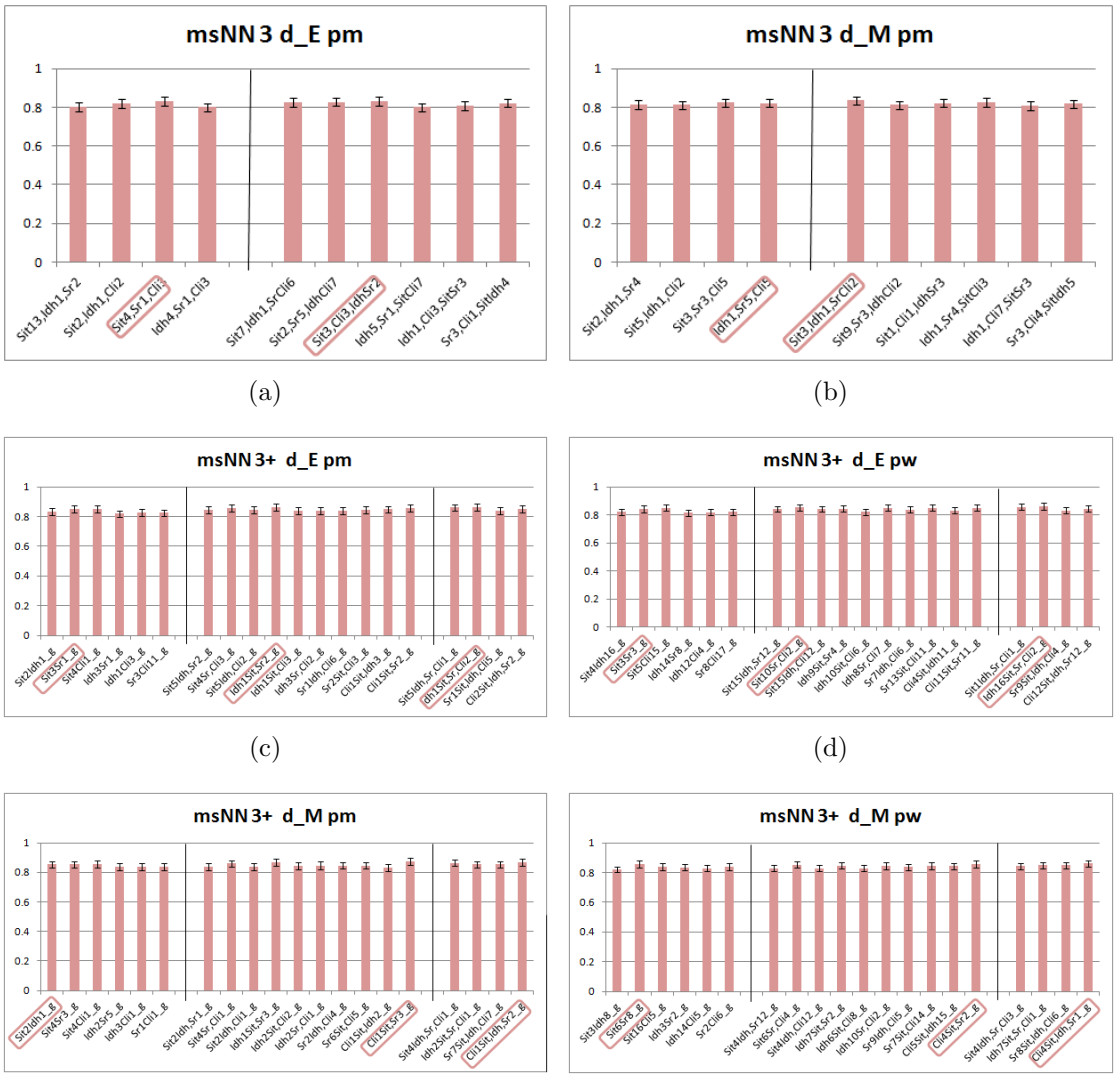


Figura B.14 - Média das acurácias para o método k-NN para abordagem com 4 classes, usando todas as amostras indenes.



(a) (b) (c) (d) (e) (f)

Figura B.15 - Média das acurácias para o método ms-NN para 2 espaços: 2 ATR (a) e (b); 1 ATR e 1 GEO (c), (d), (e) e (f) para abordagem com 4 classes, usando todas as amostras indenes.



(a) (b) (c) (d) (e) (f)

Figura B.16 - Média das acurácias para o método ms-NN para 3 espaços: 3 ATR (a) e (b); 2 ATR e 1 GEO (c), (d), (e) e (f) para abordagem com 4 classes, usando todas as amostras indenes.

