



Ministério da
**Ciência, Tecnologia
e Inovação**



sid.inpe.br/mtc-m19/2013/02.19.17.40 -TDI

ESTUDO DE RELAÇÕES DE PROXIMIDADE DIFUSAS APLICADAS AO RACIOCÍNIO BASEADO EM CASOS

Jonas Henrique Mendonça

Dissertação de Mestrado do Curso de Pós-Graduação em Computação Aplicada, orientada pelas Dras. Sandra Aparecida Sandri, e Maria Isabel Sobral Escada, aprovada em 31 de janeiro de 2013.

URL do documento original:

<<http://urlib.net/8JMKD3MGP7W/3DJH6UH>>

INPE
São José dos Campos
2013

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

CONSELHO DE EDITORAÇÃO E PRESERVAÇÃO DA PRODUÇÃO INTELLECTUAL DO INPE (RE/DIR-204):**Presidente:**

Marciana Leite Ribeiro - Serviço de Informação e Documentação (SID)

Membros:

Dr. Antonio Fernando Bertachini de Almeida Prado - Coordenação Engenharia e Tecnologia Espacial (ETE)

Dr^a Inez Staciarini Batista - Coordenação Ciências Espaciais e Atmosféricas (CEA)

Dr. Gerald Jean Francis Banon - Coordenação Observação da Terra (OBT)

Dr. Germano de Souza Kienbaum - Centro de Tecnologias Especiais (CTE)

Dr. Manoel Alonso Gan - Centro de Previsão de Tempo e Estudos Climáticos (CPT)

Dr^a Maria do Carmo de Andrade Nono - Conselho de Pós-Graduação

Dr. Plínio Carlos Alvalá - Centro de Ciência do Sistema Terrestre (CST)

BIBLIOTECA DIGITAL:

Dr. Gerald Jean Francis Banon - Coordenação de Observação da Terra (OBT)

REVISÃO E NORMALIZAÇÃO DOCUMENTÁRIA:

Marciana Leite Ribeiro - Serviço de Informação e Documentação (SID)

Yolanda Ribeiro da Silva Souza - Serviço de Informação e Documentação (SID)

EDITORAÇÃO ELETRÔNICA:

Maria Tereza Smith de Brito - Serviço de Informação e Documentação (SID)

Luciana Manacero - Serviço de Informação e Documentação (SID)



Ministério da
**Ciência, Tecnologia
e Inovação**



sid.inpe.br/mtc-m19/2013/02.19.17.40 -TDI

ESTUDO DE RELAÇÕES DE PROXIMIDADE DIFUSAS APLICADAS AO RACIOCÍNIO BASEADO EM CASOS

Jonas Henrique Mendonça

Dissertação de Mestrado do Curso de Pós-Graduação em Computação Aplicada, orientada pelas Dras. Sandra Aparecida Sandri, e Maria Isabel Sobral Escada, aprovada em 31 de janeiro de 2013.

URL do documento original:

<<http://urlib.net/8JMKD3MGP7W/3DJH6UH>>

INPE
São José dos Campos
2013

Dados Internacionais de Catalogação na Publicação (CIP)

Mendonça, Jonas Henrique.
M523r Estudo de relações de proximidade difusas aplicadas ao raciocínio baseado em casos / Jonas Henrique Mendonça. – São José dos Campos : INPE, 2013.
xx + 80 p. ; (sid.inpe.br/mtc-m19/2013/02.19.17.40 -TDI)

Dissertação (Mestrado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2013.

Orientadoras : Dras. Sandra Aparecida Sandri, e Maria Isabel Sobral Escada.

1. lógica difusa 2. raciocínio baseado em casos 3. agrupamento
4. redes neurais artificiais . I.Título.

CDU 004.048



Esta obra foi licenciada sob uma Licença [Creative Commons Atribuição-NãoComercial 3.0 Não Adaptada](https://creativecommons.org/licenses/by-nc/3.0/).

This work is licensed under a [Creative Commons Attribution-NonCommercial 3.0 Unported License](https://creativecommons.org/licenses/by-nc/3.0/).

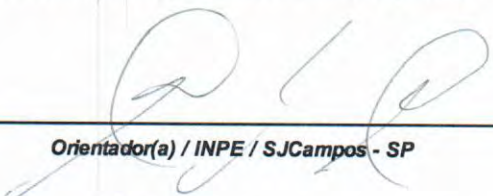
Aprovado (a) pela Banca Examinadora
em cumprimento ao requisito exigido para
obtenção do Título de **Mestre** em
Computação Aplicada

Dr. Marcos Gonçalves Quiles



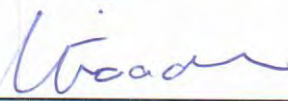
Presidente / UNIFESP / São José dos Campos - SP

Dra. Sandra Aparecida Sandri



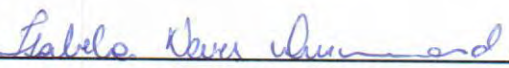
Orientador(a) / INPE / SJC Campos - SP

Dra. Maria Isabel Sobral Escada




Orientador(a) / INPE / SJC Campos - SP

Dra. Isabela Neves Drummond



Convidado(a) / UNIFEI / Itajubá - MG

Dr. Ricardo Coelho Silva



Convidado(a) / UNIFESP / São Paulo - SP

Este trabalho foi aprovado por:

maioria simples

unanimidade

Aluno (a): **Jonas Henrique Mendonça**

São José dos Campos, 31 de Janeiro de 2013

“ À medida que a complexidade aumenta, as declarações precisas perdem relevância e as declarações relevantes perdem precisão”.

LOTFI ZADEH

A meus pais Ivaldo e Doralice

AGRADECIMENTOS

Todo grande sonho não nasce sozinho. Eu tenho um sonho, não tão grandioso, e as poucas partes desse sonho que já foram realizadas aconteceram não somente por minha vontade, mas pela sorte de ter a família, os amigos e os mestres que tenho. Sem essas pessoas, eu não seria nada.

A começar por minha família que é meu porto seguro, não importa onde, nem quando, ela está sempre lá para me dar segurança de seguir esse meu sonho. Devo muito a meus pais, meus exemplos de garra e coragem para enfrentar os maiores desafios da vida. Agradeço, a meus irmãos, Isleyde, Daniela, Ana Paula e Pablo, por serem os melhores irmãos que alguém pode ter e a meus sobrinhos a quem tantas horas de brincadeiras e jogos foram adiadas.

A minha orientadora, Dra. Sandra Sandri, pela orientação e incentivo constante, dividindo o seu profundo conhecimento, sua sabedoria e amizade. A Dra. Maria Isabel Sobral Escada e a Flávia Toledo Martins-Bedê pelo apoio para a realização deste projeto. A todos os professores do INPE que de uma forma ou de outra me conduziram para a finalização de mais uma etapa da minha vida.

Aos membros da banca examinadora pela disposição em analisar este trabalho. Ao Conselho Nacional de Pesquisa pelo auxílio financeiro, ao Instituto Nacional de Pesquisas Espaciais (INPE) pela oportunidade e apoio.

Só depois de anos de convivência a gente percebe que seu melhor amigo e você podem fazer qualquer coisa, ou nada, e terem bons momentos juntos. Eder, Marina, Maria Teodora, Lígia, Daniel, André, Michelle, Ivana, Sabrina, Carol, Érica, Felipe, Marluce, Pedro, Marlon e Wanderson vocês são pessoas formidáveis.

Finalmente, agradeço a Deus pela saúde e fé na vida. Muito Obrigado.

RESUMO

O presente trabalho apresenta uma abordagem de raciocínio baseado em casos utilizando relações de proximidade difusas. A ideia básica do RBC é resolver um problema a partir do conhecimento de problemas passados, comparando-os com o novo problema, adaptando assim uma nova solução. Considera-se aqui que pesos podem ser associados aos casos porém, este processo pode ser computacionalmente custoso. Para isso, uma metodologia para cálculo de agrupamentos foi estendida com a finalidade de se calcular os pesos a partir de fragmentos da base de casos. A partir da metodologia para cálculo de agrupamentos, foi proposta uma tipologia tanto para treinamento e aprendizado dos vetores de pesos quanto para cálculo dos resultados. Esta extensão proposta foi aplicada a dois estudos de casos. No primeiro, para estimar a prevalência da esquistossomose no estado de Minas Gerais e, no segundo, a metodologia foi aplicada para classificar padrões de desmatamento em Terra do Meio no estado do Pará. Os resultados obtidos foram aplicados medidas de qualidade da classificação de dados e propôs-se uma maneira de analisar a classificação de dados temporais.

ESTUDY OF PROXIMITY FUZZY RELATIONS APPLIED TO THE CASE-BASED REASONING

ABSTRACT

This work presents an approach of case-based reasoning using fuzzy similarity relations. The basic idea of CBR is to solve a problem from the knowledge of past problems, comparing them with the new problem, thus customizing a new solution. Within this context, it presents a brief description of case-based reasoning and fuzzy logic. Weights can be attached to cases however, this process can be computationally expensive. For this, a method for calculating cluster was extended to calculate the weights from fragments of case base. This methodology was applied to two case studies: to estimate the prevalence of schistosomiasis in the state of Minas Gerais and to classify patterns of deforestation in Terra do Meio in the state of Pará.

LISTA DE FIGURAS

	<u>Pág.</u>
2.1 Principais t-normas	6
2.2 Principais t-conormas	7
2.3 Ciclo de funcionamento de sistemas RBC	9
2.4 Exemplo de hipergrafo	11
4.1 Tipos de Treinamento	20
4.2 Agrupamentos para cálculo das soluções	20
4.3 Tipos de experimentos propostos	21
5.1 Municípios mineiros cuja prevalência de esquistossomose é conhecida. Fonte: (MARTINS et al., 2008)	23
5.2 Regionalização obtida através do algoritmo SKATER	25
6.1 Mapa da Terra do Meio	33
6.2 Avaliação do tamanho das células para a definição de tipologia de padrões de ocupação.	35
6.3 Descrição de padrões de desmatamento e tipologia de ocupação identificados nas análises de imagens de satélites e trabalhos de campo.	36
6.4 Árvore de decisão utilizada para a classificação das células em padrões de ocupação nos anos de 1997, 2000, 2003, 2006 e 2009.	37
6.5 Resultado da classificação por células da Terra do Meio nos anos de 1997, 2000, 2003, 2006 e 2009.	38
7.1 Função de pertinência- a) A_1 b) A_2 c) A_3	45
7.2 Funções de similaridade a) $W + R+$ b) $WR+$ c) $W + +R+$ d) $W - R$	46
7.3 Invólucro convexo da função de pertinência da estimativa	49
7.4 Função de similaridade caso C31L16 a) 1997 b) 2000	50
8.1 Estimativas B_\bullet e B_\blacktriangle	52
8.2 Estimativas B^* , B_\bullet e B_\blacktriangle	53
8.3 Convexidade	54
8.4 Grafo e autômato para estudo de caso padrões de desmatamento	56
8.5 Estimativas temporais	57

LISTA DE TABELAS

	<u>Pág.</u>
2.1 Principais T-normas e t-conormas Fonte: (SANDRI; CORREA, 1999)	6
5.1 Classificação, para conjuntos de treinamento e teste, com aprendizado: R-Reg (base regional e regressão), G-Reg (base global e regressão) e G-DT (base global e árvore de decisão). Fonte: (MARTINS-BEDÊ et al., 2009)	26
5.2 Classificação com os modos de aprendizado R-Reg (Base regional e re- gressão), G-Reg (Base global e regressão), R-Sim (Base regional e simi- laridade) e G-Sim (Base global e similaridade) Fonte: (MARTINS-BEDÊ et al., 2009)	27
5.3 Classificação para a abordagem regional para os experimentos W-R, W- R+, WR, WR+, W+R+, W++R+ (Casos de treinamento).	27
5.4 Classificação para a abordagem global para os experimentos W-R, W- R+, WR, WR+, W+R+, W++R+ (Casos de treinamento)	27
5.5 Classificação para a abordagem regional para os experimentos W-R, W- R+, WR, WR+, W+R+, W++R+ (Casos de teste)	28
5.6 Classificação para a abordagem global para os experimentos W-R, W- R+, WR, WR+, W+R+, W++R+ (Casos de teste)	28
5.7 Classificação dos casos da região R3 para os experimentos W-R, W-R+, WR, WR+, W+R+, W++R+ utilizando fator de coesão. (Fonte: (??) .	29
5.8 Resultados obtidos quando os agrupamentos foram calculados utilizando a rede fuzzy-ART (Casos de teste) Fonte: (??)	30
6.1 Evolução dos padrões de ocupação em % do total da área analisada entre 1997 e 2009. Fonte: (LOBO; ESCADA, 2010)	37
6.2 Matriz de confusão da validação da classificação por células utilizando o classificador GeoDMA. Fonte: (LOBO; ESCADA, 2010)	39
6.3 Resultados obtidos a partir das relações de proximidade difusas - Trei- namento 1	39
6.4 Resultados obtidos a partir das relações de proximidade difusas - Teste 1	40
6.5 Resultados obtidos a partir das relações de proximidade difusas - Trei- namento 2	40
6.6 Resultados obtidos a partir das relações de proximidade difusas - Teste 2	40
7.1 Acurácia global média da classificação efetuada no segundo estudo de caso	47
7.2 Precisão global média da classificação efetuada no segundo estudo de caso	47

7.3	Qualidade global média da classificação efetuada no segundo estudo de caso	47
7.4	Qualidade média por experimento	48
8.1	Qualidade média da classificação efetuada no segundo estudo de caso considerando a evolução temporal	59
8.2	Qualidade média da classificação efetuada no segundo estudo de caso considerando a evolução temporal	59
.1	Matrizes de confusão - Região 1 - Abordagem Regional	67
.2	Matrizes de confusão - Região 1 - Abordagem Global	67
.3	Matrizes de confusão - Região 2 - Abordagem Regional	68
.4	Matrizes de confusão - Região 2 - Abordagem Global	68
.5	Matrizes de confusão - Região 3 - Abordagem Regional	69
.6	Matrizes de confusão - Região 3 - Abordagem Global	69
.7	Matrizes de confusão - Região 4 - Abordagem Regional	70
.8	Matrizes de confusão - Região 4 - Abordagem Global	70
.1	Matrizes de confusão - Casos de Treinamento - 1997	71
.2	Matrizes de confusão - Casos de Teste - 1997	72
.3	Matrizes de confusão - Casos de Treinamento - 2000	72
.4	Matrizes de confusão - Casos de Teste - 2000	73
.5	Matrizes de confusão - Casos de Treinamento - 2003	73
.6	Matrizes de confusão - Casos de Teste - 2003	74
.7	Matrizes de confusão - Casos de Treinamento - 2006	74
.8	Matrizes de confusão - Casos de Teste - 2006	75
.9	Matrizes de confusão - Casos de Treinamento - 2009	75
.10	Matrizes de confusão - Casos de Teste - 2009	76

SUMÁRIO

	<u>Pág.</u>
1 INTRODUÇÃO	1
2 FUNDAMENTAÇÃO TEÓRICA	5
2.1 Teoria dos conjuntos difusos	5
2.1.1 Operadores lógicos difusos	6
2.1.2 Relações difusas	7
2.2 Raciocínio baseado em casos	8
2.3 Hipergrafos	10
3 METODOLOGIA DE CLASSIFICAÇÃO	13
3.1 Definições básicas	13
3.2 Relações de semelhança difusa	14
3.3 Obtendo as relações de semelhança clássica	14
3.4 Usando atributos ponderados	15
3.5 Obtendo agrupamentos para uma base de casos	16
3.6 Calculando uma solução para um novo problema de acordo com um agrupamento	16
3.7 Determinando a força de um agrupamento em relação a um novo problema	17
4 PROPOSTAS DE TIPOLOGIA DE EXPERIMENTOS E DE DETERMINAÇÃO DE AGRUPAMENTOS	19
4.1 Cálculo de agrupamentos com Redes Neurais Artificiais	19
4.2 Tipologia para Treinamento	19
5 ESTUDO DE CASO: PREVALÊNCIA DA ESQUISTOSSOMOSE	23
5.1 Esquistossomose	23
5.2 Experimentos originais	24
5.3 Experimentos e resultados utilizando a abordagem de proximidade difusa ponderada	26
5.4 Extensões	28
5.4.1 Fator de Coesão	28
5.4.2 Agrupamento utilizando RNA Fuzzy-ART	29
5.5 Análise	30

6 ESTUDO DE CASO: PADRÕES DE DESMATAMENTO	33
6.1 Tipos de Desmatamento	33
6.2 Experimentos Originais	36
6.3 Experimentos e resultados utilizando a abordagem de proximidade difusa ponderada	38
7 QUALIDADE DE ESTIMATIVAS E TRATAMENTO DE DADOS TEMPORAIS	43
7.1 Medidas de qualidade de distribuições de possibilidade	43
7.1.1 Acurácia e Precisão	43
7.2 Aplicação da Metodologia	46
7.3 Convexidade como medida de qualidade de distribuições de possibilidade	48
8 MEDIDA DE QUALIDADE DE EVOLUÇÃO TEMPORAL	51
8.1 Medidas gerais de consistência de evolução temporal	51
8.2 Medidas de inconsistência para aplicações com evolução temporal monotônica	54
8.2.1 Verificação de evolução temporal monotônica	55
8.2.2 Verificação de consistência de evolução temporal monotônica	56
8.3 Experimentos	59
9 CONCLUSÕES	61
REFERÊNCIAS BIBLIOGRÁFICAS	63
ANEXO A - Matrizes de confusão - Estudo de Caso: Prevalência da Esquistossomose	67
ANEXO B - Matrizes de confusão - Estudo de Caso: Padrões de Desmatamento	71

1 INTRODUÇÃO

A técnica de *Raciocínio Baseado em Casos* (RBC) (KOLODNER, 1993) se propõe a resolver um problema usando um princípio que pode ser declarado como “problemas semelhantes tem soluções semelhantes” (AAMODT; PLAZA, 1994). A base consiste em problemas resolvidos modelados como pares (problema, solução) e é utilizada para determinar a solução para um novo problema. O primeira etapa deste procedimento consiste em recuperar problemas na base que são semelhantes ao problema considerado: ela determina os casos da base que são relevantes para a solução do problema. A segunda etapa consiste em reutilizar as soluções desses problemas relevantes, adaptando-os para o problema considerado.

É possível associar-se um peso a cada caso de maneira que os considerados mais importantes para uma determinada aplicação tenham pesos mais elevados. Além disso, pode-se associar um vetor individual de pesos nas variáveis a cada caso o que recai em ponderar as variáveis independentemente dos casos. Mas, pode-se atribuir vetores de pesos individuais para cada caso, de modo que mais atributos significativos no interior de um caso recebam pesos mais elevados. Em (ARMENGOL et al., 2004) e (MARTINS-BEDÊ et al., 2009), relações de proximidade difusa associadas a cada variável de descrição e solução foram utilizadas para derivar vetores de pesos individuais, através do algoritmo de aprendizagem proposto em (TORRA, 2000). Foi mostrado que o uso de vetores de pesos individuais para cada caso do conjunto de treinamento tende a levar a melhores resultados que a utilização de casos não ponderados.

O problema com a utilização dos pesos é que o processo de aprendizado geralmente é computacionalmente caro, o que pode impossibilitar sua utilização em grandes bases de casos. Uma abordagem para permitir a aprendizagem ponderada em bases de dados grandes consiste na extração de fragmentos da base de dados e na obtenção de pesos para cada um desses fragmentos. O cálculo da solução para um novo caso utiliza os fragmentos cujos problemas são semelhantes aos casos em questão. Um modelo para fragmentar as bases foi proposto em (FANOIKI et al., 2010). O método proposto é baseado na proximidade binária entre casos, chamado relação de semelhança entre casos, do inglês *Case Resemblance Relation* (CRR) ¹ que leva em conta tanto a semelhança nas entradas quanto nas saídas do problema. Esta medida define uma relação binária entre casos. O grafo de casos correspondente é então explorado

¹Os novos termos citados no contexto deste trabalho estão em inglês garantindo coerência com a notação utilizada em artigos já publicados

e decomposto em grupos de casos similares tanto na descrição do problema quanto na solução. Para calcular a solução de novos problemas usa-se o agrupamento mais semelhante a este caso.

Esta abordagem foi generalizada para o caso onde a medida de proximidade entre os casos, definida como a agregação entre as proximidades de entrada e de saída entre os casos, não é binária, mas apresenta valores no intervalo $[0, 1]$ definindo uma relação difusa (SANDRI, 2012). O problema é, em seguida, como extrair agrupamentos desta relação difusa. Para lidar com este problema, em primeiro lugar, os correspondentes cortes de nível são obtidos a partir da relação difusa, criando conjuntos de relações clássicas. Cada relação clássica é uma CRR, e, assim, o processo proposto para a abordagem clássica é aplicado.