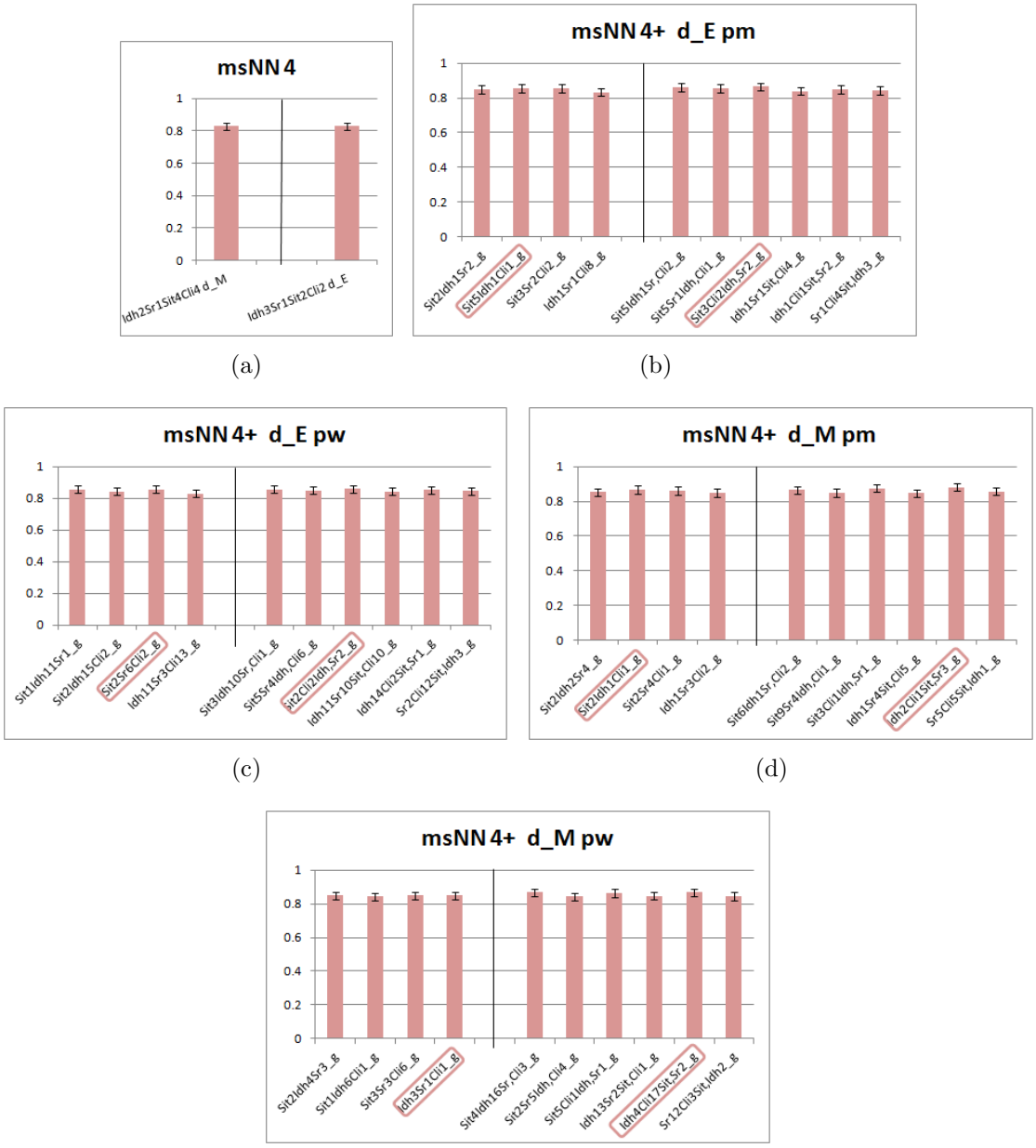
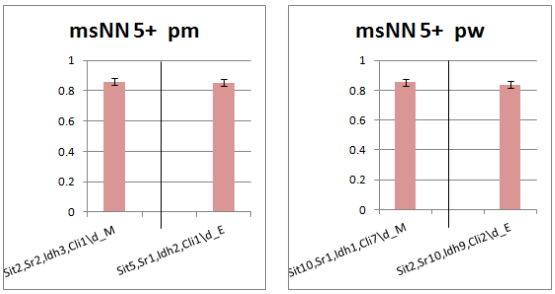


Figura B.17 - Média das acurácias para o método ms-NN para 4 espaços: 4 ATR (a); 3 ATR e 1 GEO (b), (c) e (d) para abordagem com 4 classes, usando todas as amostras indenes.

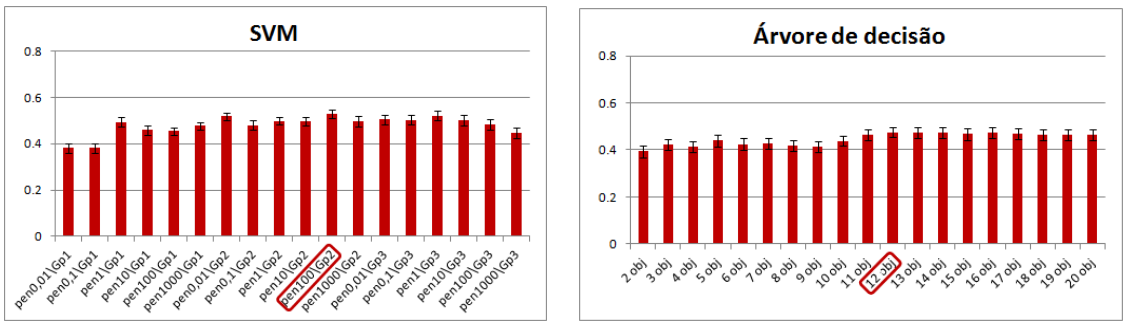


(a) (b)

Figura B.18 - Média das acurácias para o método ms-NN para 5 espaços, 4 ATR e 1 GEO para abordagem com 4 classes, usando todas as amostras indenes.

APÊNDICE C - GRÁFICOS SCHISTO LOCALIDADES

As classificações por árvore de decisão e SVM a nível local podem ser vistas na Figura C.1. Na Figura C.2 podem ser vistas as classificações pre selecionadas pelo número de vizinhos. Os gráficos das classifiáções pre-selecioandas do ms-NN estão nas Figuras C.3, C.4 e C.5.



(a) (b)

Figura C.1 - Média das acurácias para o método SVM e Árvore de decisão para schisto em nível local.

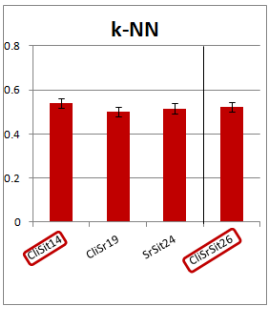
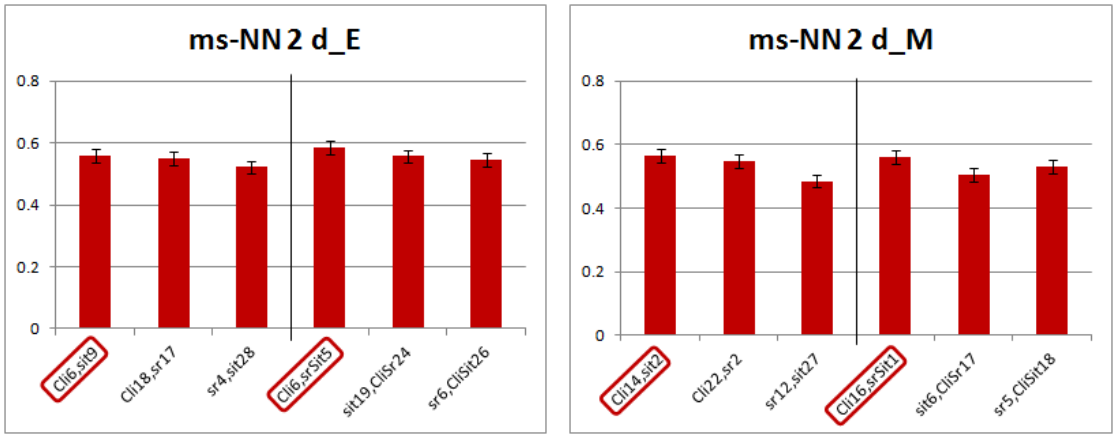
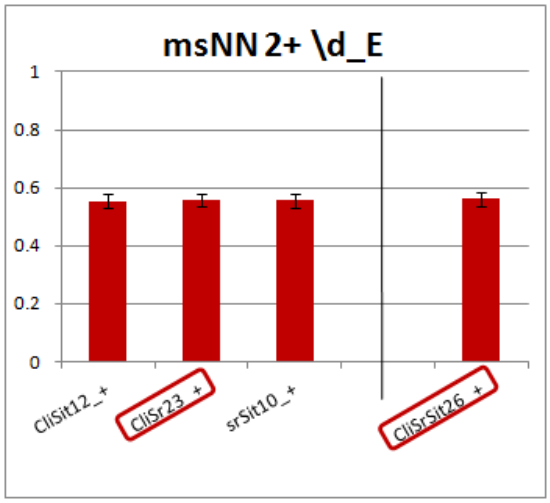


Figura C.2 - Média das acurácias para o método k-NN para Schisto em nível local.

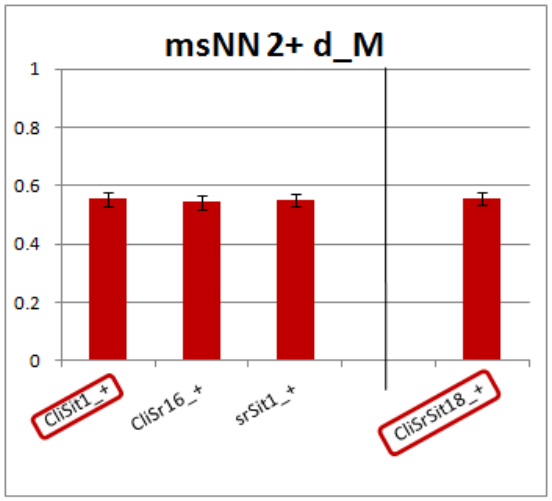


(a)

(b)

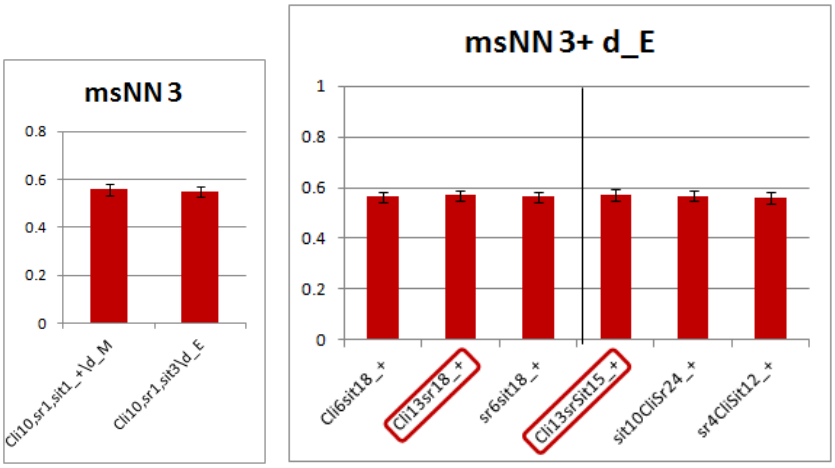


(c)

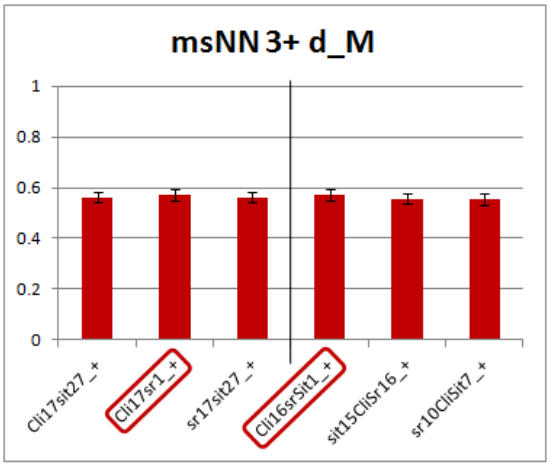


(d)

Figura C.3 - Média das acurácias para o método ms-NN para 2 espaços: 2 ATR (a) e (b); 1 ATR e 1 GEO (c), (d), para Schisto em nível local.