O primeiro objetivo principal deste trabalho é estudar os efeitos do uso de casos ponderados e dos agrupamentos na tarefa de classificação. Para isso, foram estudadas duas maneiras para cálculo dos agrupamentos uma utilizando a abordagem difusa descrita acima e outra utilizando-se redes neurais artificiais do tipo fuzzy-ART. Propõe-se aqui também uma tipologia de experimentos considerando qual base de treinamento é utilizada para encontrar os vetores de pesos para os casos dentro de um agrupamento e a maneira com que estes pesos são utilizados para o cálculo da solução. Nesta tipologia os experimentos podem ser realizados sem utilização de pesos ($W-$), com pesos aprendidos para toda a base (W), ou com pesos aprendidos com a base fragmentada em agrupamentos ($W + eW + +$) (Figura 4.1). Por sua vez, podem ser calculadas soluções para cada agrupamento ($R+$), que podem então competir entre si pela solução final, ou serem agregadas em uma única solução, ou considerando-se somente um agrupamento, a própria base (R)

Com o objetivo de se obter um enriquecimento prático acerca da metodologia proposta, foi implementado um sistema de classificação não supervisionada de dados. Toda implementação foi feita utilizando-se a plataforma Eclipse com a linguagem de programação Java, que possui características de orientação a objetos. Para cálculo dos vetores de pesos utilizou-se os conceitos apresentados em (TORRA, 2000).

A validação desta tipologia foi feita através de dois estudos de casos. O primeiro consiste em estimar a prevalência de esquistossomose no estado de Minas Gerais. O segundo classifica padrões de ocupação na região da Terra do Meio (Pará). O segundo objetivo principal é avaliar como a metodologia proposta se comporta na análise de dados temporais. Para tanto, propõem-se neste trabalho medidas de qualidade de classificação.

Uma vez determinados os agrupamentos de uma base de casos o primeiro passo do processo de aprendizado de pesos consiste em definir-se a base de treinamento para cada agrupamento. A base de treinamento para se obter os pesos dos atributos para cada caso em um agrupamento pode ser o próprio agrupamento ou uma versão expandida deste. Neste trabalho, foram elaborados diversos tipos de experimentos para classificação das aplicações utilizando a tipologia proposta.

Quando os casos apresentam uma evolução temporal pode-se expandir esta abordagem para que seja feita análise das alterações nas soluções de cada caso. Esta análise é feita comparando-se a semelhança de cada um dos problemas que compõem a base de casos com os agrupamentos obtidos através da abordagem apresentada. Porém, as possíveis soluções destes casos devem apresentar uma correlação temporal de maneira que uma solução inicial evolua para as demais.

A esquistossomose é uma doença com características sociais e de comportamento. Caramujos da espécie *Biomphalaria*, o hospedeiro intermediário da doença, utilizam a água como meio para infectar o homem. No Brasil, 6 milhões de pessoas estão infectadas, principalmente em regiões pobres do país. De acordo com o Sistema de Informação de Agravos de Notificação (SINAN) do Ministério da Saúde, de 1995 a 2005, mais de 1 milhão de casos foram diagnosticados, 27% deles no estado de Minas Gerais.

Em (MARTINS et al., 2008), os autores apresentam uma classificação da prevalência da esquistossomose no estado de Minas Gerais usando variáveis de sensoriamento remoto, climáticas, socioeconômicas e características de vizinhança. Duas abordagens foram utilizadas, uma global e outra regional. Na primeira, um único modelo de regressão foi gerado e usado para estimar o risco da doença para todo o estado. Na segunda, o estado foi dividido em 4 regiões e um modelo foi gerado para cada uma delas. Neste trabalho, para realização dos experimentos propostos adotou-se a mesma regionalização da base de casos.

A Terra do Meio compreende os Municípios de São Félix do Xingu, Tucumã e Altamira, e localiza-se entre dois importantes rios na região central do Estado do Pará, o Rio Xingu, um dos maiores tributários do Rio Amazonas, e o Rio Iriri. A área inclui uma frente de desmatamento, e sua ocupação está associada à presença de diferentes tipos de atividades econômicas, refletindo transformações na paisagem e perdas significativas da cobertura florestal (ESCADA M. I AND PINHO et al., 2010) (Escada et al, 2010). Em (Lobo et. al., 2010) a tarefa de classificação dos padrões de desmatamento foi realizada utilizando-se árvores de decisão. Como resultado, foram

obtidos cinco mapas para cada ano analisado, contendo os padrões de interesse. A partir dos mapas de padrões foi possível traçar as principais trajetórias dos padrões de ocupação ao longo do tempo. As mesmas variáveis aplicadas em (LOBO; ESCADA, 2010) foram utilizadas no processo de classificação aqui adotado.

Para avaliar como a metodologia proposta se comporta na análise de dados temporais, neste trabalho são propostas medidas de qualidade da classificação. Estas medidas podem se basear na diferença entre a função que modela a realidade e a função que modela a estimativa. Porém, isto nem sempre é adequado. Por isso, propõem-se o uso de medidas que considerem a suavidade das funções (convexidade e regularização de 2ª ordem de Tikhonov) e uma medida que considera aplicações onde a evolução temporal é necessariamente monotônica. Tais medidas foram aplicados ao estudo de caso que compreende a classificação de padrões de desmatamento.

Este trabalho está organizado da seguinte maneira: o Capítulo 2 apresenta os conceitos teóricos utilizados para elaboração deste trabalho. As metodologias utilizadas são descritas no capítulo 3. O Capítulo 4 traz as contribuições deste trabalho. O primeiro estudo de caso, referente à prevalência de esquistossomose, e seus respectivos resultados são exibidos no Capítulo 5. O Capítulo 6 traz o segundo estudo de casos, que se destina a classificação de áreas de desmatamento. O Capítulo 7 trata a qualidade da classificação de padrões de desmatamento, com um estudo de caso sobre a Terra do Meio. Medidas de qualidade da classificação temporal dos casos foram propostas e são mostradas no Capítulo 8. Finalmente, no Capítulo 9 são descritas as conclusões obtidas durante a realização deste trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo limita-se à apresentação dos principais conceitos teóricos necessários ao desenvolvimento deste trabalho. Inicia-se com a definição de lógica difusa e relações de proximidade. Na seção seguinte, define-se o raciocínio baseado em casos, descrevendo-se suas características e etapas. O capítulo se encerra com a apresentação dos hipergrafos.

2.1 Teoria dos conjuntos difusos

Em 1965, Lotfi Zadeh propôs a *Teoria dos Conjuntos Difusos* (FUZZY...) que deu origem à *Lógica Difusa*¹². Um conjunto clássico A definido em um dado domínio X , pode ser modelado utilizando-se uma função característica, que associa o grau 1 a todos os elementos do domínio que pertencem ao conjunto e o grau 0 aos demais. Esta função característica pode ser definida como um mapeamento $A : X \rightarrow \{0, 1\}$.

Na *Teoria dos Conjuntos Difusos*, utiliza-se uma função de pertinência para modelar os conjuntos difusos, como um mapeamento $A : X \rightarrow [0, 1]$. Um elemento com grau de pertinência 1 (respectivamente 0) é completamente compatível (respectivamente incompatível) com o conceito expresso pelo conjunto difuso; valores de pertinência entre 0 e 1 indicam compatibilidade parcial. Na literatura mais antiga da área, utiliza-se usualmente o símbolo μ_A para denotar a função de pertinência de um conjunto difuso A . Neste trabalho optamos por utilizar o tipo de notação mais recente, visando maior clareza na leitura.

A partir de um conjunto difuso A em X , obtemos conjuntos clássicos, através dos cortes de nível (ou α – cuts) definidos como na equação 2.1, para $\alpha \in (0, 1]$. A *cardinalidade* de um conjunto difuso A é dada pela equação 2.2, quando a função de pertinência é discreta.

$$A_\alpha = \{x \in X / A(x) \geq \alpha\} \quad (2.1)$$

$$|A| = \sum_{x \in X} A(x) \quad (2.2)$$

¹Os termos Teoria dos Conjuntos Difusos e Lógica Difusa são usados no texto de forma intercambiável.

²No Capítulo 6, o termo difuso também é utilizado como referência a um padrão de desmatamento da floresta Amazônica.

Um conjunto difuso é dito convexo se sua função de pertinência é convexa. Uma função f de $[a, b]$ em R é dita convexa se o conjunto $(x, y) \in R^2 \mid y \geq f(x)$ for um conjunto convexo. Isto equivale a afirmar que, para quaisquer x e y pertencentes a $[a, b]$ e para todo $t \in [0, 1]$, tem-se a relação mostrada na equação 2.3 (MAS-COLELL et al., 1995).

$$f(tx + (1 - t)y) \leq tf(x) + (1 - t)f(y) \quad (2.3)$$

2.1.1 Operadores lógicos difusos

Na *lógica difusa*, as *t-normas* e *t-conormas* são os operadores usados para modelar os operadores “e” e “ou” (conjunções e disjunções) da *lógica clássica*. Um operador $\top : [0, 1]^2 \rightarrow [0, 1]$ é chamado uma t-norma se é comutativo, associativo, monotônico e seu elemento neutro é 1. Um operador $\perp : [0, 1]^2 \rightarrow [0, 1]$ é chamado uma t-conorma se é comutativo, associativo e monotônico com elemento neutro igual a 0. Podemos citar as operações mínimo e produto como exemplo de t-normas e as operações de máximo e soma limitada como exemplo de t-conormas. A Tabela 6.2 indica as t-normas e t-conormas mais utilizadas e as Figuras 2.1 e 2.2 ilustram alguns destes operadores, em relação a um exemplo com dois conjuntos difusos.

Tabela 2.1 - Principais T-normas e t-conormas Fonte: (SANDRI; CORREA, 1999)

T-normas	T-conormas	Nome
$\min(a, b)$	$\max(a, b)$	Zadeh
$a \cdot b$	$a + b - ab$	Probabilística
$\max(a + b - 1, 0)$	$\min(a + b, 1)$	Lukasiewicz
$\begin{cases} a & \text{se } b = 0; \\ b & \text{se } a = 0; \\ 0 & \text{senão.} \end{cases}$	$\begin{cases} a & \text{se } b = 0; \\ b & \text{se } a = 0; \\ 1 & \text{senão.} \end{cases}$	Weber

Uma das formas de modelar o operador de implicação da lógica clássica (\rightarrow) no contexto da lógica difusa consiste no uso de um operador de implicação residual ϕ_T , baseado em uma t-norma T , definido como na equação 2.4.

Alguns exemplos muito conhecidos destes operadores incluem:

- implicação de Gödel, resíduo de $T = \min$, definido como $a^* \rightarrow b = 1$ se $a \leq b$ e $a^* \rightarrow b = b$, caso contrário;

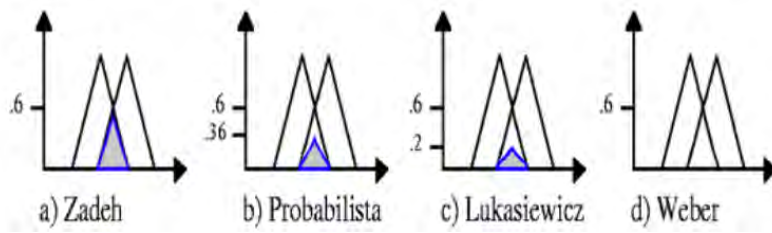


Figura 2.1 - Principais t-normas

Fonte: (SANDRI; CORREA, 1999)

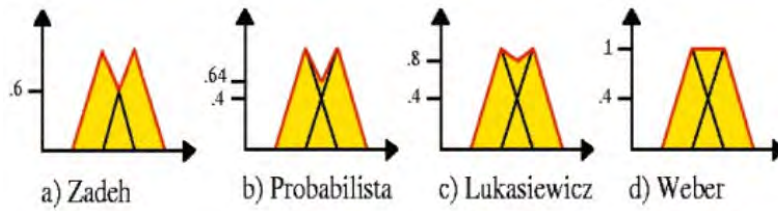


Figura 2.2 - Principais t-conormas

Fonte: (SANDRI; CORREA, 1999)

- implicação de Goguen, definido como $a^* \rightarrow b = 1$ se $a \leq b$ e $a^* \rightarrow b = b/a$ caso contrário;
- implicação de Lukasiewicz, definido como $a^* \rightarrow b = \min(1 - a + b, 1)$.

$$\phi_T(a, b) = a^* \rightarrow_T b = \sup_{c \in [0,1]} T(a, c) \leq b \quad (2.4)$$

2.1.2 Relações difusas

Na *lógica clássica*, relações são conjuntos definidos em um universo multidimensional $X = \{X_1 \times \dots \times X_n\}$. Como o próprio nome indica, uma relação implica na presença ou ausência de associação entre elementos dos diferentes universos de discurso que compõem X . Na *lógica difusa*, uma relação difusa é modelada por um conjunto difuso, na qual a função de pertinência indica o grau de associação entre elementos de X (DUBOIS et al., 1998). Em uma relação difusa R , cada ênupla $(x_1, \dots, x_n) \in X$ está associada a um grau de pertinência entre 0 e 1, i. e. $R : X \rightarrow [0, 1]$.

Se R_1 e R_2 são relações difusas, respectivamente em $X \times Y$ e em $Y \times Z$, a composição de R_1 e R_2 , denotada por $R_1 \circ R_2$, resulta em uma nova relação que associa

diretamente X a Z . A composição sup^* pode ser definida conforme a equação 2.5.

$$R_1 \circ R_2(x, z) = sup_{y \in Y} [R_1(x, y) * R_2(y, z)] \quad (2.5)$$

onde $x \in X$, $z \in Z$ e $*$ é uma t-norma. A composição mais usual é a *sup-min*.

Uma *relação de proximidade difusa* S em um domínio A é um mapeamento $S : \Omega \times \Omega \rightarrow [0, 1]$, que atribui a cada par ordenado (w, w^*) de elementos de A um valor que mede o quanto w e w^* são similares. Estas relações foram originalmente introduzidas por (ZADEH, 1971) como uma generalização da definição clássica de *relações de equivalência*.

Segundo (RUSPINI et al., 1998), (DUBOIS et al., 1998), pode-se dizer que uma relação binária difusa S em um universo A é uma relação de similaridade em A se as propriedades de simetria ($S(x, y) = S(y, x)$), reflexividade ($S(x, x) = 1$) e transitividade ($S(x, y) * S(y, z) \leq S(x, z)$) forem satisfeitas para todo $x, y, z \in X$, onde $*$ é uma t-norma.

Quando somente a reflexividade e a simetria são obedecidas, a relação é usualmente chamada *relação de proximidade*. No entanto, na literatura também se utiliza o termo *relações de proximidade* para relações difusas nas quais somente a transitividade não é necessariamente satisfeita (GODO; SANDRI, 2002).

2.2 Raciocínio baseado em casos

O Raciocínio Baseado em Casos (RBC), surgiu como uma técnica para solução automática de problemas e consiste em utilizar um conjunto de soluções anteriores, com ou sem alterações dentro de um determinado domínio, para solucionar novos problemas (ABEL, 1996).

A definição clássica de um sistema RBC foi elaborada por (RIESBECK; R.C., 1989): “Um sistema RBC resolve problemas, adaptando soluções que foram utilizadas para resolver problemas anteriores”.

Dentre as características do funcionamento de um sistema RBC estão:

- A extração do conhecimento a partir de casos ou experiências com que o próprio sistema se depara.
- A identificação das características mais significativas dos casos apresentados

a fim de devolver uma melhor solução (resposta).

- O armazenamento do caso e sua respectiva solução.

De acordo com (AAMODT; PLAZA, 1994), o RBC, de uma forma generalizada, pode ser dividido nas etapas explicadas abaixo e ilustradas pela Figura 2.3.

- Recuperação: a partir da apresentação ao sistema de um novo problema é feita a recuperação na base de casos daquele mais parecido com o problema em questão. Isto é feito a partir da identificação das características mais significativas em comum entre os casos;
- Reuso: a partir do caso recuperado é feita a reutilização da solução associada àquele caso. Geralmente a solução do caso recuperado é transferida ao novo problema diretamente como sua solução;
- Revisão: é feita quando a solução não pode ser aplicada diretamente ao novo problema. O sistema avalia as diferenças entre os problemas (o novo e o recuperado), quais as partes do caso recuperado são semelhantes ao novo caso e podem ser transferidas adaptando assim a solução do caso recuperado da base à solução do novo caso;
- Retenção: é o processo de armazenar o novo caso e sua respectiva solução para futuras recuperações. O sistema irá decidir qual informação armazenar e de que forma.

Uma vez que existe disponível uma base de dados para ser investigada, é possível investigá-la para extrair conhecimento a ser aplicado na tomada de novas decisões. A construção de um sistema de RBC a partir de uma base de dados passa pela definição de técnicas e formas de implementação de cada um dos componentes do sistema. Conforme (WANGENHEIM; WANGENHEIM, 2003), as etapas mais importantes do processo de desenvolvimento de um sistema RBC são:

- Aquisição de Conhecimento
- Representação de Caso
- Indexação
- Recuperação de Casos

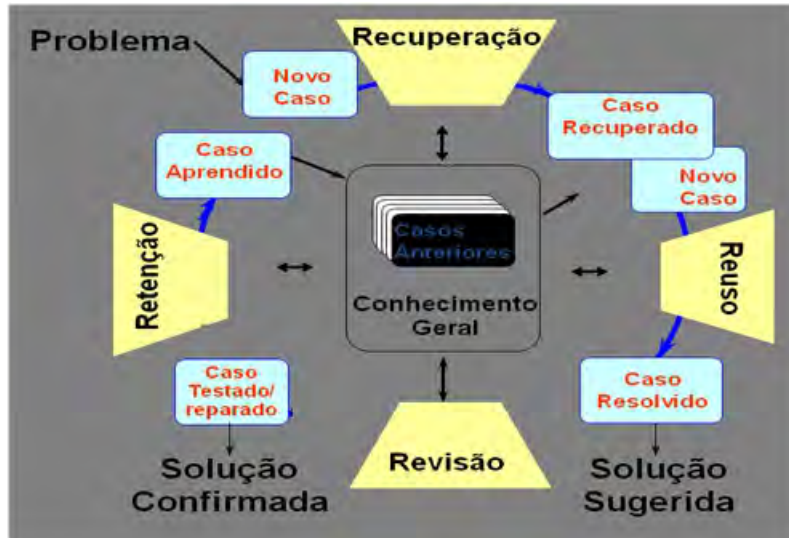


Figura 2.3 - Ciclo de funcionamento de sistemas RBC
 Fonte: (AAMODT; PLAZA, 1994)

- Adaptação de Casos.

O uso da metodologia de RBC e sua aplicação em ambientes de aprendizagem são embasados por uma ampla teoria cognitiva: o processo de lembrar, como fenômeno na resolução de problemas. O processo de reutilizar episódios passados corresponde a uma forma frequente e poderosa do raciocínio humano.

A utilização dessa técnica fica limitada apenas ao acesso às bases de dados completas, corretas e confiáveis que contenham entre as informações armazenadas, a descrição completa de problemas e das soluções que foram aplicadas em algum momento, pois esta é a matéria prima inicial e básica para a construção de sistemas baseados em casos.

2.3 Hipergrafos

Um hipergrafo é uma generalização de um grafo não direcionado, onde arestas podem conectar um número qualquer de vértices (BERGE, 1973). Formalmente, isto pode ser representado como um par, $H = (N, E)$, onde N é um conjunto de vértices e E é um conjunto de subconjuntos não vazios de N chamadas hiperarestas. O conjunto de hiperarestas E é então um subconjunto de $2^N - \{\emptyset\}$, onde 2^N é o conjunto potência de N .

Um “grafo comum” é então um hipergrafo no qual todas hiperarestas têm no máximo

2 elementos. Por outro lado, dado um hipergrafo $H = (N, E)$, uma hiperaresta $A \in E$ é dita ser máxima quando $\nexists B \in S, B \neq A/A \subset B$. Um exemplo de hipergrafo pode ser visto claramente na figura 2.4 onde temos $V = \{v_1, v_2, v_3, v_4, v_5, v_6\}$, $E = \{e_1, e_2, e_3, e_4\} = \{\{v_1, v_2, v_3\}, \{v_2, v_3\}, \{v_3, v_5, v_6\}, \{v_4\}\}$.

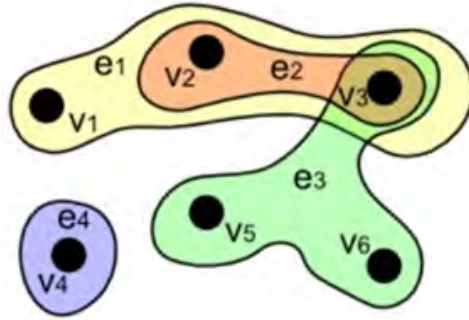


Figura 2.4 - Exemplo de hipergrafo
Fonte: (AZEREDO et al., 2008)

3 METODOLOGIA DE CLASSIFICAÇÃO

Este capítulo apresenta a metodologia de agrupamentos difusos utilizados neste trabalho, proposta originalmente em (FANOIKI et al., 2010) e posteriormente estendida em (SANDRI, 2012) (vide também (SANDRI et al., 2012)).

Na abordagem de agrupamentos difusos, cria-se primeiramente uma relação difusa entre os casos da base. Toma-se então um corte de nível desta relação, gerando assim uma relação clássica. Usando esta relação clássica, são obtidos os agrupamentos, a partir dos quais são calculadas soluções para um novo problema. Uma solução única é obtida a partir das soluções calculadas para cada agrupamento.