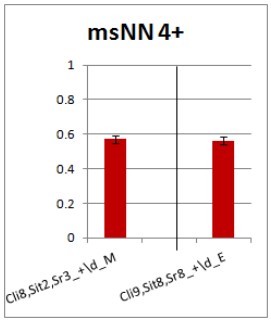


(a) (b)



(c)

Figura C.4 - Média das acurácias para o método ms-NN para 3 espaços: 3 ATR (a); 2 ATR e 1 GEO (b) e (c) para Schisto em nível local.



(a)

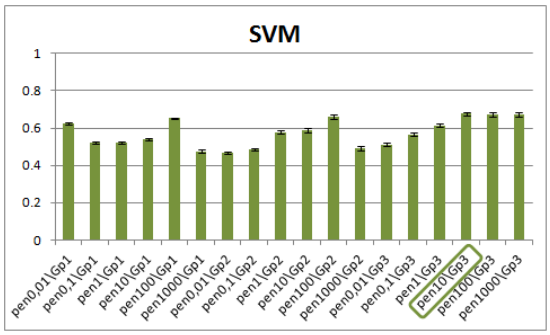
Figura C.5 - Média das acurácias para o método ms-NN para 4 espaços: 4 ATR (a); 3 ATR e 1 GEO para Schisto em nível local.

APÊNDICE D - GRÁFICOS TAPAJÓS

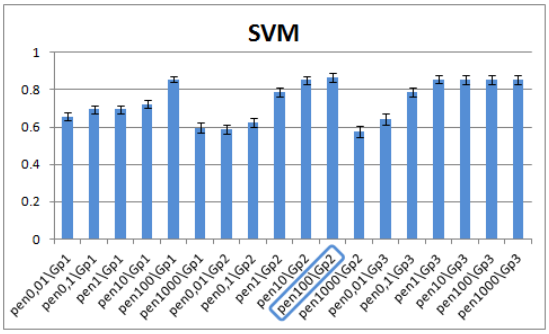
Neste Apêndice são apresentados os gráficos com:

- as classificações resultantes dos métodos SVM, árvore de decisão e k-NN (Figura D.1, D.2 e D.3);
- as classificações pre-selecionadas do método ms-NN (Figuras D.4, D.5 e D.6) para a distância euclidiana (d_E) e para a distância de mahalanobis (d_M) .

Nos gráficos com as classificações resultantes dos métodos SVM, árvore de decisão e k-NN estão destacadas os a classificação que foram selecionadas como representante de cada método. Os gráficos apresentam as médias de acurácia para cada configuração do classificador usada. No eixo horizontal dos gráficos do SVM estão as configurações dos parâmetros usados neste método: a penalidade (Pen) e o grau do polinômio (Gp). Para a árvore de decisão no eixo horizontal estão o número de objeto mínimo por folha, que foi usado como parâmetro de poda. Para o k-NN, no eixo horizontal estão o número de casos vizinhos.