3.1 Definições básicas

Um caso c é definido como um par ordenado $c = (p, o) \in P \times O$ onde p é a descrição do problema e o sua solução, sendo $P = \{P_1 \times \dots \times P_n\}$ e O o espaço de descrição do problema e o espaço de soluções, respectivamente. Considera-se aqui que a solução de um caso é modelada por uma única variável, ao contrário da descrição de um caso, que usualmente depende de diversas variáveis (atributos).

Dado um conjunto de variáveis de descrição de problema V , cria-se uma relação de proximidade S_k para cada variável $v_k \in V$, cada qual definida em um domínio P_{v_k} . Estas relações podem ser obtidas, por exemplo, variando-se o parâmetro λ da família de relações R_λ , conforme equação 3.1, com.

$$R_\lambda(a, b) = \max(0, 1 - \frac{|a - b|}{\lambda}) \quad (3.1)$$

Seja $S_{in} \subseteq P^2$ uma relação difusa reflexiva sobre todo o espaço do problema, que mede a similaridade entre as descrições de problema de dois casos da base. S_{in} pode ser obtido a partir do conjunto de relações de proximidade $\{S_1, \dots, S_N\}$, correspondentes às variáveis que descrevem o problema. Sejam $c_k = (p_k, o_k)$ e $c_j = (p_j, o_j)$ dois casos pertencentes a C , com $p_k = (p_{k1}, \dots, p_{kn})$ e $p_j = (p_{j1}, \dots, p_{jn})$, respectivamente. Foram consideradas duas maneiras para calcular S_{in} , uma não ponderada (equação 3.2) e outra ponderada (equação 3.3).

$$S_{in}(p_k, p_j) = \frac{1}{n} \sum_{l=1}^n S_l(p_{kl}, p_{jl}) \quad (3.2)$$

$$S_{in}^w(p_i, p_j) = \sum_{k=1}^n w_k \times S_k(p_{ik}, p_{jk}) \quad (3.3)$$

Define-se também uma relação de proximidade $S_{out} : O^2 \rightarrow [0, 1]$ no espaço de soluções, fazendo-se $S_{out}(o_1, o_2) = S_{vo}(o_1, o_2)$, onde S_{vo} é a relação de proximidade obtida para a variável que modela a solução. A relação S_{out} mede a similaridade das soluções de dois casos da base.

3.2 Relações de semelhança difusa

Seja ϕ um operador de implicação residual. Define-se uma relação de semelhança difusa (FCRR) $F_\phi : C^2 \rightarrow [0, 1]$ pela equação 3.4 (SANDRI et al., 2012).

$$F_\phi(c_1, c_2) = \begin{cases} 0, & \text{se } S_{in}(p_1, p_2) = 0 \\ \phi(S_{in}(p_1, p_2), S_{out}(o_1, o_2)), & \text{senão} \end{cases} \quad (3.4)$$

A relação é uma formalização gradual do princípio básico do RBC: “problemas semelhantes tem soluções semelhantes”. Em particular, a similaridade entre dois casos e é considerada nula quando a similaridade entre as descrições de problema e é nula e/ou quando a similaridade entre as soluções e é nula. ¹.

3.3 Obtendo as relações de semelhança clássica

Uma relação de casos de semelhança difusa F_ϕ não é necessariamente clássica. Como a abordagem de agrupamento baseia-se em uma relação de casos clássica, toma-se um corte de nível da relação difusa, como descrito na equação 3.5.

$$\forall \alpha \in (0, 1], F_{\phi, \alpha}(c_i, c_j) = \begin{cases} 1, & \text{se } F_\phi(c_i, c_j) \geq \alpha \\ 0, & \text{senão} \end{cases} \quad (3.5)$$

Cada $F_{\phi, \alpha}$ é chamada uma relação de semelhança clássica entre casos, do inglês *Crisp Case Resemblance Relation* (CCRR). Como há infinitos valores no intervalo $[0, 1]$, existe um número infinito de CCRRs $F_{\phi, \alpha}$ derivadas a partir de F_ϕ . No entanto, como o número de casos na base é finito, existe um conjunto finito de FCRRs relevantes, dado por $Z = \{\alpha \in (0, 1] / F_\phi(c_1, c_2) = \alpha\}$.

¹A primeira propriedade decorre obviamente da primeira condição de 3.4. A segunda é consequência de uma propriedade dos operadores residuados ($\phi(a, 0) = 0$ quando $a > 0$), vinda da lógica clássica.

Os dois valores extremos para α são relevantes: α_{\min} e $\alpha = 1$. Na formulação original de (FANOIKI et al., 2010), uma CRRR S_{res} é definida diretamente (sem a criação da relação difusa) como $S_{res}(c_a, c_b) = 1$ se e somente se $0 < S_{in}(p_a, p_b) \leq S_{out}(o_a, o_b)$. Esta abordagem é um caso particular da abordagem difusa, já que $S_{res} = F_{\phi,1}$ para qualquer operador residuado ϕ . No outro extremo, tem-se $\alpha_{\min} = \inf Z$. Pode-se provar que $F_{\phi, \alpha_{\min}}(c_1, c_2) = \min(S_{in}(p_a, p_b), S_{out}(o_a, o_b))$, para qualquer operador residuado ϕ . Em (SANDRI et al., 2012), $F_{\phi,1}$ e $F_{\phi, \alpha_{\min}}$ são denotados por $F_{\phi, \uparrow}$ e $F_{\phi, \downarrow}$, respectivamente. Nos experimentos descritos neste trabalho, os melhores resultados obtidos foram aqueles utilizando $F_{\phi, \downarrow}$.

3.4 Usando atributos ponderados

A solução para um agrupamento e a força do agrupamento podem ser determinadas usando operadores ponderados. É possível associar-se um peso a cada caso de maneira que os considerados mais importantes para uma determinada aplicação tenham pesos mais elevados. Além disso, pode-se associar um vetor individual de pesos nas variáveis a cada caso o que recai em ponderar as variáveis independentemente dos casos. Mas, pode-se atribuir vetores de pesos individuais para cada caso, de modo que mais atributos significativos no interior de um caso recebam pesos mais elevados. Em (ARMENGOL et al., 2004) e (MARTINS-BEDÊ et al., 2009), relações de proximidade difusa associadas a cada variável de descrição e solução foram utilizadas para derivar vetores de pesos individuais, através do algoritmo de aprendizagem proposto em (TORRA, 2000). Foi mostrado que o uso de vetores de pesos individuais para cada caso do conjunto de treinamento tende a levar a melhores resultados que a utilização de casos não ponderados.

Muitas funções de agregação, como médias, t-normas e t-conormas, possuem versões ponderadas. Por exemplo, usando a abordagem com vetores de pesos atribuídos a cada um dos casos, o operador de agregação média ponderada é dado pela Equação 3.6 onde w é um vetor de pesos definido para todo k , $w_k \in [0, 1]$ e $\sum_k w_k = 1$.

$$S_{means}^w(p_i, p_j) = \sum_k w_k \times S_k(p_{ik}, p_{jk}) \quad (3.6)$$

Uma versão ponderada de S_{in} pode ser utilizada para calcular os agrupamentos. Utilizando um vetor de pesos para cada caso, a relação resultante é possivelmente assimétrica, que no entanto pode ser transformada em simétrica (FANOIKI et al., 2010). No presente trabalho, as relações clássicas não sofreram transformações para

se tornar simétricas.

3.5 Obtendo agrupamentos para uma base de casos

Sejam $c_a = (p_a, o_a)$ e $c_b = (p_b, o_b)$ dois casos em C . Seja $R = F_{\phi, \alpha}$ uma FCRR obtida como um corte de nível, para um ϕ e um $\alpha \in (0, 1]$ dados.

Baseado na relação de semelhança clássica R , os casos em C podem ser organizados em conjuntos de agrupamentos, formando hipergrafos. Dizemos que um hipergrafo $H = (C, E)$, $E \subseteq C^2$ é compatível com a FCRR R quando obedece as seguintes condições (SANDRI et al., 2012):

- $\forall c_a, c_b \in C$, se $R(c_a, c_b) = 1$, então $\exists h \in E$ tal que $(c_a, c_b) \subseteq h$
- $\forall c_a, c_b \in C$, se $R(c_a, c_b) = 0$, então $\nexists h \in E$ tal que $(c_a, c_b) \subseteq h$

Em outras palavras, se dois casos estão relacionados através de R , existirá pelo menos uma hiperaresta em E que os contém. Por outro lado, se dois casos não estão relacionados, eles não estarão contidos numa mesma hiperaresta de E .

É fácil provar que o hipergrafo obtido a partir de R que contém somente arestas maximais, é compatível com C . Este hipergrafo é dado por $H_{\max} = (C, E_{\max})$, onde $E_{\max} = \{A \in E / \nexists B \subseteq E, B \neq A \wedge A \subseteq B\}$.

Neste trabalho, foram somente utilizados os hipergrafos de arestas maximais de cada base de casos. Outras metodologias podem ser utilizadas para obtenção de agrupamentos, partindo de uma relação R em C . Pode-se ainda obter conjuntos de agrupamentos cujo hipergrafo não é necessariamente compatível com R , como por exemplo através de redes neurais artificiais (vide 4.1).

3.6 Calculando uma solução para um novo problema de acordo com um agrupamento

Dada uma base de casos C , medidas de proximidade S_j escolhidas para cada variável v_j , medidas de proximidades globais S_{in} e S_{out} e um hipergrafo $H = (C, E)$ compatível com $R = F_{\phi, \alpha}$ para um operador residual ϕ e um valor $\alpha \in (0, 1]$, a questão é como calcular uma solução o^* apropriada para um novo problema p^* . Em (FANOIKI et al., 2010) e (SANDRI et al., 2012), esta solução é calculada a partir dos casos contidos nos agrupamentos cujas descrições de problema são, de alguma maneira, similares a p^* , conforme denotado na equação 3.7. Para cada $h = \{c_1, c_2, \dots, c_r\} \in E^*$, é calcu-

lada a solução para p^* , denotada por o_h^* , usando uma função de agregação apropriada que leva em conta tanto o conjunto de soluções o_i quanto a similaridade entre cada p_i e p^* , considerando os casos (p_i, o_i) em h . Por exemplo, se a função de agregação é a média ponderada e as similaridades são agregadas usando S_{in} , utiliza-se a equação 3.8.

$$E^* = \{h \in E \mid \forall c_i = (p_i, o_i) \in h, S_{in}(p_i, p^*) > 0\} \quad (3.7)$$

$$o_h^* = \sum_{i=1}^r \frac{S_{in}(p_i, p^*) \times o_i}{\sum_{i=1}^r S_{in}(p_i, p^*)} \quad (3.8)$$

3.7 Determinando a força de um agrupamento em relação a um novo problema

Seja O^* o conjunto de soluções para p^* obtidas pelos agrupamentos em E^* (vide seção 3.6). Para selecionar a solução final o^* a partir de O^* , pode-se agregar as soluções em O^* , ou pode-se simplesmente assumir como solução aquela gerada pelo agrupamento mais fortemente relacionado com p^* .

Neste trabalho, adotamos a opção onde os agrupamentos competem entre si para fornecer a solução final. A força de um agrupamento $h = \{c_1, c_2, \dots, c_r\}$ em E^* , onde $c_i = (p_i, o_i)$, em relação ao problema p^* , é calculada como na Equação 3.9 onde f é uma função de agregação adequada, como por exemplo uma média, uma t-norma ou uma t-conorma, por exemplo (FANOIKI et al., 2010).

$$str_f(h, p^*) = f(S_{in}(p_1, p^*), \dots, S_{in}(p_n, p^*)) \quad (3.9)$$

4 PROPOSTAS DE TIPOLOGIA DE EXPERIMENTOS E DE DETERMINAÇÃO DE AGRUPAMENTOS

Neste capítulo propomos o uso de redes neurais artificiais para determinação de agrupamentos. Propomos também uma tipologia para experimentos envolvendo vários aspectos da metodologia de agrupamentos difusa (vide Capítulo 3).

4.1 Cálculo de agrupamentos com Redes Neurais Artificiais

Para encontrar os agrupamentos em uma base de casos pode-se utilizar redes neurais artificiais (RNA).

Para realização deste trabalho efetuou-se um estudo utilizando a RNA Fuzzy-ART. Uma rede Fuzzy-ART gera agrupamentos de vetores de características difusos. Mais especificamente, segundo (SILVA, 2002), cada componente do vetor de entrada i é um valor de pertinência da função membro de uma determinada característica difusa, indicando o quanto esta característica está presente na amostra. Assim, a dinâmica de um sistema Fuzzy-ART é descrita em termos das operações da teoria de conjuntos difusos.

Apesar da rede ART ser uma rede não supervisionada, possui um mecanismo de controle do grau de similaridade que é função do parâmetro ρ (limiar de vigilância), cujo valor é especificado pelo usuário. Quando um novo padrão não se enquadra a qualquer grupo já existente, este mecanismo provoca a formação de um novo grupo, determinando se um novo padrão de entrada pode ser incluído em um dos agrupamentos.

Neste trabalho, o número de neurônios criados ao fim da etapa de treinamento da rede corresponde ao número de agrupamentos e cada caso é incluído ao grupo cujo vetor de pesos mais se assemelha a suas variáveis (??).

4.2 Tipologia para Treinamento

A abordagem apresentada em 3.1 tem dois aspectos principais, um se refere à associação de vetores de peso aos casos, proposta por Vicenç Torra (vide (TORRA, 2000)), e o outro ao uso de relações de semelhança entre casos, com a consequente criação de agrupamentos de casos, proposta inicialmente em (FANOIKI et al., 2010) e posteriormente generalizada em (SANDRI et al., 2012). Nesta seção são apresentados uma tipologia de experimentos que relaciona estes dois aspectos, descrita em (??), (??) e (SANDRI et al., 2012).

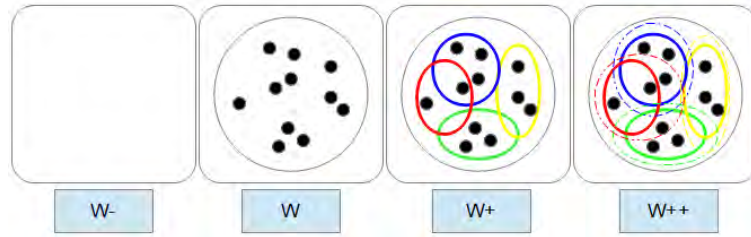


Figura 4.1 - Tipos de Treinamento

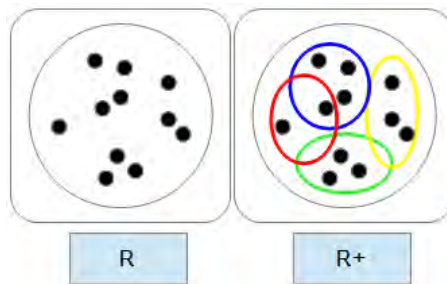


Figura 4.2 - Agrupamentos para cálculo das soluções

Uma vez determinados os agrupamentos em uma base, é necessário selecionar, para cada um deles, os casos da base que irão compor a base de treinamento daquele agrupamento, a partir da qual serão aprendidos os vetores de pesos para cada caso no agrupamento. Além disso, é preciso estabelecer como será calculada a solução de um novo problema, considerando-se somente um ou vários agrupamentos.

Os experimentos podem ser realizados sem utilização de pesos ($W-$), com pesos aprendidos para toda a base (W), ou com pesos aprendidos com a base fragmentada em agrupamentos ($W + eW + +$) (Figura 4.1). Por sua vez, podem ser calculadas soluções para cada agrupamento ($R+$), que podem então competir entre si pela solução final, ou serem agregadas em uma única solução, ou considerando-se somente um agrupamento, a própria base (R) como mostra a Figura 4.2.

Seis tipos de experimentos foram criados (Figura 4.3). As estratégias são descritas como segue, considerando-se a apresentação de um novo problema à base.

a) Uma única solução obtida, considerando-se toda a base (R):

- $W - R$: sem utilização de pesos e com o cálculo de uma única solução, obtida ao se considerar um único agrupamento, a própria base;

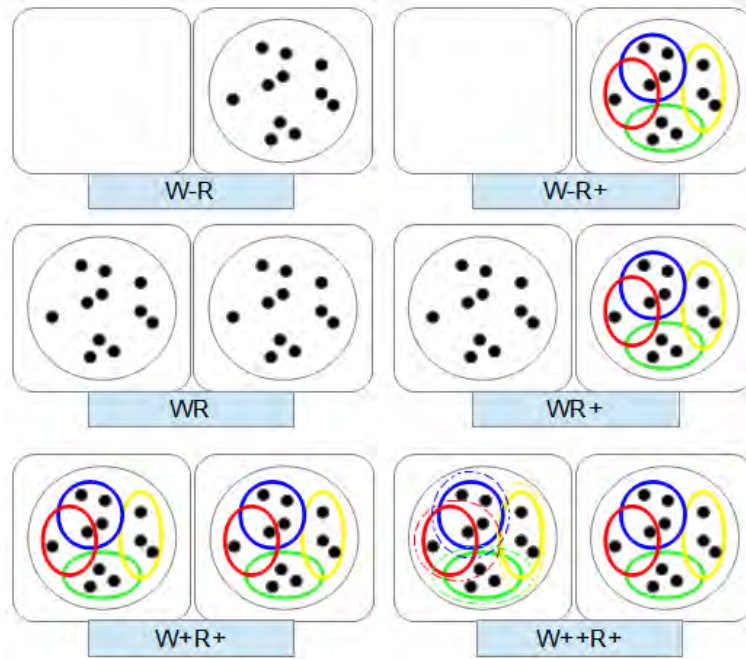


Figura 4.3 - Tipos de experimentos propostos

- WR : com utilização de pesos, aprendidos usando-se toda a base de casos como base de treinamento, e calculando-se uma única solução, obtida ao se considerar um único agrupamento (a própria base);
- b) Várias soluções obtidas, uma para cada um de vários agrupamentos ($R+$):
- $W - R+$: sem utilização de pesos e com o cálculo de várias soluções, cada qual obtida a partir de um dos vários agrupamentos;
 - $WR+$: com utilização de pesos, aprendidos usando-se toda a base de casos como base de treinamento, e com o cálculo de várias soluções, cada qual obtida a partir de um dos vários agrupamentos;
 - $W + R+$: com pesos aprendidos em cada um dos vários agrupamentos, considerando-se os próprios casos do agrupamento como base de treinamento, e com o cálculo de várias soluções, cada qual obtida a partir de um dos vários agrupamentos;
 - $W + +R+$: com pesos aprendidos em cada um dos vários agrupamentos, considerando-se os próprios casos do agrupamento mas também outros casos cuja descrição de problemas seja similar aos casos que compõem o

agrupamento, e com o cálculo de várias soluções, cada qual obtida a partir de um dos vários agrupamentos.

Para implementar a estratégia $W++R+$ é preciso encontrar meios para expandir os agrupamentos, i.e., determinar quais outros casos da base, além daqueles já pertencentes a um agrupamento, devem ser usados no aprendizado dos vetores de peso dos casos do agrupamento.

Queremos construir uma base de treinamento C^* para um agrupamento h^* , obtido a partir de uma base C . Uma abordagem aceitável consiste em incluir em C^* cada caso $c_0 = (p_0, o_0)$, cuja descrição de problema p_0 está de alguma maneira relacionada às descrições de problema dos casos em h^* . Pode-se implementar esta abordagem de pelo menos duas maneiras, com $C^{*\forall} = \{c_0 \in C / \forall c_i = (p_i, o_i) \in h^*, S_{in}(p_i, p_0) > 0\}$ e com $C^{*\exists} = \{c_0 \in C / \exists c_i = (p_i, o_i) \in h^*, S_{in}(p_i, p_0) > 0\}$. Na primeira opção, com $C^{*\forall}$, para fazer parte da base de treinamento de um agrupamento h^* , um caso c_0 que não pertença a h^* tem q ter sua descrição de problema similar às descrições de problema de todos os casos em h^* . Na segunda opção, com, basta que a descrição de problema de seja similar à descrição de problema de ao menos um componente de h^* . Obviamente, $h^* \subseteq C^{*\forall} \subseteq C^{*\exists}$, e portanto, a base de treinamento de $C^{*\exists}$ é maior que $C^{*\forall}$, que por sua vez é maior que o próprio h^* . Note que, como h^* não é vazio, também não o são $C^{*\exists}$ e $C^{*\forall}$.

Nos experimentos realizados em (SANDRI et al., 2012) e também no presente trabalho, adotou-se a primeira abordagem, ou seja, para que c_0 fosse incluído em C^* , este deveria apresentar um grau de similaridade maior do que 0 com todos os casos de h^* .

Contrariamente à $W+R+$, que implementa $C^* = h^*$, a opção $W++R+$ significa que a informação negativa (ou seja, casos que têm descrições de problemas semelhantes, mas soluções diferentes) é levada em conta para calcular os pesos em um agrupamento, o que intuitivamente deveria induzir a criação de vetores de pesos melhores.

Duas combinações, $W+R$ e $W++R$, não foram abordadas neste trabalho. Tanto em $W+$ como em $W++$, um conjunto de vetores de peso é associado a cada caso, sendo um vetor para cada agrupamento a que pertença o caso. Em R , uma solução é calculada considerando-se todos os casos da base. Portanto, para se criar experimentos do tipo $W+R$ e $W++R$, seria necessário primeiramente transformar os vários vetores associados a um dado caso da base em um único vetor. Aqui também

seria possível agregar os vários vetores de peso ou fazê-los competir entre si. Estas abordagens, por serem mais elaboradas, foram deixadas como trabalhos futuros.

5 ESTUDO DE CASO: PREVALÊNCIA DA ESQUISTOSSOMOSE

O estado de Minas Gerais possui 853 municípios, sendo que a prevalência de esquistossomose é conhecida para 197 deles (Figura 5.1). Nesta seção, apresenta-se um estudo de caso onde busca-se estimar a prevalência da esquistossomose para as cidades onde a prevalência não é conhecida, baseando-se nas características das cidades para as quais a prevalência é conhecida. Foram utilizados dados cedidos pela Secretaria do Estado de Minas Gerais que correspondem à prevalência de esquistossomose no Estado, apresentados originalmente em (GUIMARAES et al., 2005).

Neste estado, a distribuição da esquistossomose não é regular. 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).

5.1 Esquistossomose

A esquistossomose é uma doença produzida por trematódeos do gênero *Schistosoma* que, para o homem, tem como principais agentes etiológicos as espécies *S. mansoni*, *S. haematobium* e *S. japonicum* (KATZ; ALMEIDA, 2003).

Segundo (NEVES, 2001), a esquistossomose é uma patologia endêmica dos países

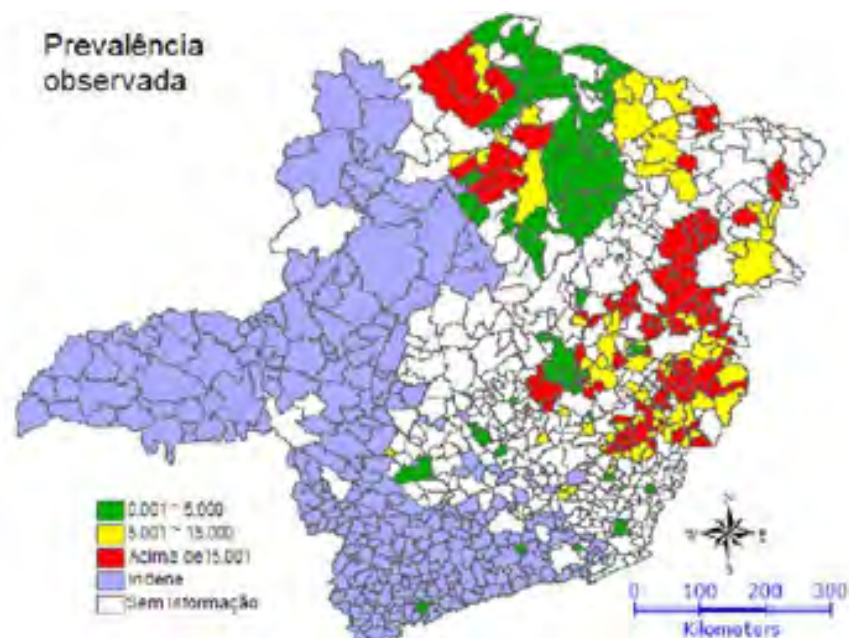


Figura 5.1 - Municípios mineiros cuja prevalência de esquistossomose é conhecida. Fonte: (MARTINS et al., 2008)

subdesenvolvidos ou em desenvolvimento. De acordo com (KATZ; ALMEIDA, 2003), a doença acomete 200 milhões de pessoas em 74 países. No Brasil, estima-se em cerca de seis milhões o número de infectados com a esquistossomose mansoni, que se concentram principalmente nos Estados do Nordeste e em Minas Gerais.

A doença instala-se basicamente por saneamento precário ou inexistente. As pessoas se contaminam através do contato com água natural infestada por cercárias que são eliminadas na água através de hospedeiros intermediários, moluscos límnicos do gênero *Biomphalaria* (*B.glabrata*, *B.tenagophila*, *B.straminea*) (DOUMENGE et al., 1987).

O ciclo biológico de transmissão da esquistossomose é descrito da seguinte forma. Os ovos do *S. mansoni* são eliminados pelas fezes do hospedeiro humano infectado e, se as fezes são lançadas nas coleções de água doce, eles eclodem liberando uma larva ciliada, denominada miracídio, responsável por infectar o hospedeiro intermediário. As larvas abandonam o caramujo e ficam livres na água, na forma de cercaria. Se o homem tiver contato com águas infectadas pelas cercarias, estas penetram ativamente, pela pele e mucosas, fazendo com que o indivíduo adquira a infecção. O homem infectado pode transmitir a doença por muitos anos, eliminando ovos de *S. mansoni* nas fezes (SILVA et al., 2005).

5.2 Experimentos originais

Nos experimentos originais (MARTINS et al., 2008), 86 variáveis independentes de vários tipos foram utilizadas para calcular a prevalência: variáveis de sensoriamento remoto (22), climáticas (6), socioeconômicas (34) e características de vizinhança (24). As variáveis de sensoriamento remoto são provenientes dos sensores MODIS (Moderate Resolution Imaging Spectroradiometer) e SRTM (Shuttle Radar Topography Mission), e supostamente relacionado ao tipo de habitat do caracol. As variáveis climáticas foram obtidas a partir do Centro Previsão de Tempo e Estudos Climáticos (CPTEC) do Instituto Nacional de Pesquisas Espaciais (INPE) e refletem as condições de sobrevivência de caramujos e as várias formas de larvas do *Schistosoma mansoni*. As variáveis socioeconômicas foram obtidas a partir do SNIU (Sistema Nacional de Indicadores Urbanos), tais como acesso a água potável e saneamento básico. As variáveis de características de vizinhança medem a disparidade entre os municípios vizinhos com relação as variáveis de renda, acesso a esgoto, educação, água e acúmulo de água.

Das 86 variáveis originais, um conjunto menor foi selecionado de acordo com testes



Figura 5.2 - Regionalização obtida através do algoritmo SKATER
 Fonte: (MARTINS et al., 2008)

utilizando regressão linear múltipla (MARTINS et al., 2008); as variáveis independentes escolhidas foram aquelas que apresentaram alta correlação com a variável dependente e baixa correlação com as outras variáveis. Duas abordagens principais foram utilizadas: i) uma global, em que todos os municípios com prevalência da doença conhecida foram utilizados, tanto para construção ou validação de um modelo de regressão linear e, ii) uma regional, em que o estado foi dividido em quatro regiões homogêneas e um modelo de regressão linear foi criado pra cada um deles.

O algoritmo SKATER (Assunção, 2006) foi utilizado para obter as regiões homogêneas no modelo regional; este algoritmo cria regiões de forma que as áreas vizinhas com características semelhantes pertencem à mesma região (Figura 5.2).

Nos experimentos relatados em (MARTINS-BEDÊ et al., 2009), foram utilizados conjuntos diferentes de variáveis independentes. Foram utilizadas 5 variáveis na abordagem global. Na abordagem regional, foram utilizadas 2 variáveis para a região R1, 5 para a região R2, 4 para a região R3 e 3 para a região R4. Em ambas as abordagens, global e regional, aproximadamente 2/3 das amostras foram utilizadas como conjunto de treinamento e 1/3 foi utilizado como conjunto de teste. As esti-

mativas de prevalência foram classificadas como baixa ($[0,5\%)$), média ($[5,15\%)$) e alta ($[15,100\%)$), como preconizado pela OMS.

Em (MARTINS-BEDÊ et al., 2009), utilizou-se árvore de decisão e regressão para nas abordagens global e regional, tanto para os dados de treinamento quanto para os dados de validação. Como a quantidade total de amostras da região R1 era muito pequena (16), os autores optaram por utilizar todas as amostras exclusivamente para treinamento. A Tabela 5.1 reproduz os resultados deste trabalho, tal como detalhados em (SANDRI et al., 2012). Podemos ver que, apesar de os resultados de treinamento para algumas regiões serem muito bons, os resultados do teste não são satisfatórios, com exceção da região R4.

Tabela 5.1 - Classificação, para conjuntos de treinamento e teste, com aprendizado: R-Reg (base regional e regressão), G-Reg (base global e regressão) e G-DT (base global e árvore de decisão). Fonte: (MARTINS-BEDÊ et al., 2009)

Treinamento	R1 (16)	R2 (59)	R3 (44)	R4 (28)
R-Reg	50,00%	42,37%	54,55%	57,14%
G-Reg	56,25%	54,24%	59,09%	60,71%
G-DT	62,50%	64,41%	77,27%	78,57%
Teste	R1 (0)	R2 (27)	R3 (14)	R4 (9)
R-Reg	-	37,04%	28,57%	77,78%
G-Reg	-	29,63%	42,86%	100,0 %
G-DT	-	29,63%	35,71%	44,44%

Na tabela 5.1 pode-se observar que os experimentos utilizando árvore de decisão apresentaram os melhores resultados para todas as regiões considerando-se apenas os casos de treinamento. Para os casos de teste, os melhores resultados foram obtidos a partir do uso de regressão.

Os dados usados em (MARTINS-BEDÊ et al., 2009) foram reutilizados com a abordagem difusa ponderada proposta em (FANOIKI et al., 2010), exposta no Capítulo 3. A Tabela 5.2 reproduz os primeiros resultados obtidos (MARTINS-BEDÊ et al., 2009), comparados aos resultados de (MARTINS-BEDÊ et al., 2009) para regressão, considerando-se a acurácia total (treinamento e teste). Verifica-se na tabela que os resultados da abordagem difusa foram equivalentes em alguns poucos experimentos, produzindo em geral resultados inferiores aos de regressão.

Tabela 5.2 - Classificação com os modos de aprendizado R-Reg (Base regional e regressão), G-Reg (Base global e regressão), R-Sim (Base regional e similaridade) e G-Sim (Base global e similaridade) Fonte: (MARTINS-BEDÊ et al., 2009)

Treinamento	R1 (16)	R2(59)	R3(44)	R4(28)
R-Reg	56,00%	51,00%	72,00%	76,00%
G-Reg	50,00%	40,00%	48,00%	59,00%
R-Sim	56,00%	56,00%	62,00%	38,00%
G-Sim	56,00%	49,00%	71,00%	65,00%

5.3 Experimentos e resultados utilizando a abordagem de proximidade difusa ponderada

Os experimentos apresentados nesta seção foram realizados usando as mesmas relações de proximidade utilizadas em (MARTINS-BEDÊ et al., 2009). As relações de proximidade (equação 3.1) foram calculadas através dos mesmos valores usados em (MARTINS-BEDÊ et al., 2009). A relação foi calculada agregando-se as relações através da média aritmética. Os agrupamentos foram obtidos como descrito na seção 3.4. A força dos agrupamentos foi determinada através de uma média aritmética entre a similaridade de entrada do problema a ser resolvido e os casos que compunham cada agrupamento. Os resultados obtidos para os casos de treinamento são apresentados nas Tabelas 5.3 (abordagem regional) e 5.4 (abordagem global).

Tabela 5.3 - Classificação para a abordagem regional para os experimentos W-R, W-R+, WR, WR+, W+R+, W++R+ (Casos de treinamento).

Experimentos	W-R	W-R+	WR	WR+	W+R+	W++R+
R1 (16)	72,72%	72,72%	36,36%	72,72%	72,72%	72,72%
R2 (59)	52,54%	71,18%	47,00%	50,00%	50,00%	60,00%
R3 (44)	54,54%	75,00%	61,00%	61,00%	63,00%	65,00%
R4 (28)	35,71%	89,28%	57,00%	71,00%	64,00%	71,00%

Os resultados obtidos para os casos de teste são apresentados nas Tabelas 5.5 (abordagem regional) e 5.6 (abordagem global).

A partir dos resultados apresentados nas Tabelas 5.3, 5.4, 5.5 e 5.6 observou-se que a abordagem regional apresenta melhores resultados. Outro ponto a se destacar é que os melhores resultados são alcançados, geralmente, quando a base de casos é dividida em agrupamentos. Comparando-se com os resultados apresentados em

Tabela 5.4 - Classificação para a abordagem global para os experimentos W-R, W-R+, WR, WR+, W+R+, W++R+ (Casos de treinamento)

Experimentos	W-R	W-R+	WR	WR+	W+R+	W++R+
R1 (16)	36,36%	72,72%	18,18%	36,36%	36,36%	36,36%
R2 (59)	18,76%	60,23%	37,60%	46,15%	62,50%	42,43%
R3 (44)	18,17%	65,63%	69,71%	48,80%	63%	65%
R4 (28)	59,53%	89,28%	57,00%	59,17%	53,33%	35,49%

Tabela 5.5 - Classificação para a abordagem regional para os experimentos W-R, W-R+, WR, WR+, W+R+, W++R+ (Casos de teste)

Experimentos	W-R	W-R+	WR	WR+	W+R+	W++R+
R1 (5)	40%	20%	40%	20%	40%	40%
R2 (27)	51,85%	48,15%	37,04%	48,15%	29,63%	51,85%
R3 (14)	42,86%	57,14%	50 %	71,42%	14,29%	57,14%
R4 (9)	33,33%	55,56%	33,33%	66,67%	66,67%	66,67%

Tabela 5.6 - Classificação para a abordagem global para os experimentos W-R, W-R+, WR, WR+, W+R+, W++R+ (Casos de teste)

Experimentos	W-R	W-R+	WR	WR+	W+R+	W++R+
R1 (5)	20%	20%	20%	20%	20%	20%
R2 (27)	18,51%	40,74%	29,63%	18,51%	37,04%	33,33%
R3 (14)	14,28%	42,86%	57,14%	64,29%	64,29%	57,14%
R4 (9)	55,56%	66,67%	33,33%	77,78%	55,56%	33,33%

(MARTINS-BEDÊ et al., 2009), usando as mesmas relações de proximidade, vemos que a abordagem apresentada neste trabalho gerou melhores resultados.

5.4 Extensões

Analisando os resultados apresentados na seção 5.3 observou-se que a melhor região para análise é a região R3 devido ao maior número de casos e uma melhor distribuição destes entre as classes que compõem o problema. Esta região foi então utilizada para analisar o efeito de outras relações de proximidade e métodos para cálculo de agrupamentos que pudessem ser usadas na metodologia proposta.

5.4.1 Fator de Coesão

Neste caso, para calcular a força do agrupamento foi aplicado um fator de coesão entre os casos de cada agrupamento gerando uma nova força de um agrupamento em

relação ao problema p^* (Equação 5.2). As mesmas relações utilizadas em (SANDRI et al., 2012) foram empregadas para cálculo das relações difusas.

O fator de coesão é calculado da seguinte forma:

$$coe(h, p^*) = \frac{1}{n} \sum_{i=1}^n S_{in}(c_i, p^*) \quad (5.1)$$

Nesta extensão a força de um agrupamento passa a ser:

$$str - coe_f = str_f * coe(h, p^*), \quad (5.2)$$

onde str_f é calculado conforme mostrado na Equação ?? e f é uma função de agregação adequada.

Em (??), a função f utilizada no fator de coesão foi a similaridade de entrada média entre todos os casos que compõem o agrupamento. Os resultados obtidos são apresentados na Tabela 5.7.

Tabela 5.7 - Classificação dos casos da região R3 para os experimentos W-R, W-R+, WR, WR+, W+R+, W++R+ utilizando fator de coesão. (Fonte: ??)

Experimentos	W-R	W-R+	WR	WR+	W+R+	W++R+
R3 (14)	42,86%	57,14%	42,86%	57,14%	64,29%	57,14%

Os resultados obtidos através desta abordagem foram semelhantes aos calculados sem a utilização do fator de coesão porém, para a abordagem W+R+ a melhora foi significativa.

5.4.2 Agrupamento utilizando RNA Fuzzy-ART

A RNA fuzzy-ART foi usada para cálculo dos agrupamentos. A rede foi criada com 3 neurônios iniciais e utilizou-se limiar de vigilância igual a 0,45. O valor do limiar de vigilância e dos demais parâmetros da rede foi ajustado baseando-se na literatura e em experimentações previamente realizadas. O número inicial de neurônios foi determinado baseando-se no número de classes que compõem a base de casos.

Foram realizados 3 treinamentos (T_1, T_2, T_3) para a RNA. No primeiro, a rede foi treinada com os atributos que formavam cada um dos casos. Depois, efetuou-se o

treinamento a partir da relação de proximidade Sin, considerando-se que os casos deveriam ser agrupados com base somente em sua similaridade de entrada. Por último, foi usada a relação que efetua uma combinação entre as similaridades de entrada e de saída. O número de agrupamentos para cada um dos treinamentos foi 4, 5 e 3 respectivamente. Ressaltando que, nestes experimentos, não utilizou-se o fator de coesão para cálculo da força dos agrupamentos.

A Tabela 5.8 mostra os resultados obtidos quando os agrupamentos foram calculados usando o rede Fuzzy-ART sem o fator de coesão entre os agrupamentos gerados.

Tabela 5.8 - Resultados obtidos quando os agrupamentos foram calculados utilizando a rede fuzzy-ART (Casos de teste) Fonte: (??)

Experimentos	W-R	W-R+	WR	WR+	W+R+	W++R+
T_1	42,86%	42,86%	42,86%	64,29%	42,86%	42,86%
T_2	42,86%	64,29%	42,86%	35,71%	50%	42,86%
T_3	42,86%	57,14%	42,86%	21,43%	35,71%	42,86%

A utilização rede neural artificial Fuzzy-ART se mostrou eficiente para calcular os agrupamentos e, esta abordagem, considera a informação negativa. Porém, sua utilização é mais indicada quando a base de casos é previamente conhecida, pois o processo de treinamento pode ser mais custoso e gera agrupamentos desnecessários caso seus parâmetros sejam mal ajustados. Observou-se também que os resultados, quando esta técnica foi utilizada para cálculo dos agrupamentos, não foram, em geral, tão bons quanto os obtidos quando hipergrafos foram utilizados para cálculo dos agrupamentos.

5.5 Análise

Para este trabalho, primeiramente, fez-se um estudo a respeito de como os cortes de nível nas relações de proximidade influenciariam no cálculo dos agrupamentos. Foram realizados experimentos com os cortes de nível 0.3, 0.6 e 0.9. O uso de cortes de nível gerou um aumento no número de agrupamentos. Este aumento pode ser benéfico para a classificação, porém, para as relações de agregação utilizadas neste trabalho, este aumento de agrupamentos fez com que o processo de classificação se tornasse inviável pelo conseqüente aumento do custo computacional.

As relações de proximidade usadas foram escolhidas a partir dos experimentos realizados. Além das relações descritas na Seção 5.3, para agregar as relações de entrada e

de saída foram testadas as relações: implicação de Gödel, implicação de Goguen e produto porém, o uso destas relações não gerou um melhor resultado no processo de classificação.

Observando os resultados apresentados nas Tabelas 5.5 e 5.6 verificou-se que os experimentos utilizando pesos apresentaram resultados, no mínimo, equivalentes aos experimentos com similaridade não ponderada. Outra constatação é que a divisão da base de casos em agrupamentos melhora os resultados obtidos.

A exceção é a região 2 cujo melhor resultado foi obtido utilizando similaridade não ponderada. Porém, através de uma análise mais detalhada, pode-se verificar que 26 dos 27 casos são classificados como pertencentes a classe Média, o que inevitavelmente coincide com a maioria dos casos que compõem a região.

Considerando-se como sucesso apenas a porcentagem de acertos pode-se verificar que a abordagem regional apresenta, em geral, melhores resultados que a abordagem global.

Adotando-se o mesmo critério e comparando-se as Tabelas 5.3, 5.4, 5.5 e 5.6 observou-se que, com exceção da região R4, os resultados obtidos através da abordagem proposta neste trabalho foram melhores que os apresentados em (MARTINS-BEDÊ et al., 2009).

6 ESTUDO DE CASO: PADRÕES DE DESMATAMENTO

O banco de dados de áreas desmatadas do PRODES (MELLO et al., 2002) fornece informações que possibilitam detectar os diferentes tipos de ocupação da terra. Localizada no estado do Pará, a Terra do Meio compreende os Municípios de São Félix do Xingu, Tucumã e Altamira (Figura 6.1).