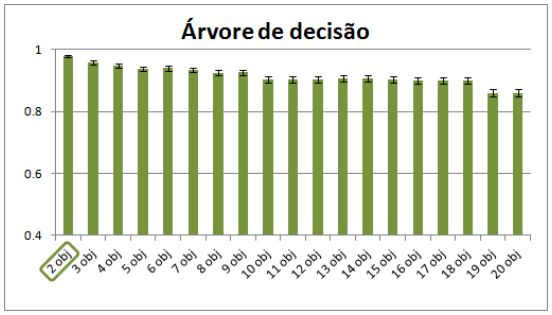


(a) Pixel

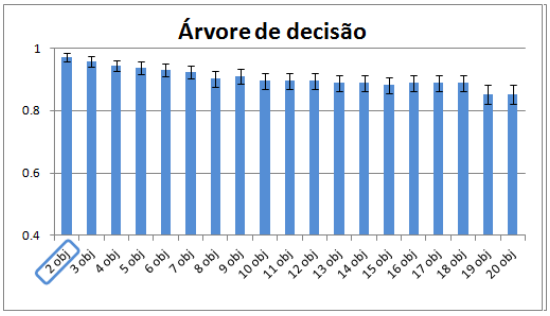


(b) Segmento

Figura D.1 - Média das acurácias para o método SVM

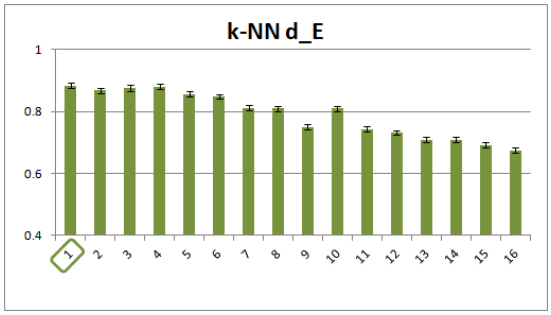


(a) Pixel

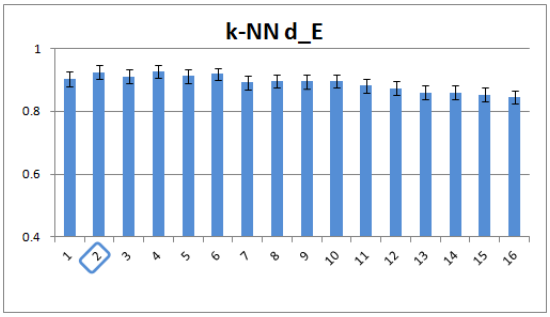


(b) Segmento

Figura D.2 - Média das acurácias para o método árvore de decisão

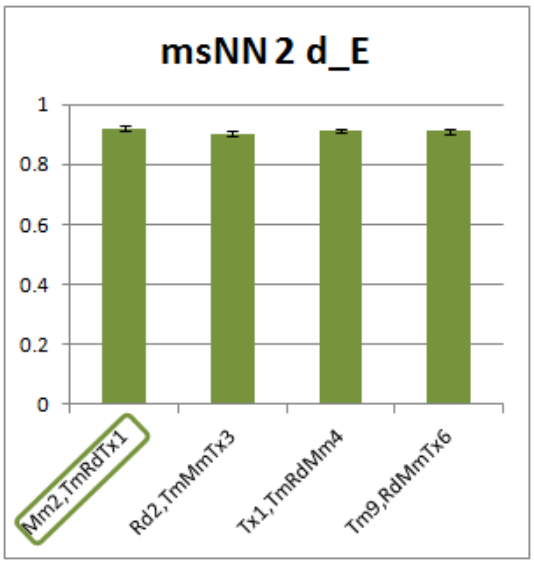


(a) Pixel

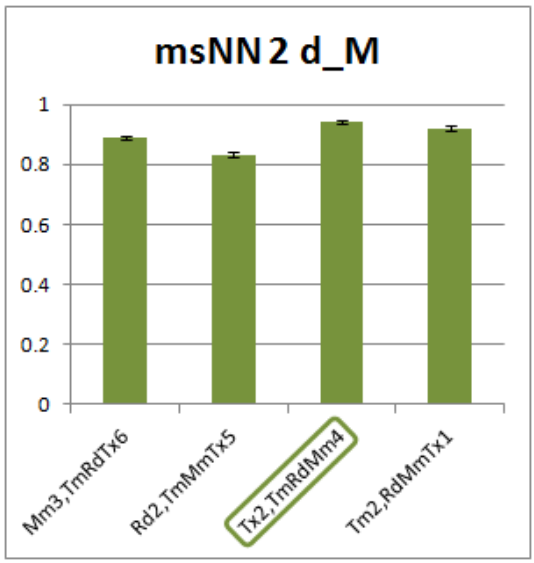


(b) Segmento

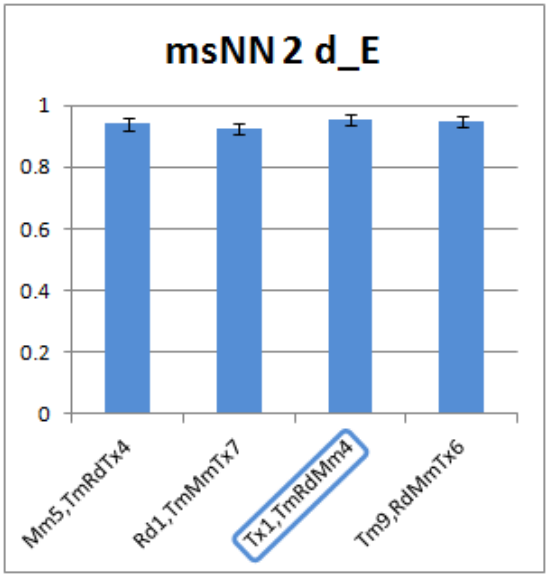
Figura D.3 - Média das acurácias para o método k-NN