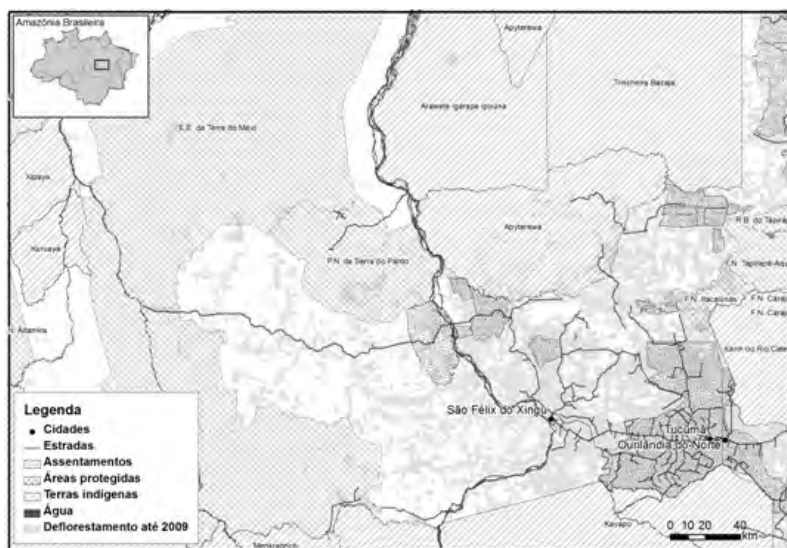


Figura 6.1 - Mapa da Terra do Meio
Fonte: (LOBO; ESCADA, 2010)

Foram coletados dados de desmatamento do PRODES para a Terra do Meio relativos aos anos de 1997, 2000, 2003, 2006 e 2009. Utilizando estes dados, alguns trabalhos foram realizados para análise dos padrões de ocupação da terra tais como em (LOBO; ESCADA, 2010), (??), (ESCADA; CARRIELO F., 2007) e (ESCADA M. I AND PINHO et al., 2010).

6.1 Tipos de Desmatamento

Técnicas de mineração de dados têm sido utilizadas para extrair informações de grandes bases de dados tais como do PRODES (SILVA et al., 2005) (??). Essas técnicas permitem explorar um conjunto de dados, extraindo ou ajudando a evidenciar padrões de interesse. (SILVA et al., 2005) (??) propuseram uma metodologia baseada em técnicas de mineração de dados para identificar diferentes padrões de desmatamento na Amazônia. Os autores utilizaram uma abordagem por objetos de análise representados pelos polígonos de desmatamento do PRODES. Em (LOBO; ESCADA,

2010) foi realizado um estudo para avaliar a evolução da ocupação da região da Terra do Meio no período de 1997 a 2009. Para isso, seu trabalho foi dividido em 4 etapas:

- reunião dos dados de desmatamento do PRODES para os anos de 1997, 2000, 2003, 2006 e 2009;
- definir, a partir dos mapas de desmatamento obtidos na primeira atividade, uma tipologia de padrões de ocupação representados por células;
- seleção de casos de treinamento representando todos os padrões de desmatamento, e utilizados na classificação de padrões da região Terra do Meio.
- análise dos padrões de ocupação e de sua evolução no tempo a partir da definição de trajetórias de padrões.

Para a análise da paisagem utilizando os dados de desmatamento pode-se trabalhar com objetos individuais, representados por cada um dos polígonos de desmatamento (??) ou com células onde cada uma é representada por um conjunto de polígonos de desmatamento (AZEREDO et al., 2008).

Na análise baseada em células a área de estudo é subdividida em pequenas regiões regulares. No espaço celular os polígonos são agregados em unidades maiores, e cada célula representa uma porção da paisagem. Assim, cada célula desta grade é associada a um padrão de desmatamento, onde cada padrão é descrito por um conjunto de métricas da paisagem. O tamanho da célula de 10 km X 10 km, adotado, foi o que melhor representou os padrões de ocupação observados conforme mostra a Figura 6.2. Este tamanho foi definido a partir da análise visual dos dados de desmatamento representados por células de diferentes tamanhos (LOBO; ESCADA, 2010).

Depois de definir do tamanho das células, foi elaborada a tipologia de padrões de ocupação. Cada padrão foi associado semanticamente a diferentes agentes sociais e estágios de ocupação da fronteira agropecuária. Dados de levantamentos de campo deram subsídios à definição da tipologia de padrões de ocupação (ESCADA; CARRI-ELLO F., 2007). Essa tipologia é apresentada na Figura 6.3 e descrita a seguir:

- Padrão Difuso ¹. Caracterizado por pequenas manchas isoladas de desmatamento. São áreas geralmente ocupadas por famílias que praticam a agri-

¹Nos capítulos anteriores o termo difuso também é utilizado como sinônimo de *Fuzzy Logic*

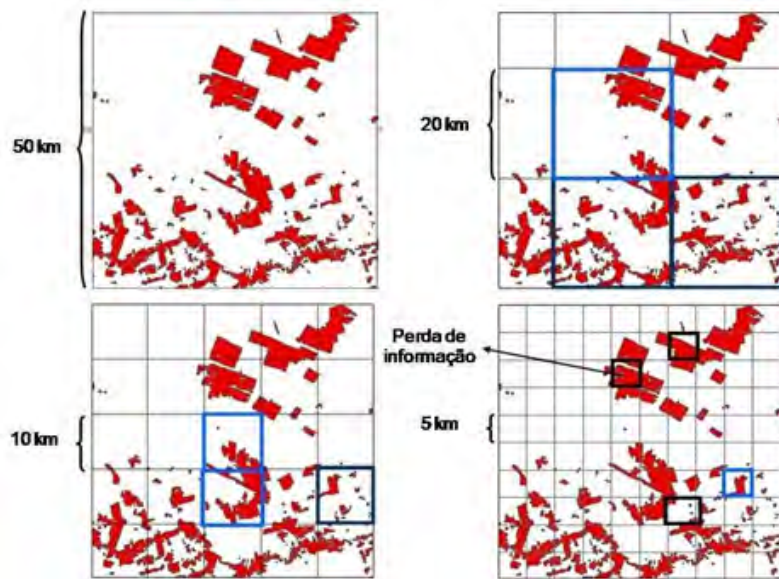


Figura 6.2 - Avaliação do tamanho das células para a definição de tipologia de padrões de ocupação.

Fonte: (AZEREDO et al., 2008)

cultura familiar ou pecuária para subsistência sem o uso de maquinários e tratores.

- Padrão Geométrico. Caracterizado por desmatamentos extensos, com forma regular, associados às grandes fazendas de gado, onde são utilizados maquinários e tratores. Esse padrão ocorre, muitas vezes, em áreas com pouco acesso as estradas.
- Padrão Multidirecional. Este padrão representa estágios intermediários a avançados de ocupação, onde ocorre intensificação do processo de desmatamento e de ocupação. Esse padrão se origina a partir dos padrões Difuso e Geométrico. Apresenta um ambiente com remanescentes florestais fragmentados e com diversos agentes sociais atuando simultaneamente.
- Padrão Consolidado. Representa estágio avançados de ocupação de uma maneira geral, quase toda área da célula é desmatada e ocupada por pastagens, restando poucos fragmentos florestais.

Além destes padrões, foram mapeadas áreas de floresta, representadas por células que não apresentaram nenhuma alteração de sua cobertura florestal.

	Padrão	Descrição (1:100.000)	Semântica
	Difuso	<ul style="list-style-type: none"> Manchas pequenas Manchas isoladas Baixa a média densidade Distribuição uniforme 	<ul style="list-style-type: none"> Início da ocupação Ocupação espontânea (não Planejada) Pequenos produtores rurais Ocupação Ribeirinha (+ Rios)
	Geométrico Grande	<ul style="list-style-type: none"> Manchas médias a grandes e isoladas Forma geométrica regular Baixa a média densidade 	<ul style="list-style-type: none"> Estágios iniciais de ocupação Fazendas Médias e Grandes
	Multidirecional Desordenado	<ul style="list-style-type: none"> Manchas pequenas a médias que se uniram Formas variadas (irregular, geométrica, linear) Média a alta densidade Multidirecional 	<ul style="list-style-type: none"> Ocupação em expansão, inicialmente espontânea Pode haver concentração fundiária Pequenos e médios produtores rurais
	Consolidado	<ul style="list-style-type: none"> Manchas grandes e contínuas Densidade baixa e áreas pequenas de remanescentes florestais Manchas compactas e contínuas 	<ul style="list-style-type: none"> Estágios avançados de ocupação Concentração fundiária Pequenos, médios e grandes produtores rurais Esgotamento da floresta Ocupação consolidada

Floresta
 Desmatamento

Figura 6.3 - Descrição de padrões de desmatamento e tipologia de ocupação identificados nas análises de imagens de satélites e trabalhos de campo.

Fonte: (LOBO; ESCADA, 2010)

6.2 Experimentos Originais

Para efetuar a classificação dos padrões, em (LOBO; ESCADA, 2010) foi utilizada uma árvore de decisões (Figura 6.4) considerando as métricas $Percent_{Land}$, LSI e MPS.

Segundo a árvore de decisão da Figura 6.4 a porcentagem de desmatamento acima de 69% ($Percent_{Land}$) separa o padrão Consolidado dos demais padrões. Este padrão representa o estágio mais avançado de ocupação, com maior proporção de área desmatada. Em seguida, o LSI, que mede a complexidade da forma dos polígonos de desmatamento e classifica as células que apresentaram LSI maior que 3,54 como padrão Multidirecional. Esse padrão representa, na maioria das vezes, ocupação não planejada com uma grande heterogeneidade de formas e tamanhos de manchas relacionadas aos diferentes tipos de produtores rurais. Esse padrão apresenta maior complexidade que o Difuso e o Geométrico. A métrica LSI apresenta valor próximo a 1 quando a maioria dos polígonos presentes na célula apresentam formas que se aproximam de retângulos ou de círculos, como é o caso dos padrões Geométrico (retângulo) e Difuso (círculos), assim, as células com LSI menor ou igual a 3,54 foram classificadas como padrão Geométrico ou Difuso. A discriminação desses dois padrões é feita com a métrica MPS (tamanho médio das manchas). O padrão Geométrico apresenta manchas de desmatamento maiores do que o padrão Difuso, assim, quando a célula apresenta MPS maior que 127 ha, esta é classificada como padrão

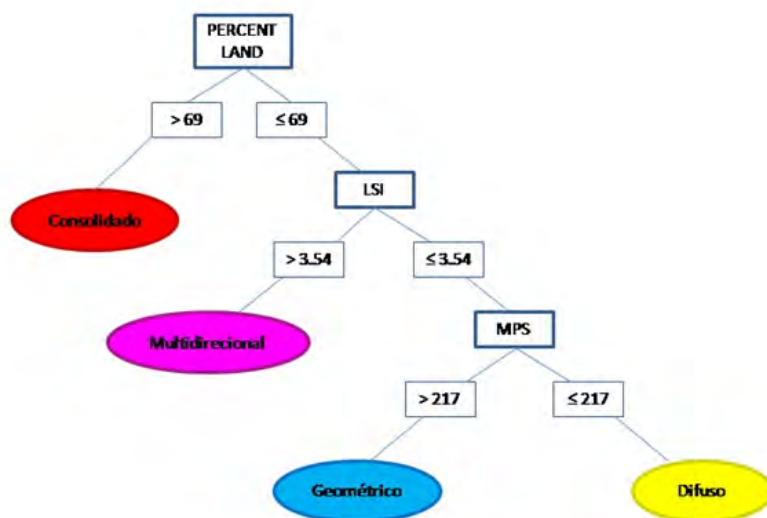


Figura 6.4 - Árvore de decisão utilizada para a classificação das células em padrões de ocupação nos anos de 1997, 2000, 2003, 2006 e 2009.

Fonte: (LOBO; ESCADA, 2010)

Geométrico. Caso contrário, atribui-se a ela a classe do padrão Difuso. Os resultados obtidos são mostrados na Figura 6.5 (LOBO; ESCADA, 2010).

Segundo (LOBO; ESCADA, 2010), a evolução temporal de cada um dos padrões de desmatamento analisados mostra que o padrão Difuso que representa os estágios iniciais de ocupação ocupou maior área (31%) quando comparado com os outros padrões, em todo período analisado (6.1), sendo um padrão onde grande parte da ocupação se dá de forma espontânea, não planejada. O padrão Geométrico variou entre 6 e 9% nas três primeiras datas atingindo um patamar de 11% das células entre 2003 e 2009. O padrão Multidirecional passou a representar 11% das células analisadas em 2009, enquanto que o padrão Consolidado que representava em 1997 apenas 1% da área total, aumentou para 6% em 2009 como pode ser observado na tabela 5.1.

Após a classificação das células nas cinco datas estudadas, (LOBO; ESCADA, 2010) sorteou aleatoriamente 100 células, de diversos anos, para a realização da classificação visual. Durante a seleção das amostras procurou-se garantir que todos os padrões fossem amostrados. (LOBO; ESCADA, 2010) observou ainda que a confusão foi maior entre os padrões Geométrico e Difuso. Isso pode ter ocorrido devido à heterogeneidade de algumas das células selecionadas que apresentaram polígonos relativos aos dois padrões de desmatamento, entretanto, o desempenho do classificador foi de 86%

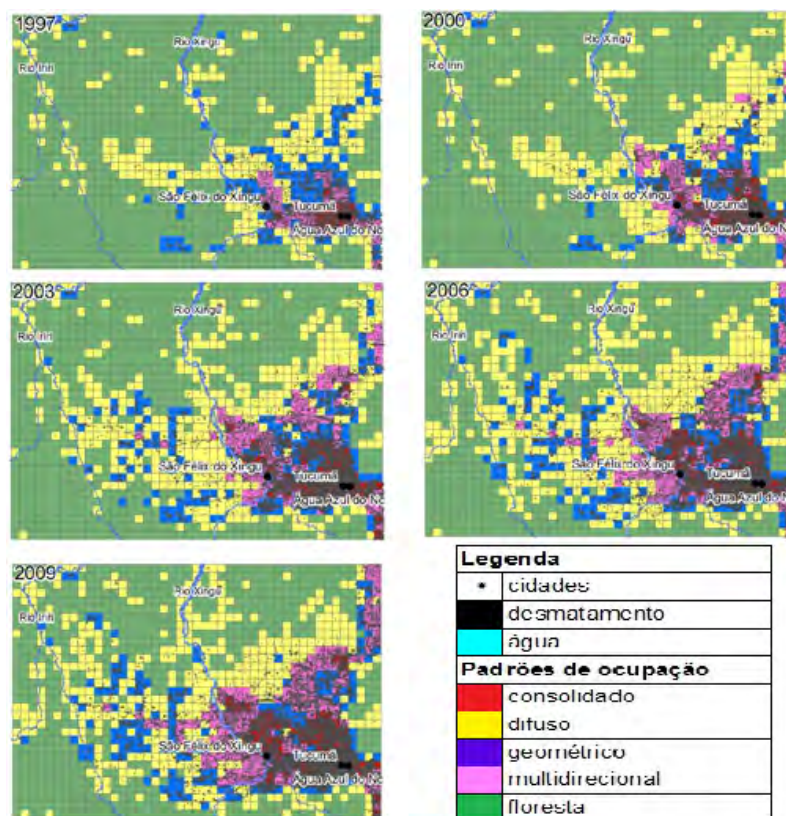


Figura 6.5 - Resultado da classificação por células da Terra do Meio nos anos de 1997, 2000, 2003, 2006 e 2009.

Fonte: (LOBO; ESCADA, 2010)

Tabela 6.1 - Evolução dos padrões de ocupação em % do total da área analisada entre 1997 e 2009. Fonte: (LOBO; ESCADA, 2010)

Ano	1997	2000	2003	2006	2009
Floresta	67%	63%	51%	43%	41%
Difuso	21%	23%	29%	32%	31%
Geométrico	8%	6%	9%	11%	11%
Multidirecional	3%	5%	7%	9%	11%
Consolidado	1%	3%	5%	6%	6%

de acerto conforme matriz de confusão apresentada na Tabela ??.

6.3 Experimentos e resultados utilizando a abordagem de proximidade difusa ponderada

A metodologia proposta neste trabalho foi aplicada a esta base de casos referentes aos anos de 1997, 2000, 2003, 2006 e 2009. Os casos foram divididos em treinamento

Tabela 6.2 - Matriz de confusão da validação da classificação por células utilizando o classificador GeoDMA. Fonte: (LOBO; ESCADA, 2010)

-	Difuso	Geométrico	Multidirecional	Consolidado
Difuso	35	3	2	0
Geométrico	2	19	1	2
Multidirecional	1	0	16	2
Consolidado	0	0	1	16

e teste de duas maneiras diferentes:

- Os mesmos casos para treinamento e teste utilizados em (LOBO; ESCADA, 2010). Ou seja, utilizou-se 20 casos, selecionados aleatoriamente por (LOBO; ESCADA, 2010), referentes a cada um dos anos que compreendem a base de casos. Os resultados obtidos são apresentados nas Tabelas 6.3 e 6.4.
- Em uma segunda etapa, os casos foram separados na seguinte proporção: 70% utilizados para treinamento e 30% para teste. Os casos foram selecionados aleatoriamente. Os resultados obtidos são apresentados nas Tabelas 6.5 e 6.6.

Em ambas abordagens os parâmetros selecionados foram os mesmos utilizados por (LOBO; ESCADA, 2010). Vale destacar que o número de casos pertencentes a cada classe pode variar de acordo com o ano analisado. Para calcular as relações de proximidade S_λ (equação 3.1) o valor de λ foi definido como $\lambda_i = 0,3(\max(v_{i,1}, \dots, v_{i,n}) - \min(v_{i,1}, \dots, v_{i,n}))$ onde $v_{i,j} \in \mathbb{N}$ são as variáveis analisadas e n é o número de casos que compõem a base de casos. A relação S_{in} foi calculada agregando-se as relações S_λ através da média aritmética. Os agrupamentos foram obtidos como descrito na seção 3.5. A força dos agrupamentos foi determinada pela similaridade de entrada máxima entre o problema a ser resolvido e os casos que compunham cada agrupamento.

Para realizar os experimentos, cujos resultados foram mostrados neste capítulo, utilizou-se as mesmas variáveis usadas por (LOBO; ESCADA, 2010). Utilizou-se um corte de nível com α próximo a 0 para cálculo da relação a partir da qual gerou-se os agrupamentos. A força de um agrupamento foi determinada através da operação máximo. Optou-se por estas medidas através de experimentos previamente realizados.

Tabela 6.3 - Resultados obtidos a partir das relações de proximidade difusas - Treinamento 1

Experimentos	W-R	W-R+	WR	WR+	W+R+	W++R+
1997 (432)	73%	100%	68%	100%	95%	98%
2000 (435)	66%	100%	63%	100%	100%	100%
2003 (573)	65%	100%	57%	95%	100%	95%
2006 (659)	63%	100%	61%	100%	100%	100%
2009 (670)	58%	100%	68%	100%	98%	100%

Tabela 6.4 - Resultados obtidos a partir das relações de proximidade difusas - Teste 1

Experimentos	W-R	W-R+	WR	WR+	W+R+	W++R+
1997 (20)	68%	90%	64%	75%	68%	72%
2000 (20)	70%	100%	66%	95%	95%	95%
2003 (20)	72%	100%	68%	94%	92%	90%
2006 (20)	60%	95%	54%	90%	86%	86%
2009 (20)	60%	100%	68%	85%	90%	82%

Tabela 6.5 - Resultados obtidos a partir das relações de proximidade difusas - Treinamento 2

Experimentos	W-R	W-R+	WR	WR+	W+R+	W++R+
1997 (315)	73%	100%	68%	100%	100%	100%
2000 (317)	66%	100%	65%	100%	100%	100%
2003 (415)	65%	100%	59%	100%	100%	100%
2006 (479)	63%	100%	61%	100%	100%	100%
2009 (484)	58%	100%	64%	100%	100%	100%

Tabela 6.6 - Resultados obtidos a partir das relações de proximidade difusas - Teste 2

Experimentos	W-R	W-R+	WR	WR+	W+R+	W++R+
1997 (137)	72%	91%	67%	78%	72%	76%
2000 (138)	65%	97%	65%	93%	92%	93%
2003 (178)	66%	93%	66%	91%	90%	89%
2006 (200)	65%	96%	63%	94%	90%	90%
2009 (206)	58%	98%	65%	83%	88%	82%

Neste estudo de casos foram realizados experimentos, além daqueles cujos resultado foram apresentados neste capítulo, calculando-se a força de cada agrupamento usando as operações mínimo, média aritmética, média geométrica e OWA. Além disso, realizou-se estudos acerca do uso de cortes de nível.

A metodologia apresentou uma taxa de acerto elevada na maioria dos experimentos realizados. Pode-se observar que o uso de agrupamento para cálculo das soluções foi o fator que mais colaborou para melhora dos resultados apresentados. A ponderação pouco influenciou nos resultados uma vez que tanto para os casos de treinamento quanto para teste os melhores resultados foram obtidos no experimento $W - R+$.

Os erros apresentados são aceitáveis pois as classes desta base de casos apresentam um grau de similaridade (cada uma tende a evoluir para uma outra em um determinado período de tempo).

7 QUALIDADE DE ESTIMATIVAS E TRATAMENTO DE DADOS TEMPORAIS

Existem várias maneiras de se medir a qualidade de um conjunto de resultados obtidos através da aplicação de um método sobre um conjunto de dados. Em tarefas de classificação, onde o sucesso é aferido pela capacidade do método de produzir exatamente o resultado esperado, as medidas mais utilizadas são a taxa de acerto e/ou erro, e o índice kappa. Em métodos que produzem como resultado estimativas modeladas em algum tipo de modelo de incerteza (por exemplo, distribuições de probabilidade, de possibilidade, etc), as medidas de qualidade são específicas ao modelo de incerteza adotado. Este capítulo traz um estudo sobre a qualidade dos resultados produzidos pela abordagem de agrupamentos utilizando relações difusas, apresentadas ao longo deste trabalho. Em particular, estuda-se o comportamento da medida de qualidade de estimativas possibilistas proposta em (SANDRI et al., 1997), aplicada aos dados de padrões de desmatamento da Terra do Meio (vide Capítulo 6). Além disso, o capítulo trata da qualidade da abordagem tendo em vista a evolução temporal dos dados e propõe outros índices para medir a qualidade das estimativas.