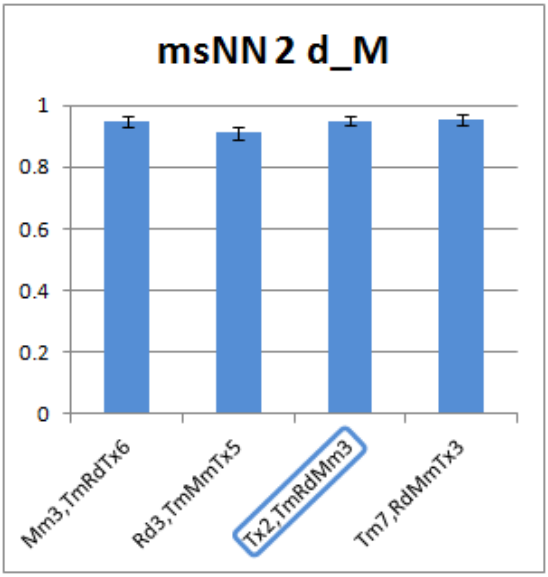
(a) Pixel - com dist. euclidiana



(b) Pixel - com dist. de mahalanobi



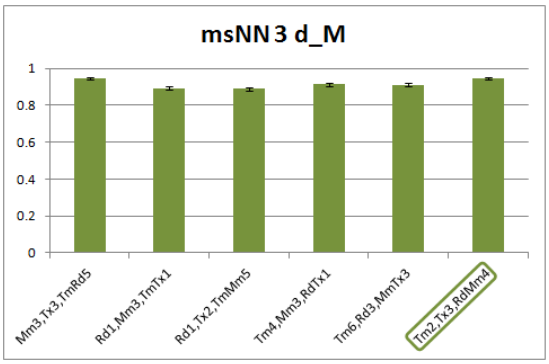
(c) Segmento - com dist. euclidiana



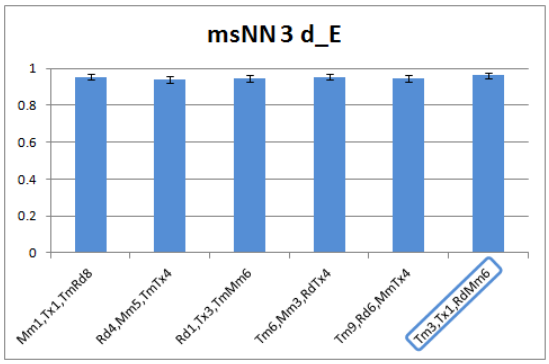
(d) Segmento - com dist. de mahalanobi

Figura D.4 - Média das acurácias para o método ms-NN para 2 espaços de ATR, usando distância Euclidiana (a), (c) e (e) e Mahalanobis (b), (d) e (f)

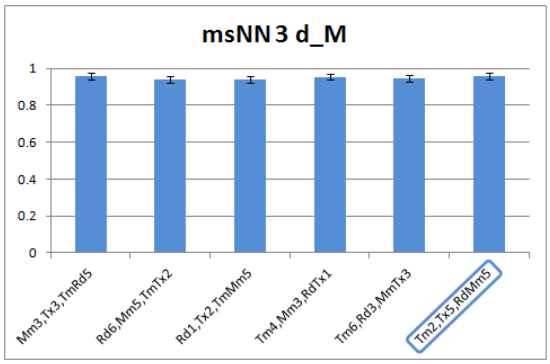
(a) Pixel - com dist. euclidiana



(b) Pixel - com dist. de mahalanobis

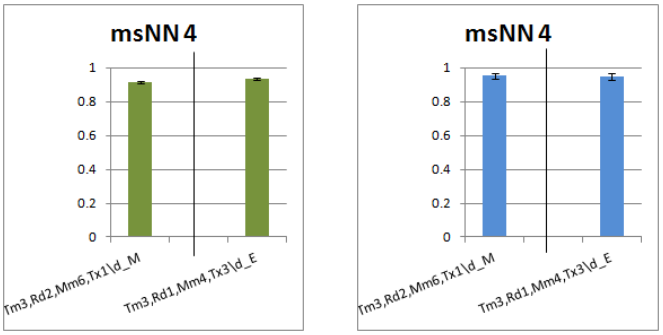


(c) Segmento - com dist. euclidiana



(d) Segmento - com dist. de mahalanobis

Figura D.5 - Média das acurácias para o método ms-NN para 3 espaços de ATR, usando distância Euclidiana (a), (c) e (e) e Mahalanobis (b), (d) e (f)

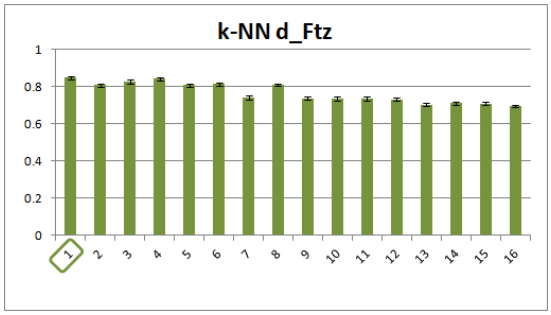


(a) Pixel (b) Segmento

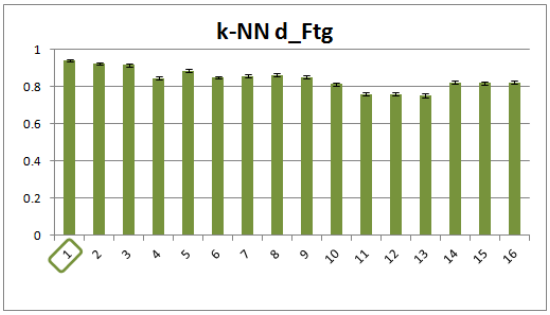
Figura D.6 - Média das acurácias para o método ms-NN para 4 espaços de ATR, usando distância Euclidiana (d_E) e Mahalanobis (d_M)

APÊNDICE E - GRÁFICOS TAPAJÓS COM RELAÇÕES DIFUSAS

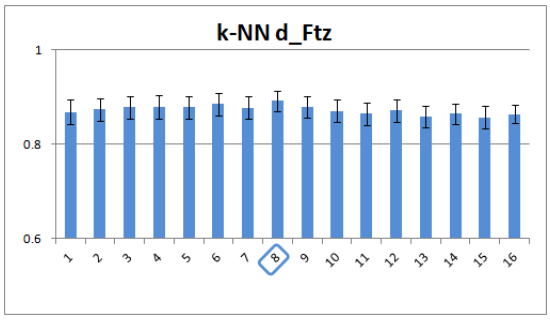
Neste Apêndice são apresentados os gráficos dos métodos k-NN e ms-NN em que são usadas uma função baseada em relações difusas e partições fuzzy trapezoidal (d_Ftz) triangular (d_Ftr).