7.1 Medidas de qualidade de distribuições de possibilidade

7.1.1 Acurácia e Precisão

Nesta seção serão apresentados os conceitos utilizados para medir a qualidade dos resultados da abordagem de agrupamentos com relações difusas descrita neste trabalho.

Segundo (MIKHAIL; ACKERMAN, 1976) acurácia é o grau de proximidade de uma estimativa com seu parâmetro (ou valor verdadeiro), enquanto precisão expressa o grau de consistência da grandeza medida com sua média. Acrescenta-se ainda que a acurácia reflete a proximidade de uma grandeza estatística ao valor do parâmetro para o qual ela foi estimada e que precisão está diretamente ligada com a dispersão da distribuição das observações.

Por sua vez, uma estimativa deve ser o mais precisa possível para facilitar a tarefa de decisão, com escolha de um único valor final. Uma estimativa completamente imprecisa, embora acurada, pode ser de muito pouca utilidade. Por exemplo, uma estimativa dada na forma de uma distribuição de probabilidade uniforme relativa a uma dada variável aleatória x pode ser considerada acurada, pois a probabilidade do valor real de x , não é menor que a dos outros valores do domínio. No entanto,

como a entropia é máxima, o resultado é inútil. Uma boa medida de qualidade deve portanto levar em conta, tanto a acurácia quanto a precisão.

Em (SANDRI et al., 1997) foi proposta uma medida de qualidade para distribuições de possibilidade, que nada mais são que conjuntos difusos que estimam o valor de uma dada variável. A medida de qualidade de uma distribuição de possibilidade é o produto de sua acurácia e precisão. A acurácia é definida como o grau de pertinência da distribuição no valor conhecido da variável. A precisão, por sua vez, é definida como complemento da imprecisão, esta última tomada como a razão entre a cardinalidade da distribuição (área sob a curva) e a cardinalidade do domínio da variável. A abordagem é descrita em termos formais no que segue, usando-se os termos conjunto difusos e distribuição de possibilidade de maneira intercambiável. Uma estimativa perfeita neste formalismo ocorre quando o valor real da variável tem grau de pertinência 1 e todos os demais valores do domínio tem grau 0.

Seja uma variável x definida no universo X , cujo valor x^* é conhecido. Suponhamos que a estimativa do valor real de x é dado por um especialista e (por exemplo, um humano ou um sistema informático) como um conjunto difuso A , cuja função de pertinência $A : X \rightarrow [0, 1]$ é normalizada. A acurácia (acc), precisão (p) e qualidade (q) de uma estimativa possibilista são dadas por:

$$acc(e, x) = \mu_A(x^*) \quad (7.1)$$

$$p(e, x) = 1 - \frac{|A|}{|X|} \quad (7.2)$$

$$q(e, x) = acc(e, x) \cdot p(e, x) \quad (7.3)$$

O termo $\frac{|A|}{|X|}$ em 7.2 é baseado no índice de Jacquard, que mede a razão entre a cardinalidade da intersecção de dois conjuntos A e B e de sua união ($\frac{|A \cap B|}{|A \cup B|}$).

Por exemplo, suponhamos que tenhamos três estimativas distintas para (A_1, A_2, A_3) o valor da variável ω , definida em universo de discurso $\Omega = \{a, b, c, d\}$, definidas pelos conjuntos de pares ordenados $\{(a, 1), (b, .90), (c, .85), (d, .80)\}$, $\{(a, .10), (b, 1), (d, .50), (c, .25)\}$ e $\{(a, .30), (b, 1), (c, .65), (d, .50)\}$, respectivamente (vide Figura 7.2). Suponhamos que $\omega^* = a$.

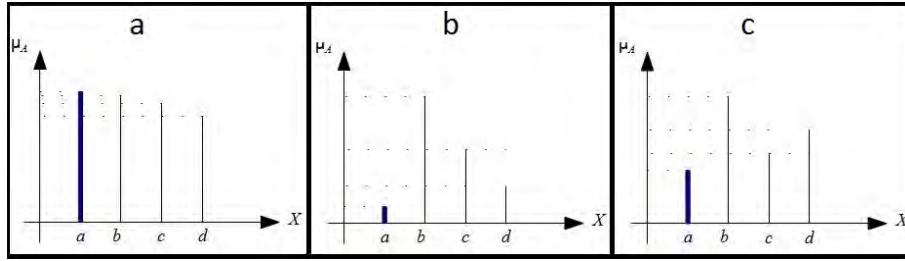


Figura 7.1 - Função de pertinência- a) A_1 b) A_2 c) A_3

Para A_1 temos $\text{acc}(A_1)=1$, $\text{prec}(A_1)=.1125$, $q(A_1) = .1125$, para A_2 temos $\text{acc}(A_2)=.1$, $\text{prec}(A_2)= .5375$, $q(A_2) = .0537$ e para A_3 temos $\text{acc}(A_3)=.3$, $\text{prec}(A_3)= .3875$ e $q(A_3) = .1162$. Verifica-se visualmente que as medidas de acurácia e precisão são adequadas. O conjunto difuso A_1 é mais acurado que A_2 , que por sua vez é mais acurado que A_3 . Já A_2 é mais preciso que A_3 , e ambos são mais precisos que A_1 . No entanto, o exemplo mostra que a medida de qualidade das distribuições pode nem sempre ser adequada. No exemplo, a qualidade de A_1 e A_3 são praticamente iguais. Em determinadas aplicações, pode valer a pena modificar a medida de qualidade, inclusive privilegiando a acurácia sobre a precisão ou vice-versa. Além disso, em algumas aplicações pode ser interessante levar em conta a convexidade da estimativa. Por exemplo, a estimativa A_3 não é convexa, ao contrário de A_1 e A_2 , mas esta característica não é considerada na medida de qualidade.

Quando tem-se um conjunto de variáveis $\{x_1, \dots, x_m\}$ definidas em $\{X_1, \dots, X_m\}$, pode-se obter, usando uma média aritmética, medidas de acurácia, precisão e de qualidade globais conforme Equações 7.4, 7.5 e 7.6, respectivamente.

$$A(e) = \frac{1}{m} \sum_{j=1}^m \text{acc}(e, x_j) \quad (7.4)$$

$$P(e) = \frac{1}{m} \sum_{j=1}^m p(e, x_j) \quad (7.5)$$

$$Q(e) = \frac{1}{m} \sum_{j=1}^m q(e, x_j) \quad (7.6)$$

No presente trabalho, foi utilizado a formulação acima com uma pequena modificação. A precisão foi modificada para fazer com que uma estimativa perfeita tivesse

sua medida de precisão igual a 1. A medida de precisão utilizada foi então:

$$p(\{e, x\}) = \left(1 - \frac{|A|}{|X|}\right) \frac{|X|}{|X - 1|} = \frac{|X| - |A|}{|X - 1|} \quad (7.7)$$

No entanto, como a função original foi multiplicada por uma constante, a relação de ordem entre estimativas para uma variável permanece a mesma. Porém, no caso de termos um conjunto de variáveis $\{x_1, \dots, x_m\}$ definidas em universos de discurso de tamanhos diferentes $\{X_1, \dots, X_m\}$, esta mudança pode influenciar na ordem de qualidade global dos especialistas.

7.2 Aplicação da Metodologia

A partir dos experimentos realizados e descritos no Capítulo 6, sobre padrões de desmatamento, foram obtidas funções de similaridade entre os casos e as classes que compõem o estudo realizado. A Figura 7.2 mostra as funções obtidas em alguns dos experimentos realizados.

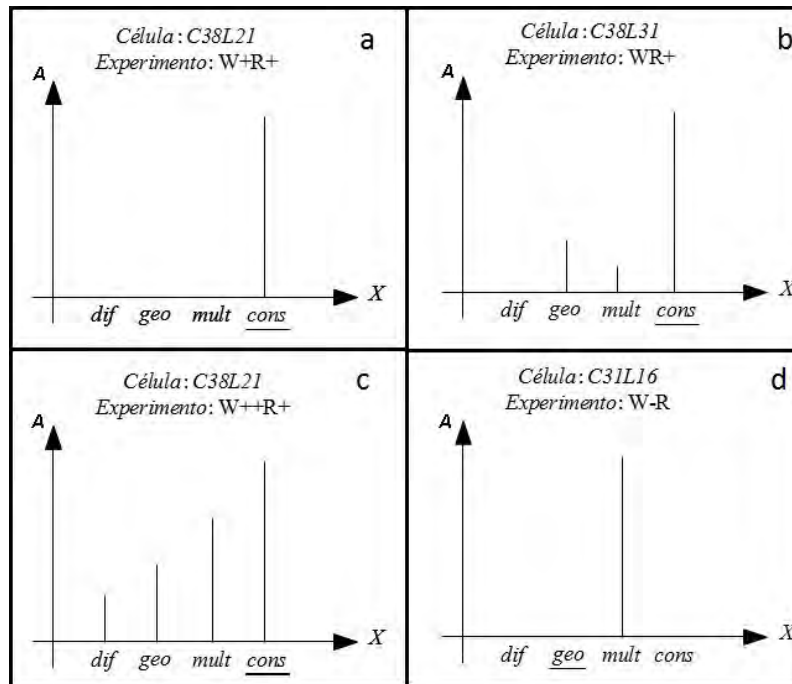


Figura 7.2 - Funções de similaridade a) $W + R+$ b) $WR+$ c) $W + +R+$ d) $W - R$

O melhor resultado possível ($q = 1$) é o apresentado na Figura 7.2a e o pior ($q = 0$) o apresentado na Figura 7.2d. A Figura 7.2b mostra um resultado acurado, preciso

e não convexo enquanto a Figura 7.2c apresenta uma classificação acurada, imprecisa e convexa. Para abordagem de medida de qualidade utilizada neste trabalho, a convexidade não está sendo considerada. Portanto, o resultado apresentado em 7.2b seria dito melhor que o visto em 7.2c.

Após calcular as medidas de qualidade para cada um dos anos foram medidas a acurácia, a precisão e a qualidade médias de cada um dos experimentos efetuados. As Tabelas 7.1, 7.2 e 7.3 mostram, respectivamente, a acurácia, a precisão e a qualidade global das classificações para cada um dos anos analisados considerando-se apenas os casos de teste. Aqui, considerou-se apenas os experimentos cujos resultados foram apresentados na Tabela 6.6.

Tabela 7.1 - Acurácia global média da classificação efetuada no segundo estudo de caso

Ano/Exp	W-R	W-R+	WR	WR+	W+R+	W++R+
1997	0,89	0,98	0,82	0,98	0,97	0,98
2000	0,86	0,98	0,79	0,98	0,98	0,99
2003	0,86	0,99	0,80	0,98	0,97	0,97
2006	0,73	0,98	0,82	0,97	0,98	0,97
2009	0,78	0,97	0,83	0,98	0,96	0,98

Tabela 7.2 - Precisão global média da classificação efetuada no segundo estudo de caso

Ano/Exp	W-R	W-R+	WR	WR+	W+R+	W++R+
1997	0,23	0,48	0,13	0,48	0,35	0,34
2000	0,24	0,41	0,16	0,39	0,46	0,48
2003	0,24	0,43	0,24	0,46	0,42	0,44
2006	0,16	0,42	0,25	0,38	0,38	0,43
2009	0,17	0,43	0,27	0,40	0,40	0,39

Tabela 7.3 - Qualidade global média da classificação efetuada no segundo estudo de caso

Ano/Exp	W-R	W-R+	WR	WR+	W+R+	W++R+
1997	0,17	0,48	0,08	0,49	0,35	0,35
2000	0,20	0,41	0,12	0,40	0,46	0,48
2003	0,21	0,42	0,12	0,40	0,45	0,43
2006	0,22	0,42	0,22	0,38	0,38	0,43
2009	0,14	0,43	0,25	0,40	0,36	0,39

Analisando os resultados obtidos observou-se que a acurácia foi elevada para todos os anos analisados e em todos os experimentos realizados. Comparando com os resultados apresentados no Capítulo 6 vemos que os resultados são condizentes pois, a taxa de acerto foi elevada.

Observa-se também que a precisão não foi tão elevada, o que prejudicou as medidas de qualidade apresentadas pelo sistema de classificação. A baixa precisão pode ser explicada, em partes, pelo grau de similaridade natural que ocorre entre as classes que compõem o estudo de casos.

A partir das análises de qualidade verifica-se, assim como nas análises anteriores, que os experimentos cujos resultados foram calculados a partir de uma base fragmentada geram os melhores resultados.

Uma outra forma de se avaliar a qualidade da classificação realizada é medir a qualidade média da classificação de um caso específico para cada experimento ao longo do tempo. A Tabela 7.4 mostra a qualidade média obtida para dois casos distintos da base de casos.

Tabela 7.4 - Qualidade média por experimento

Caso/Exp	W-R	W-R+	WR	WR+	W+R+	W++R+
C37L22	0,34	0,39	0,27	0,43	0,42	0,48
C29L17	0,26	0,43	0,38	0,34	0,32	0,30

A partir da Tabela 7.4 podemos observar que a qualidade média variou entre 0,277 e 0,481 para o caso C37L22. Para o caso C29L17, esta variação ocorreu entre 0,261 e 0,427. Nesta análise também observou-se que o uso de agrupamentos para cálculo dos resultados gerou melhores resultados.

Porém, este tipo de medição não foi aplicada pois, neste trabalho, as classificações foram realizadas ano a ano fazendo com que os atributos de um caso em um período de tempo $t + 1$ não fossem levados em consideração em um período de tempo t e vice-versa.

7.3 Convexidade como medida de qualidade de distribuições de possibilidade

Suponhamos que o universo de discurso Ω da variável ω acima seja completamente ordenado, i.e. $a < b < c < d$. Por exemplo, suponhamos que ω se refere à quantidade de ovos que Hans comerá em seu café da manhã (de 0 a 3), segundo as opiniões de três sistemas baseados em casos e_1 , e_2 e e_3 . Estes sistemas, ou especialistas, forneceram como resultado as distribuições A_1 , A_2 e A_3 . Um exame visual mostra que, ao contrário de A_1 e A_2 , a estimativa A_3 não é convexa. Ou seja, o sistema e_3 considera que existe uma maior possibilidade de Hans comer 1 ou 3 ovos, mas não 2. Se se considera aceitáveis somente estimativas em torno de um valor, a qualidade da estimativa de e_3 deveria ser penalizada pelo fato de A_3 não ser convexa. Este é o caso por exemplo de aplicações onde existe um componente de evolução temporal, como desflorestamento ao longo dos anos.

Propõe-se introduzir o uso do conceito de convexidade para medir a qualidade de uma classificação, comparando a cardinalidade da estimativa com o invólucro convexo (*convex hull*) desta estimativa conforme mostrado no Capítulo 8. A figura 7.3 mostra o invólucro convexo da estimativa A_2 .

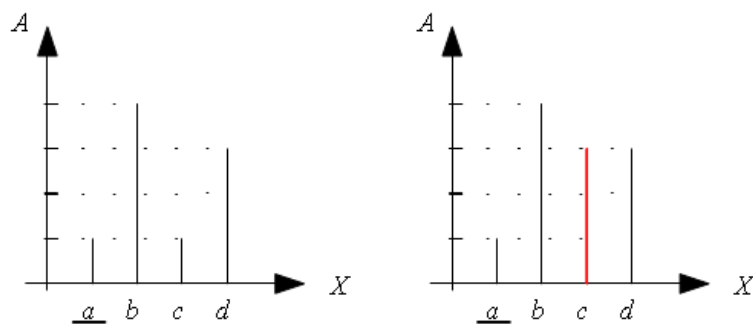


Figura 7.3 - Invólucro convexo da função de pertinência da estimativa

Para medir a qualidade da classificação, de maneira similar ao índice de precisão, pode-se tomar o índice de *Jacquard*. Seja $A_e : X \rightarrow [0, 1]$ uma função qualquer (não necessariamente uma função de pertinência a um conjunto difuso), fornecida por um especialista e como estimativa do valor de uma variável x em X . Temos então:

$$\text{conv}(A_e) = \frac{|A_e|}{|\text{hull}(A_e)|} \quad (7.8)$$

onde $hull(A)$ é o invólucro convexo de A . Quando uma estimativa coincide com seu invólucro convexo, $conv(e, x) = 1$, como esperado.¹

A medida de qualidade q , dada na Equação 7.8, pode então ser estendida como q' :

$$q'(e, x) = acc(e, x) \times p(e, x) \times conv(e, x) \quad (7.9)$$

Por sua vez, pode ser usada para se obter uma medida de qualidade global de um especialista, nos moldes da equação 7.6.

Para o estudo de casos, referente ao desmatamento, o conceito de convexidade poderia ser aplicado a análise da qualidade da classificação realizada. O grau de pertinência de um caso a um conjunto de classes pode ser uma função convexa ou não (Figura 7.4).

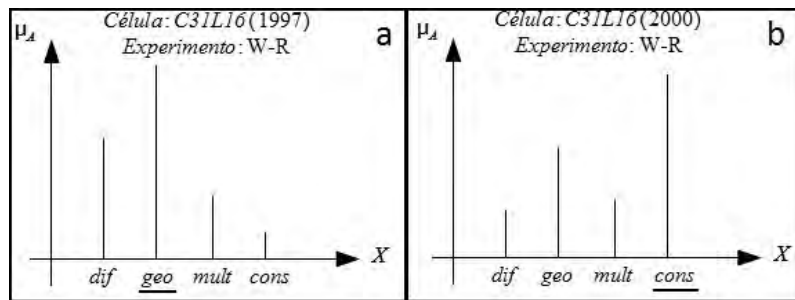


Figura 7.4 - Função de similaridade caso C31L16 a) 1997 b) 2000

Observa-se na figura que a classificação obtida pelo caso C31L16 (base de casos dos padrões de desmatamento) foi acurada e imprecisa para ambos os anos. Porém, ela é convexa para o ano 1997 e não convexa para o ano 2000. No primeiro caso a medida de qualidade apresentada na seção 7.1 seria multiplicada por 1, o que não interferiria na qualidade calculada. No segundo caso (ano 2000), a qualidade seria multiplicada por um valor .

Especificamente para o estudo de casos referente aos padrões de desmatamento, para fazer uma análise temporal das classificações obtidas, as classes difuso e geométrico deveriam ser consideradas como uma única classe. Esta junção ocorre pois ambas as classes representam um estado inicial de desmatamento provenientes de tipos de

¹A função $conv$ é aqui definida como a convexidade de um conjunto difuso. No entanto, ela pode ser usada para qualquer função A .

intervenções diferentes. Mas, ambas evoluem para a classe multidirecional.

8 MEDIDA DE QUALIDADE DE EVOLUÇÃO TEMPORAL

O uso de outras medidas pode ser interessante em aplicações que envolvem o conceito de evolução temporal, como por exemplo, a aplicação em padrões de desmatamento abordada no capítulo 6. Uma medida de consistência entre a evolução temporal real e uma estimativa (não difusa), poderia se basear na soma da diferença ponto a ponto entre a função que modela a realidade e a que modela a estimativa. Uma medida de consistência nestes moldes teria que ser normalizada para que a melhor estimativa tivesse grau máximo. Mas uma função assim nem sempre seria adequada, sobretudo em aplicações onde a evolução temporal esperada é monotônica.

Na Seção 8.1 estudamos algumas medidas gerais de inconsistência para contradomínios nos naturais e na Seção 8.2, estudamos medidas para aplicações onde a evolução temporal é necessariamente monotônica. Por fim, as medidas de qualidade propostas foram aplicadas ao estudo de casos apresentado no Capítulo 6 e os resultados obtidos são apresentados na Seção 8.3.

8.1 Medidas gerais de consistência de evolução temporal

Seja Ω um subconjunto dos números naturais, representando um conjunto de classes e seja ω a variável de classificação. Seja $T = \{t_1, \dots, t_n\}$ uma escala temporal, onde cada t_i indica um momento no tempo e seja $B : T \rightarrow \Omega$, uma função que modela a evolução de uma dada variável x , como por exemplo a classe do padrão de desmatamento de uma dada célula.

Sejam $B^* : T \rightarrow \Omega$ e $B_e : T \rightarrow \Omega$, as funções que modelam a realidade e a estimativa dada por um especialista, respectivamente. No restante do texto usaremos também a notação $B = (b_1, \dots, b_n)$, para denotar a sequência temporal $\langle B(t_1), B(t_2), \dots, B(t_n) \rangle$, onde $b_i = B(t_i)$.