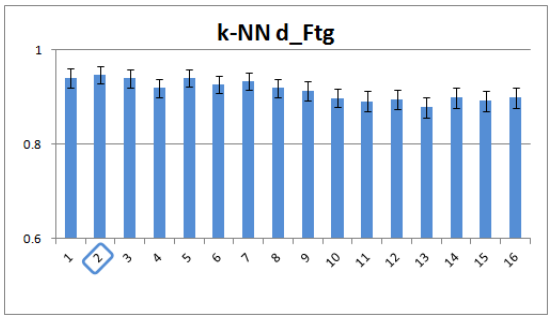
(a) Pixel



(b) Pixel

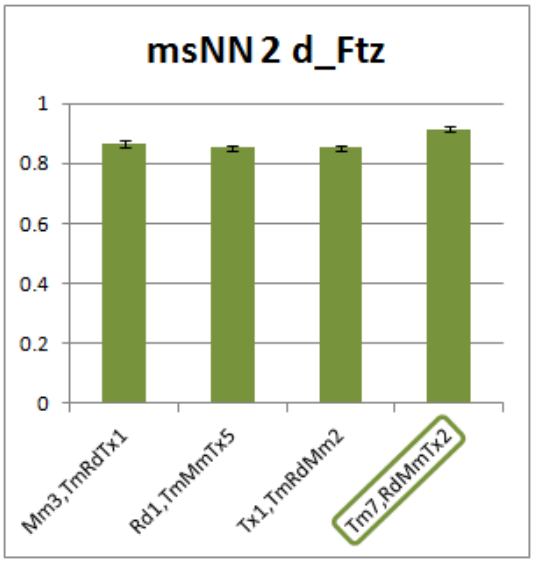


(c) Segmento

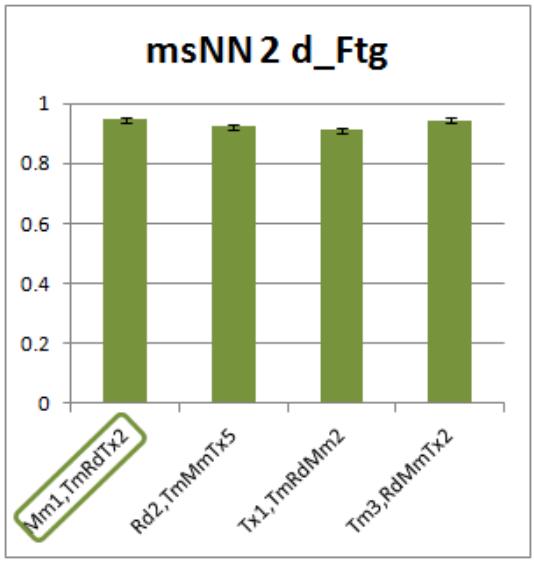


(d) Segmento

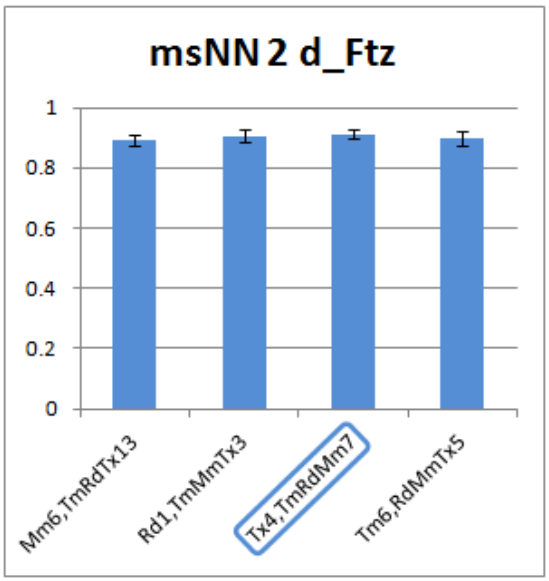
Figura E.1 - Média das acurácias para o método k-NN, usando partições fuzzy trapezoidal (a), (c) e (e) e partições fuzzy triangular (b), (d) e (f)