Um possível índice de inconsistência temporal seria dado por *inctg1*¹ :

$$inctg1(e, x) = \sum_{t \in T} | B^*(t) - B_e(t) | \quad (8.1)$$

No entanto, esta medida não é capaz de distinguir as estimativas B_\bullet e B_\blacktriangle , mostradas na Figura 8.1, fornecidas por dois especialistas e_\bullet e e_\blacktriangle , respectivamente. Se a oscilação entre os resultados não é considerada desejável, a estimativa B_\bullet deveria ser

¹O termo *inctg1* é um acrônimo para medida de inconsistência temporal geral.

considerada superior à B_{\blacktriangle} , mesmo tendo o mesmo valor para $inctg1$. Uma possível forma de tratar este problema seria aliar um índice de convexidade à medida de consistência temporal.

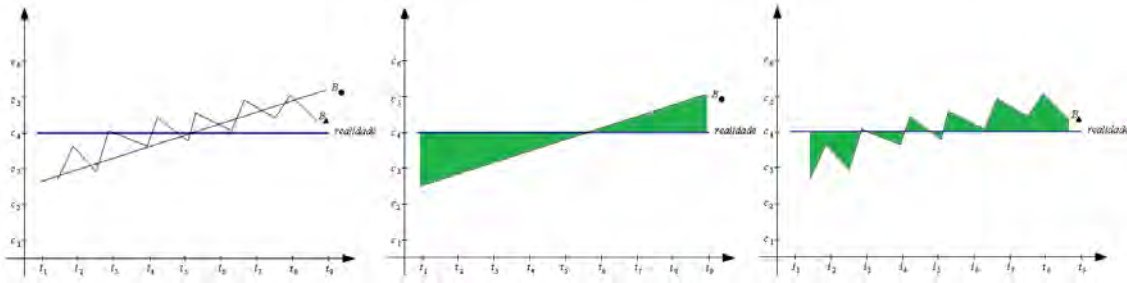


Figura 8.1 - Estimativas B_{\bullet} e B_{\blacktriangle}

Um segundo índice de inconsistência temporal $inctg2$ seria dado por:

$$inctg2(e, x) = \sum_{t \in T} | B^*(t) - B_e(t) | \times conv(B_e) \quad (8.2)$$

Esta medida seria capaz de distinguir entre as estimativas B_{\bullet} e B_{\blacktriangle} , penalizando B_{\blacktriangle} . No entanto, esta medida nem sempre é muito sensível.

Outras funções podem ser utilizadas no lugar da convexidade para medir suavidade, como a regularização de 2ª ordem de Tikhonov (A.N.TIKHONOV; V.S.ARSENIN., 1977). A regularização de 2ª ordem de Tikhonov, de um um vetor de R posições p é definida por:

$$\Omega(p) = \sum_{i=2}^{R-1} (p(i-1) + p(i+1) - 2p(i))^2 \quad (8.3)$$

Esta função é mínima quando p descreve uma reta, ou seja quando a suavidade é máxima. Usando a formalização acima, esta regularização seria então dada por

$$tik(B_e) = \sum_{i=2}^{R-1} (B_e(t-1) + B_e(t+1) - 2B_e(t))^2 \quad (8.4)$$

com $R = |T|$.

Uma terceira medida de consistência temporal seria dada então por

$$inctg3(e, x) = \sum_{t \in T} (B^*(t) - B_e(t)) \times tik(B_e) \quad (8.5)$$

Suponhamos que estamos interessados na evolução temporal em uma aplicação por um período de 6 anos, entre 1991 e 1996, com $T = \{91, 92, 93, 94, 95, 96\}$, e que o contradomínio é dado por $\Omega = \{1, 2, 3, 4, 5\}$.

Seja $B^* = (1, 1, 2, 3, 3, 5)$ a evolução real e $B_\bullet = (1, 2, 1, 4, 2, 5)$ e $B_\blacktriangle = (2, 2, 2, 2, 4, 5)$ duas estimativas (vide Figura 8.2). Neste exemplo, elas são indistinguíveis em relação à inconsistência temporal medida por $inctg1$. No entanto, a estimativa B_\blacktriangle é suave, tal como a evolução real, enquanto que B_\bullet oscila.

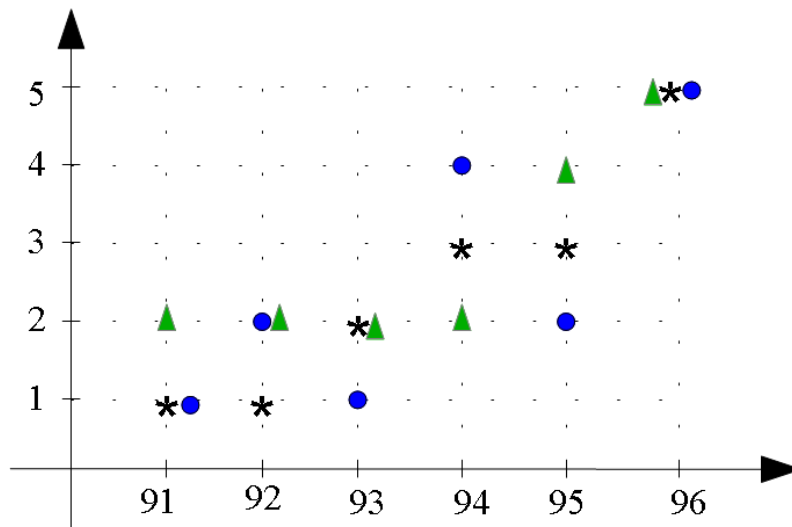


Figura 8.2 - Estimativas B^* , B_\bullet e B_\blacktriangle

Para calcular $inctg2$, temos que obter o invólucro convexo de cada estimativa para poder calcular sua medida de convexidade. Aqui temos $hull(B_\blacktriangle) = B_\blacktriangle = (2, 2, 2, 2, 4, 5)$ e $hull(B_\bullet) = (1, 2, 2, 4, 4, 5)$. Calculando, obtemos $conv(B_\bullet) = 15/18 = .83$ e $conv(B_\blacktriangle) = 17/17 = 1$, e temos então $inctg2(e_\bullet, x) = 18/5 = 4.8$ e $inctg2(e_\blacktriangle, x) = 4$. Portanto, a estimativa B_\blacktriangle é considerada melhor que B_\bullet , já que minimiza a inconsistência em relação à realidade tal como medida por $inctg2$.

Tomemos agora o período de 1991 a 1998. Seja $B^* = (2, 2, 2, 3, 3, 3, 4, 5)$, $B_\bullet = (3, 2, 2, 4, 2, 4, 3, 5)$ e $B_\blacktriangle = (2, 3, 2, 2, 2, 2, 3, 5)$ (vide figura 8.3). Usando estes vetores

obtemos $inctg1(e_{\bullet}, x) = 5 = inctg1(e_{\blacktriangle}, x)$, $inctg2(e_{\bullet}, x) = 5.95$ e $inctg2(B_{\blacktriangle}, x) = 6$. Ou seja, embora a estimativa B_{\bullet} apresente um maior volume de oscilações, ela pouco se distingue de B_{\blacktriangle} . Isto se deve a que a medida de convexidade é muito influenciada por pequenos vales muito longos entre dois picos, ainda que estes últimos sejam pequenos.

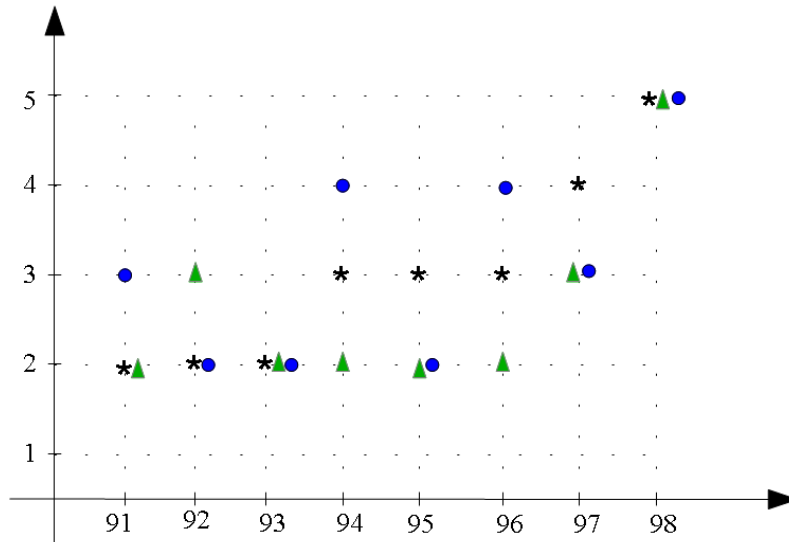


Figura 8.3 - Convexidade

Calculando a regularização de Tikhonov, obtemos $tik(e_{\bullet}) = 3$ e $tik(e_{\blacktriangle}) = 1$. Assim temos $inctg3(e_{\bullet}, x) = 15$ e $inctg3(e_{\blacktriangle}, x) = 5$. Assim sendo, $inctg3$ distingue B_{\blacktriangle} como muito mais temporalmente consistente com a realidade que a B_{\bullet} .

Os resultados acima foram obtidos transformando-se a escala Ω com valores categóricos (f, d, g, m, c) em uma escala numérica (1, 2, 3, 4, 5). Outras transformações são possíveis, podendo-se considerar, por exemplo, que a distância entre o padrão difuso e o geométrico é menor que a diferença entre os padrões geométrico e multidimensional. Mas é importante notar que diferentes escalas numéricas podem gerar ordens diferentes entre os resultados.

8.2 Medidas de inconsistência para aplicações com evolução temporal monotônica

Na Seção 8.1, estudamos algumas medidas de consistência de evolução temporal de caráter geral. No entanto, estas medidas não são por si só adequadas para lidar com aplicações onde o contradomínio de trajetória temporal é uma escala e a evolução

temporal real esperada é monotônica. Por exemplo, no problema de padrões de desmatamento apresentado no Capítulo 7, as classes estão dispostas em uma escala, variando de “floresta” a “consolidado”. A evolução temporal esperada numa célula é monotonicamente crescente, não havendo volta a um estágio anterior ². No entanto, a evolução não é necessariamente estritamente monotônica, podendo uma célula apresentar um mesmo padrão por um número indefinido de anos.

Para medir a consistência de estimativas em problemas com evolução temporal monotônica esperada, primeiramente é necessário se assegurar que a evolução temporal real é também consistente. Na Seção 8.2.1 ilustramos um procedimento para medir que a evolução temporal de uma célula é monotônica. Em seguida, serão mostradas medidas para verificar que uma estimativa de evolução temporal é consistente com a real.

8.2.1 Verificação de evolução temporal monotônica

Suponhamos que temos um conjunto de 5 classes para os padrões: floresta (f), difuso (d), geométrico (g), multidirecional (m), e consolidado (c). Suponhamos também que as seguintes restrições são dadas, considerando resolução temporal de um ano:

- No início da trajetória, uma célula pode apresentar qualquer um dos padrões.
- O padrão floresta pode evoluir para difuso, geométrico.
- O padrão difuso pode evoluir para geométrico, mas o contrário, não.
- O padrão difuso pode evoluir para multidirecional, mas não para consolidado.
- O padrão geométrico pode evoluir para multidirecional ou consolidado.
- O padrão multidirecional só pode evoluir para consolidado.
- O padrão consolidado não evolui para outros padrões e é o estágio final.
- Uma célula pode manter o mesmo padrão por tempo indeterminado.

Constrói-se o grafo $G = (N, E)$, com o conjunto de nós (vértices) $N = \{f, d, g, m, c\}$ e o conjunto de arestas E formado por pares ordenados em N que obedecem as

²Quando a evolução não é monotonicamente crescente, a trajetória da célula é considerada inválida e não é utilizada para medir estimativas.

restrições acima (vide Figura 8.4). Um caminho no grafo é denotado pela sequência $\langle p_1, \dots, p_n \rangle$, onde p_1 e p_n correspondem aos eventos inicial e final, respectivamente, e é tal que p_{i+1} é o sucessor de p_i .

É fácil verificar que os caminhos possíveis no grafo descrevem curvas monotônicas no universo $\{f, d, g, m, c\}$. Por exemplo, em um período de 5 anos podemos ter o caminho $\langle f, f, g, c, c \rangle$, mas não o caminho $\langle f, d, c, c, g \rangle$.

Para verificar a monotonicidade da evolução temporal de uma célula, basta construirmos um autômato finito, cuja linguagem reconhecida é o conjunto dos caminhos possíveis sobre o grafo G . O autômato é construído a partir de G com o conjunto de estados $Q = \{1, 2, 3, 4, 5, 6\}$, com estado inicial $q_0 = 1$, estados finais $F = \{2, 3, 4, 5, 6\}$ e com a função de transição ilustrada na figura 8.4. É fácil verificar que linguagem reconhecida pelo autômato, $L(M_G)$, corresponde ao conjunto de caminhos deriváveis do grafo G .

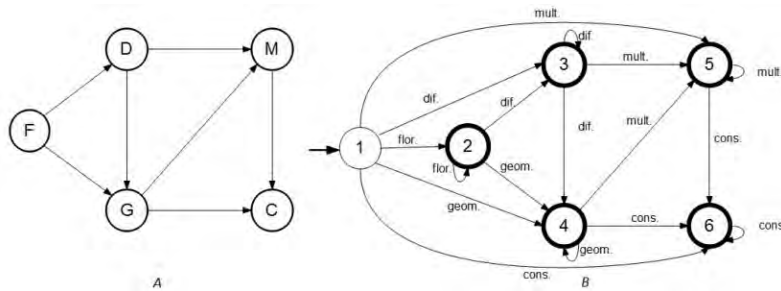


Figura 8.4 - Grafo e autômato para estudo de caso padrões de desmatamento

8.2.2 Verificação de consistência de evolução temporal monotônica

Suponhamos que temos uma célula com evolução temporal real $B^* = (m, m, m)$ e as estimativas $B_{\blacktriangle} = (m, m, c)$, $B_{\bullet} = (g, m, c)$, $B_{\blacklozenge} = (d, m, c)$, $B_{\circ} = (m, m, g)$ e $B_{\blacksquare} = (c, m, g)$ para um dado triênio. É fácil verificar que B^* pertence a $L(M_G)$, assim como as sequências temporais acima, exceto por B_{\circ} e B_{\blacksquare} , que são monotonicamente decrescentes (Figura 8.5).

Embora B_{\blacktriangle} , B_{\bullet} e B_{\blacklozenge} descrevam trajetórias válidas, elas não são perfeitamente consistentes com a trajetória real B^* . É fácil verificar que B_{\blacktriangle} é mais próxima da trajetória real que B_{\bullet} , que por sua vez é melhor que B_{\blacklozenge} . Também se verifica que embora B_{\circ} e B_{\blacksquare} não pertençam a $L(M_G)$, B_{\circ} é mais próxima de B^* que B_{\blacksquare} .

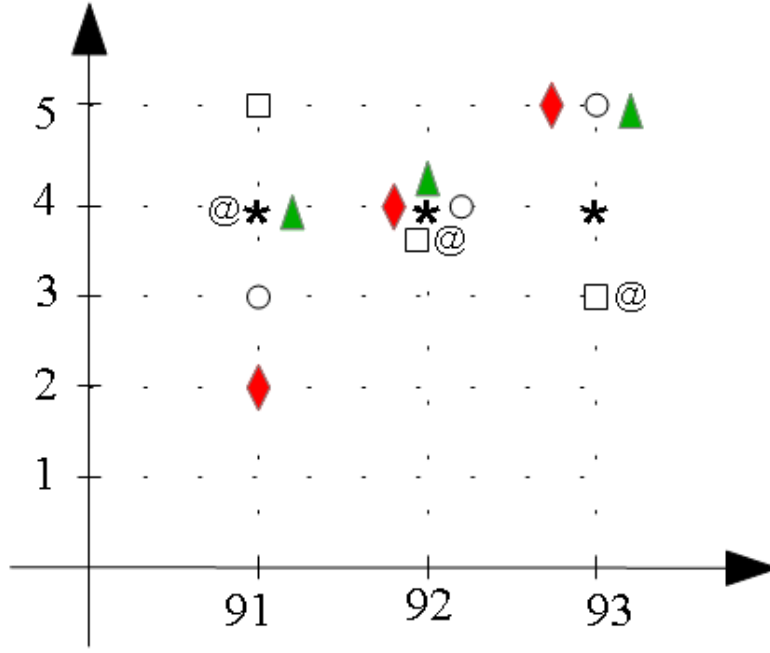


Figura 8.5 - Estimativas temporais

A abordagem mais imediata para medir a consistência temporal de uma trajetória estimada em relação a uma trajetória real dada consiste em utilizar a distância de Hamming entre dois vetores p_1 e p_2 , dada por:

$$Hamm(p_1, p_2) = \sum_{i=1}^R f_H(p_1(i), p_2(i)), \text{ onde } f_H(a, b) = \begin{cases} 1 & \text{se } a = b; \\ 0 & \text{senão.} \end{cases} \quad (8.6)$$

Usando esta distância, verificamos que $Hamm(B^*, B_{\blacktriangle}) = Hamm(B^*, B_{\circ}) = 1$, $Hamm(B^*, B_{\blacksquare}) = Hamm(B^*, B_{\bullet}) = Hamm(B^*, B_{\blacklozenge}) = 2$. Ou seja, esta medida não é capaz de distinguir que \blacktriangle é a melhor estimativa de B^* , confundindo-a com a trajetória inválida B_{\circ} . Além disso, ela não é capaz de distinguir entre B_{\bullet} , B_{\blacklozenge} e B_{\blacksquare} .

O uso das distâncias apresentadas para o caso geral no início do capítulo também não são capazes de levar em conta a exigência de monotonicidade. Suponhamos que o conjunto de classes $\{f, d, g, m, c\}$ seja transformado no conjunto de números naturais $\{1, 2, 3, 4, 5\}$. Teremos, por exemplo, $inctg1(B^*, B_{\blacktriangle}) = inctg1(B^*, B_{\circ}) = 1$, $inctg1(B^*, B_{\bullet}) = inctg1(B^*, B_{\blacksquare}) = 2$ e $inctg1(B^*, B_{\blacklozenge}) = 3$. Ou seja, a função $inctg1$ não é capaz de distinguir entre a sequência monotonicamente crescente B_{\blacktriangle} e a monotonicamente decrescente B_{\circ} . No entanto, medidas como $inctg1$ podem

ser aliadas a uma medida de monotonicidade para gerar uma medida adequada de inconsistência temporal para trajetórias reais monotônicas.

Uma função básica capaz de medir o quanto uma sequência de valores p é monotonicamente crescente, usando nossa notação, é dada por:

$$M^+(B_e) = \sum_{i=1}^{R-1} \max(0, p_i p_{i+1}) \quad (8.7)$$

Para as sequências acima, temos $M^+(B^*) = M^+(B_{\blacktriangle}) = M^+(B_{\bullet}) = M^+(B_{\blacklozenge}) = 0$, $M^+(B_{\circ}) = 1$ e $M^+(B_{\blacksquare}) = 2$. Ou seja, esta medida é capaz de distinguir as sequências monotonicamente crescente das decrescentes.

A medida acima é mínima no melhor caso, i.e. quando a sequência é monotônica, e poderia em princípio ser incorporada diretamente com índices como *inctg1*, como por exemplo na medida

$$inctm^{+'}(e, x) = \sum_{t \in T} |B^*(t) - B_e(t)| \times M^+(B_e) \quad (8.8)$$

No entanto, *inctm^{+'}* não é capaz de distinguir entre B_{\bullet} , B_{\blacktriangle} e B_{\blacklozenge} já que ambas levam a medida a 0.

Uma medida mais adequada é dada por:

$$inctm^{+''}(e, x) = \sum_{t \in T} |B^*(t) - B_e(t)| \times (1 + M^+(B_e)) \quad (8.9)$$

Neste caso, obtemos $inctm^{+''}(B_{\blacktriangle}) = 1$, $inctm^{+''}(B_{\bullet}) = 2$, $inctm^{+''}(B_{\blacklozenge}) = 3$, $inctm^{+''}(B_{\circ}) = 2$ e $inctm^{+''}(B_{\blacksquare}) = 9$. Quando a estimativa obtida é igual à trajetória real, temos $inctm^{+''}(B_{\star}) = 0$, como esperado. Embora esta medida produza bons resultados, ela confunde B_{\bullet} , que erra duas vezes mas é monotônica, com B_{\circ} , que erra somente uma vez mas não é monotônica. Além disso, embora B_{\blacklozenge} seja considerada, corretamente, como pior que B_{\bullet} , ela é considerada também pior que B_{\circ} .

Para lidar com a escolha entre distância da trajetória real e monotonicidade, propomos o uso da medida parametrizada $inctm^{+\delta}$:

$$inctm^{+\delta}(e, x) = \sum_{t \in T} |B^*(t) - B_e(t)| \times (1 + \delta M^+(B_e)), \text{ com } \delta \geq 1 \quad (8.10)$$

Com $\delta = 2$, obtemos: $inctm^{+2}(B_{\blacktriangle}) = 1$, $inctm^{+2}(B_{\bullet}) = 2$, $inctm^{+2}(B_{\blacklozenge}) = 3$, $inctm^{+2}(B_{\textcircled{a}}) = 4$ e $inctm^{+2}(B_{\blacksquare}) = 10$. A ordem entre as estimativas B_{\blacktriangle} , B_{\bullet} e B_{\blacklozenge} permanece inalterada, assim como aquela entre $B_{\textcircled{a}}$ e B_{\blacksquare} . No entanto, agora B_{\blacklozenge} é considerada mais consistente temporalmente com a trajetória real que $B_{\textcircled{a}}$.