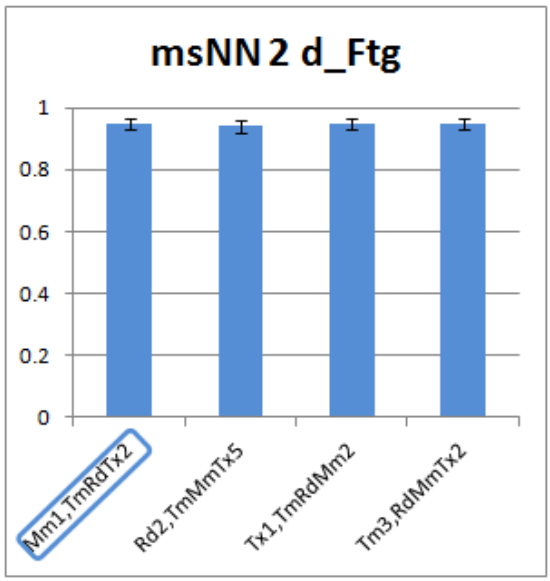
(a) Pixel - com dist. fuzzy



(b) Pixel - com dist. fuzzy

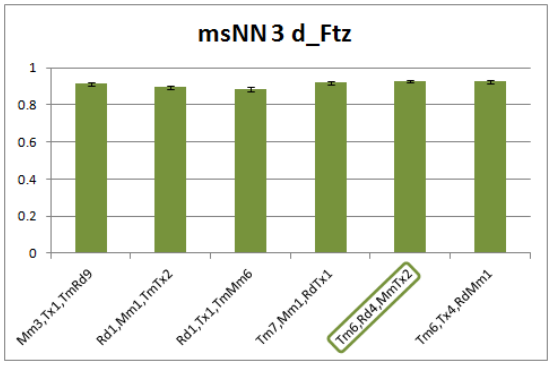


(c) Segmento - com dist. fuzzy

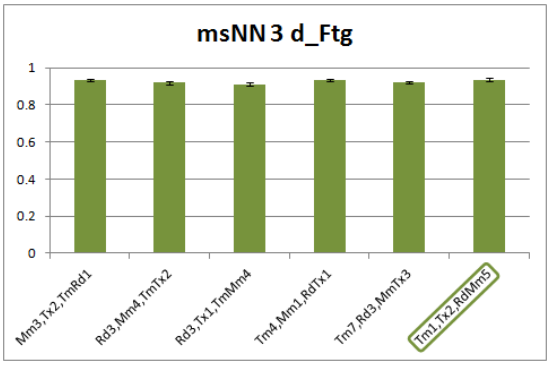


(d) Segmento - com dist. fuzzy

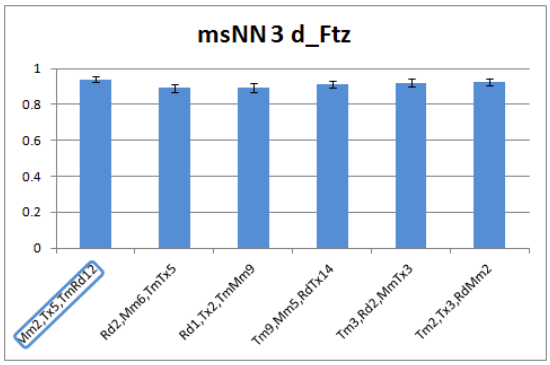
Figura E.2 - Média das acurácias para o método ms-NN para 2 espaços de ATR, usando partições fuzzy trapezoidal (a), (c) e (e) e partições fuzzy triangular (b), (d) e (f)



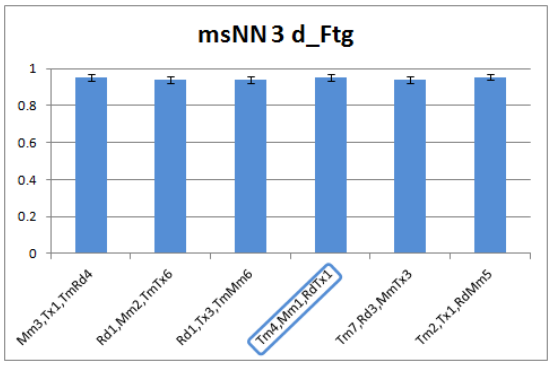
(a) Pixel - com dist. fuzzy



(b) Pixel - com dist. fuzzy

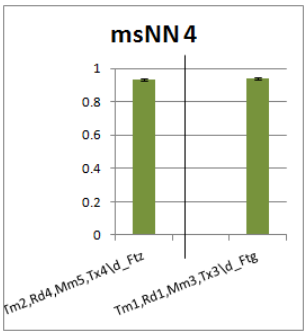


(c) Segmento - com dist. fuzzy

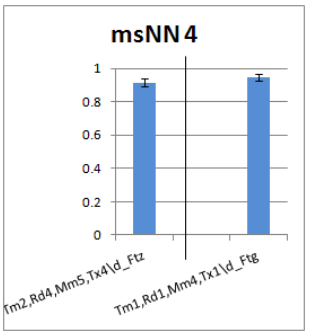


(d) Segmento - com dist. fuzzy

Figura E.3 - Média das acurácias para o método ms-NN para 3 espaços de ATR, usando partições fuzzy trapezoidal (a), (c) e (e) e partições fuzzy triangular (b), (d) e (f)



(a) Pixel - com dist. fuzzy



(b) Segmento - com dist. fuzzy

Figura E.4 - Média das acurácias para o método ms-NN para 4 espaços de ATR, usando partições fuzzy trapezoidal e triangular

PUBLICAÇÕES TÉCNICO-CIENTÍFICAS EDITADAS PELO INPE

Teses e Dissertações (TDI)

Teses e Dissertações apresentadas nos Cursos de Pós-Graduação do INPE.

Manuais Técnicos (MAN)

São publicações de caráter técnico que incluem normas, procedimentos, instruções e orientações.

Notas Técnico-Científicas (NTC)

Incluem resultados preliminares de pesquisa, descrição de equipamentos, descrição e ou documentação de programas de computador, descrição de sistemas e experimentos, apresentação de testes, dados, atlas, e documentação de projetos de engenharia.

Relatórios de Pesquisa (RPQ)

Reportam resultados ou progressos de pesquisas tanto de natureza técnica quanto científica, cujo nível seja compatível com o de uma publicação em periódico nacional ou internacional.

Propostas e Relatórios de Projetos (PRP)

São propostas de projetos técnico-científicos e relatórios de acompanhamento de projetos, atividades e convênios.

Publicações Didáticas (PUD)

Incluem apostilas, notas de aula e manuais didáticos.

Publicações Seriadas

São os seriados técnico-científicos: boletins, periódicos, anuários e anais de eventos (simpósios e congressos). Constam destas publicações o Internacional Standard Serial Number (ISSN), que é um código único e definitivo para identificação de títulos de seriados.

Programas de Computador (PDC)

São a seqüência de instruções ou códigos, expressos em uma linguagem de programação compilada ou interpretada, a ser executada por um computador para alcançar um determinado objetivo. Aceitam-se tanto programas fonte quanto os executáveis.

Pré-publicações (PRE)

Todos os artigos publicados em periódicos, anais e como capítulos de livros.