Aqui também transformações entre as escalas categóricas e numéricas pode produzir resultados diferentes.

8.3 Experimentos

As medidas de qualidade apresentadas neste capítulo foram aplicadas aos experimentos realizados no Capítulo 6. Os resultados obtidos são apresentados nas Tabelas 8.1 e 8.2. Para estas medições, foram considerados os experimentos que utilizaram os mesmos casos para treinamento e teste apresentados em (LOBO; ESCADA, 2010).

Tabela 8.1 - Qualidade média da classificação efetuada no segundo estudo de caso considerando a evolução temporal

Tipo/Exp	W-R	W-R+	WR	WR+	W+R+	W++R+
<i>inctg1</i>	2,68	0	2,64	2,84	2,88	2,78
<i>inctg2</i>	2,77	0	2,68	2,82	3,18	3,08
<i>inctg3</i>	5,40	0	6,69	13,78	14,24	13,98
<i>inctg^{+\delta}</i>	3,62	0	3,62	4,32	4,46	4,52

Tabela 8.2 - Qualidade média da classificação efetuada no segundo estudo de caso considerando a evolução temporal

Tipo/Exp	W-R	W-R+	WR	WR+	W+R+	W++R+
<i>inctg1</i>	2,79	0,62	3,00	3,55	3,58	3,52
<i>inctg2</i>	2,86	0,61	3,47	3,53	3,63	3,58
<i>inctg3</i>	4,88	2,05	6,44	11,83	11,73	11,45
<i>inctg^{+\delta}</i>	3,57	0,82	3,92	4,64	4,67	4,69

Analisando os resultados apresentados nas Tabelas 8.1 e 8.2 pode-se verificar que, assim como nos Capítulos 6 e 7, os melhores resultados são encontrados quando a base é fragmentada para cálculo das soluções. A ponderação dos casos não gerou

melhora significativa no processo de classificação.

Pode-se observar que os melhores resultados foram obtidos quando consideramos apenas a convexidade das estimativas. Isto pode ser justificado pelo fato de que, mesmo para uma classificação errônea, as curvas obtidas eram convexas, fazendo com que as diferenças encontradas fossem baixas. Porém, deve-se ressaltar que as demais medidas de qualidade mostraram-se mais sensíveis as variações observadas entre a realidade e as estimativas feitas.

9 CONCLUSÕES

Neste trabalho apresentou-se uma abordagem para raciocínio baseado em casos utilizando relações de proximidade difusa. Empregou-se a criação de agrupamentos baseando-se no princípio: “Problemas mais similares na entrada, são também mais parecidos em suas soluções”. Para isso, aplicou-se a abordagem inicialmente proposta por (FANOIKI et al., 2010) e estendida conforme proposto em (SANDRI et al., 2012) e neste trabalho.

Foi proposta uma tipologia de experimentos para efetuar a classificação de dados. Os experimentos foram realizados sem utilização de pesos (W-), com pesos aprendidos para toda a base (W), ou com pesos aprendidos com a base fragmentada em agrupamentos (W+ e W++) . As soluções podem ser calculadas para cada agrupamento (R+), que podem então competir entre si pela solução final, ou serem agregadas em uma única solução, ou considerando-se somente um agrupamento, a própria base (R). Esta tipologia foi aplicada a dois estudos de caso.

No primeiro estudo de casos (prevalência da esquistossomose), os resultados obtidos foram na maioria dos casos melhores que os apresentados na literatura para lidar com o mesmo problema. Considerou-se como critério de comparação a porcentagem de acertos obtida.

Ainda neste estudo de casos, estudou-se o efeito do uso de cortes de nível na relação de similaridade resultante da agregação entre as relações S_{in} e S_{out} . Verificou-se que os melhores resultados foram obtidos para α próximos a zero pois, neste caso, é gerado um número menor de agrupamentos.

Para este estudo de casos, foi analisada a aplicação de diversas medidas de agregação entre relações de proximidade. A força dos agrupamentos foi determinada através de uma média aritmética entre a similaridade de entrada do problema a ser resolvido e os casos que compunham cada agrupamento.

Observou-se também que o uso de cortes de nível, considerando similaridades entre casos, em valores menores do que 1 geram um aumento do número de agrupamentos. Este fato pode inviabilizar a metodologia pois os agrupamentos gerados são utilizados para cálculo dos pesos dos atributos e esta etapa é a mais custosa computacionalmente .

Uma alternativa encontrada foi o uso de RNAs Fuzzy ART para o cálculo dos agrupamentos. Porém sua utilização é indicada quando a base de casos a ser estudada e

conhecida pois agrupamentos desnecessários podem ser gerados caso os parâmetros da rede sejam mal ajustados. A principal vantagem do uso desta técnica está no fato que aqui as informações negativas (casos semelhante na descrição do problema e não na solução), são levadas em conta.

Uma outra extensão utilizada neste estudo de casos foi o fator de coesão. Sua utilização contribuiu para melhorar os resultados obtidos no experimento $W + R+$.

O segundo estudo de casos, padrões de desmatamento, mostrou que a divisão da base de casos em agrupamentos melhora os resultados da classificação e que a sobreposição das classes influencia no resultado final. Para este estudo de casos foram aplicadas diversas funções para se medir a força de um agrupamento para o processo de classificação. Foram realizados estudos usando média aritmética, média geométrica, mínimo, máximo e OWA. Os melhores resultados foram obtidos utilizando-se a função máximo.

Neste estudo de casos, utilizou-se uma medida de qualidade baseada nos conceitos de acurácia e precisão. Os resultados obtidos não apresentaram uma boa qualidade, apesar da acurácia ter sido bastante elevada. Tal fato ocorreu pois a classificação foi imprecisa. Fato que pode ser justificado, em partes, pela semelhança que ocorre entre as classes que compõem o problema (a classe multidirecional é uma evolução das classes difuso e geométrico e a classe consolidado é uma evolução de multidirecional).

Por último, foram propostas medidas de qualidade para análise da classificação em séries temporais. Estas medidas foram aplicadas aos resultados obtidos no segundo estudo de casos e confirmaram os resultados até então obtidos. Foi também verificado que a divisão da base de casos em agrupamento gera melhores resultados que os obtidos quando consideramos a base de casos completa.

A elaboração deste trabalho gerou uma série de possibilidades a serem trabalhadas tais como a validação da classificação efetuada utilizando validação cruzada, aplicar a abordagem em conjunto de dados usuais (Iris, Wine, Abalone, ...) e efetuar a classificação dos casos considerando as informações de anos anteriores.

REFERÊNCIAS BIBLIOGRÁFICAS

AAMODT; PLAZA, E. Case-based reasoning: Foundational issues, methodological variations, and system approaches. **AICom - Artificial Intelligence Communications**, v. 7, n. 1, p. 39–59, 1994. 1, 9

ABEL, M. **Um estudo sobre raciocínio baseado em casos**. Trabalho de Conclusão de Curso — Universidade Federal do Rio Grande do Sul (UFRS), Porto Alegre, 1996. Disponível em:
<<http://www.inf.ufrgs.br/bdi/wp-content/uploads/CBR-TI60.pdf>>. 8

A.N.TIKHONOV; V.S.ARSENIN. Solutions of ill-posed problems. **Winston and Sons**, p. 177–200, 1977. 52

ARMENGOL, E.; ESTEVA, F.; GODO, L.; TORRA, V. On learning similarity relations in fuzzy case-based reasoning. **Trans. on Rough Sets**, p. 14–32, 2004. 1, 15

AZEREDO, M.; ESCADA, M. I. S.; CÂMARA, G. **Mineração de dados espaciais utilizando métricas de paisagem**. São José dos Campos: INPE, 2008. 11, 34, 35

BERGE, C. **Graphs and hypergraphs**. Paris: North-Holland Pub, 1973. 10

CARVALHO, O. S.; DUTRA, L. V.; MOURA, A. C. M.; FREITAS, C. C.; AMARAL, R. S.; DRUMMOND, S. C.; FREITAS, C. R.; SCHOLTE, R. G. C.; GUIMARÃES, R. P. S.; MELO, G. R.; CORREIA, V. R. 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. **Anais do Simpósio Brasileiro de Sensoriamento Remoto**. São José dos Campos: INPE, 2005. 23

DOUMENGE, J. P.; MOTT, K. E.; CHEUNG, C.; VILLENAVE, D.; CAPUI, O.; PERRIN, M. F. **Atlas of the global distribution of schistosomiasis**. [S.l.]: Geneva: WHO-CEGET-CNRS, 1987. 24

DUBOIS, D.; ESTEVA, F.; GARCIA, P.; GODO, L.; MANTARAS, R. L.; PRADE, H. Fuzzy set modelling in case-based reasoning. **International Journal of Intelligent Systems, John Wiley and Sons, Ltda**, v. 13, p. 345373, 1998. 7, 8

ESCADA M. I AND PINHO, C.; MEDEIROS, L.; LOBO, F.; SILVA, M.; KAMPEL, S. **Estrutura, conexão e uso da terra das comunidades, vilas e povoados de São Félix do Xingu e sudeste paraense**. Relatório técnico submetido à biblioteca do INPE. — Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, 2010. 3, 33

ESCADA, M. I. S.; CARRIELO F., A. A. **Uso e Cobertura da Terra em São Félix do Xingu: Levantamento de campo**. Relatório técnico — Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, 2007. 33, 34

FANOIKI, T.; DRUMMOND, I.; SANDRI, S. Case-based reasoning retrieval and reuse using case resemblance hypergraphs. **Proc IEEE Int. Conf Fuzzy Systems**, 2010. 1, 13, 15, 16, 17, 19, 26, 61

FUZZY sets: Information and control. [S.l.: s.n.]. 5

GODO, L.; SANDRI, S. A similarity-based approach to deal with inconsistency in systems of fuzzy gradual rules. **Information Processing and Management of Uncertainty in knowledge-Based Systems**, v. 3, p. 1655, 2002. 8

GUIMARAES, R.; FREITAS, C.; DUTRA, L.; SHIMABUKURO Y.E. ANDMOURA, A.; AMARAL, R.; DRUMMOND, S.; SCHOLTE R.G.C. ANDFREITAS, C. R.; CARVALHO, O. S. Avaliação do modelo de mistura oriundos das imagens modis como uma variável para determinar a prevalência da esquistossomose no estado de minas gerais. In: SIMPÓSIO NACIONAL DE GEOGRAFIA DA SAÚDE, 2005. São José dos Campos, 2005. 23

KATZ, N.; ALMEIDA, K. Esquistossomose, xistosa, barriga d'água. **Ciência e Cultura**, 2003. 23, 24

KOLODNER, J. Cased-based reasoning. **Morgan Kaufmann**, 1993. 1

LOBO, F. L.; ESCADA, M. I. **Análise da dinâmica da paisagem na região do centro de endemismo xingu: estudo de caso na região da terra do meio, Pará**. Relatório técnico submetido à biblioteca do INPE — Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, 2010. 4, 33, 34, 36, 37, 38, 39, 40, 59

MARTINS-BEDÊ, F.; GODO, L.; SANDRI, S. A.; DUTRA, L.; FREITAS, C.; CARVALHO, O.; GUIMARAES, R.; AMARAL, R. Classification of schistosomiasis prevalence using fuzzy case-based reasoning. **International**

Work-Conference on Artificial Neural Networks, v. 5517, p. 1053–1060, 2009. 1, 15, 25, 26, 27, 28, 31

MARTINS, F.; FREITAS, C.; DUTRA, L.; SANDRI, S.; DRUMMOND, I.; FONSECA, F.; GUIMARÃES, R.; AMARAL, R.; CARVALHO, O. Risk mapping of schistosomiasis in the state of minas gerais, brazil, using modis and socioeconomic spatial data. **IEEE Trans. on Geoscience and Remote Sensing**, v. 47, n. 11, p. 3899–3908, 2008. 3, 23, 24, 25

MAS-COLELL, A.; WHINSTON, M.; GREEN, J. Microeconomic theory. **Oxford university press**, 1995. 6

MELLO, E. M. K.; MOREIRA, J. C.; SANTOS, J. r. d.; SHIMABUKURO, Y. E.; DUARTE, V.; SOUZA, I. d. M. e.; BARBOSA, C. C.; SOUZA, R. C. M. d.; PAIVA, J. A. d. C. Prodes digital: experiencia brasileira no mapeamento automatizado do desflorestamento da amazônia. In: SIMPOSIO LATINOAMERICANO DE PERCEPCION REMOTA Y SISTEMAS DE INFORMATION ESPACIAL, 10; REUNION PLENARIA DE SELPER, 21., 11-15 nov. 2002, Cochabamba, Bolivia. [S.l.], 2002. Acesso em: 10 maio 2013. 33

MIKHAIL, E.; ACKERMAN, F. Observations and least squares. **University Press of America**, p. 497, 1976. 43

NEVES, R. **Morphological aspects of Schistosoma mansoni adult worms isolated from nourished and undernourished mice: a comparative analysis by confocal laser scanning microscopy**. [S.l.]: Memorias do Instituto Oswaldo Cruz, 2001. 23

RIESBECK, C.; R.C., S. Inside case-based reasoning. **Erlbaum, Hillsdale**, v. 38, p. 149–157, 1989. 8

RUSPINI, E.; BONISSONE, P.; PEDRYCZ, W. **Handbook of fuzzy computation**. London: IOP Publishing Ltd, 1998. 8

SANDRI, S. A fuzzy residuated approach to case-based reasoning. In: 14TH INTERNATIONAL CONFERENCE ON INFORMATION PROCESSING AND MANAGEMENT OF UNCERTAINTY IN KNOWLEDGE-BASED SYSTEMS, 2012, Catania, Italy. **Proceedings...** Catania, 2012. 2, 13

SANDRI, S.; CORREA, C. Lógica nebulosa. **Escola de Redes Neurais**, p. c73:c90, 1999. 6, 7

SANDRI, S.; DUBOIS, D.; KALFSBEEK, H. W. Elicitation, assessment, and pooling of expert judgments using possibility theory. **IEEE Transactions on Fuzzy Systems**, v. 3, p. 313–335, 1997. [43](#), [44](#)

SANDRI, S.; MENDONÇA, J. H.; MARTINS-BEDÊ, F. Weighted fuzzy similarity relations case-based reasoning: a case study in classification. In: WORLD CONFERENCE COMPUTATIONAL INTELLIGENCE, 2012, Brisbane, Australia. **Proceedings...** Brisbane, 2012. [13](#), [14](#), [15](#), [16](#), [19](#), [22](#), [26](#), [28](#), [61](#)

SILVA, M. P. S.; CÂMARA, G.; ESCADA, M. I. S.; SOUZA, R. C. M.; VALERIANO, D. M. Mining patterns of change in remote sensing image databases. **IEEE International Conference on data mining**, 2005. [24](#), [33](#)

SILVA, N. C. **Utilização de operadores genéticos para otimizar classificadores neurais não-supervisionados de imagens**. 200 p. Tese de Doutorado em Geociências — Universidade de Brasília, Brasília, 2002. [19](#)

TORRA, V. On the learning of weights in some aggregation operators: the weighted mean and owa operators. **Math. and Soft Comp.**, v. 6, 2000. [1](#), [2](#), [15](#), [19](#)

WANGENHEIM, C. G.; WANGENHEIM, A. **Raciocínio baseado em casos**. Curitiba: Editora Manole, 2003. [10](#)

ZADEH, L. A. Similarity relations and fuzzy orderings. **Information Sciences, Elsevier Science Ltd.**, v. 3, n. 2, p. 177–200, 1971. [8](#)

ANEXO A - Matrizes de confusão - Estudo de Caso: Prevalência da Esquistossomose

Neste anexo, são exibidas as matrizes de confusão obtidas para o primeiro estudo de casos. A Tabela .1 mostra as matrizes de confusão dos resultados obtidos para a região 1 utilizando a abordagem regional. As linhas de cada matriz se referem aos valores desejados e as colunas aos valores obtidos. O número de casos para as classes baixa, média e alta são 1, 3 e 1, respectivamente.

Tabela .1 - Matrizes de confusão - Região 1 - Abordagem Regional

W-R	baixa	média	alta	W-R+	baixa	média	alta
baixa	1	0	0	baixa	0	1	0
média	2	1	0	média	3	0	0
alta	0	1	0	alta	0	0	1
WR	baixa	média	alta	WR+	baixa	média	alta
baixa	1	0	0	baixa	0	1	0
média	2	1	0	média	1	1	1
alta	1	0	0	alta	1	0	0
W+R+	baixa	média	alta	W++R+	baixa	média	alta
baixa	0	1	0	baixa	1	0	0
média	1	2	0	média	3	0	0
alta	0	1	0	alta	0	0	1

As matrizes de confusão para a abordagem global são mostradas na Tabela .2.

Tabela .2 - Matrizes de confusão - Região 1 - Abordagem Global

W-R	baixa	média	alta	W-R+	baixa	média	alta
baixa	1	0	0	baixa	0	0	1
média	3	0	0	média	0	1	2
alta	1	0	0	alta	0	1	0
WR	baixa	média	alta	WR+	baixa	média	alta
baixa	0	0	1	baixa	0	0	1
média	0	1	2	média	0	1	2
alta	0	1	0	alta	0	1	0
W+R+	baixa	média	alta	W++R+	baixa	média	alta
baixa	0	0	0	baixa	0	0	1
média	0	1	2	média	0	1	2
alta	0	1	0	alta	0	1	0

A Tabela .3 mostra as matrizes de confusão dos resultados obtidos para a região 2 utilizando a abordagem regional. As linhas de cada matriz se referem aos valores desejados e as colunas aos valores obtidos. O número de casos para as classes baixa, média e alta são 4, 13 e 10, respectivamente.

Tabela .3 - Matrizes de confusão - Região 2 - Abordagem Regional

W-R	baixa	média	alta	W-R+	baixa	média	alta
baixa	0	4	0	baixa	0	2	2
média	0	13	0	média	0	6	7
alta	0	9	1	alta	0	3	7
WR	baixa	média	alta	WR+	baixa	média	alta
baixa	0	3	1	baixa	0	2	2
média	0	10	3	média	1	6	6
alta	1	9	0	alta	0	3	7
W+R+	baixa	média	alta	W++R+	baixa	média	alta
baixa	3	1	0	baixa	0	3	1
média	7	4	2	média	0	8	5
alta	7	2	1	alta	0	4	6

As matrizes de confusão para a abordagem global são mostradas na Tabela .4.

Tabela .4 - Matrizes de confusão - Região 2 - Abordagem Global

W-R	baixa	média	alta	W-R+	baixa	média	alta
baixa	4	0	0	baixa	0	2	2
média	12	1	0	média	0	3	10
alta	9	1	0	alta	0	2	8
WR	baixa	média	alta	WR+	baixa	média	alta
baixa	0	1	3	baixa	0	3	1
média	0	0	13	média	0	3	10
alta	0	2	8	alta	0	1	9
W+R+	baixa	média	alta	W++R+	baixa	média	alta
baixa	0	1	3	baixa	0	0	4
média	0	1	12	média	0	0	13
alta	0	1	9	alta	0	1	9

A Tabela .5 mostra as matrizes de confusão dos resultados obtidos para a região 3 utilizando a abordagem regional. As linhas de cada matriz se referem aos valores desejados e as colunas aos valores obtidos. O número de casos para as classes baixa, média e alta são 2, 6 e 6, respectivamente.

Tabela .5 - Matrizes de confusão - Região 3 - Abordagem Regional

W-R	baixa	média	alta	W-R+	baixa	média	alta
baixa	0	2	0	baixa	2	0	0
média	0	0	6	média	2	1	3
alta	0	0	6	alta	0	1	5
WR	baixa	média	alta	WR+	baixa	média	alta
baixa	0	2	0	baixa	2	0	0
média	0	1	5	média	0	3	3
alta	0	0	6	alta	0	1	5
W+R+	baixa	média	alta	W++R+	baixa	média	alta
baixa	2	0	0	baixa	2	0	0
média	5	0	1	média	2	1	3
alta	6	0	0	alta	0	1	5

As matrizes de confusão para a abordagem global são mostradas na Tabela .6.

Tabela .6 - Matrizes de confusão - Região 3 - Abordagem Global

W-R	baixa	média	alta	W-R+	baixa	média	alta
baixa	2	0	0	baixa	2	0	0
média	6	0	0	média	1	0	5
alta	6	0	0	alta	1	0	5
WR	baixa	média	alta	WR+	baixa	média	alta
baixa	0	2	0	baixa	0	2	0
média	0	3	3	média	0	4	2
alta	0	1	5	alta	0	1	5
W+R+	baixa	média	alta	W++R+	baixa	média	alta
baixa	0	2	0	baixa	0	2	0
média	0	5	1	média	0	3	3
alta	0	2	4	alta	0	1	5

A Tabela .7 mostra as matrizes de confusão dos resultados obtidos para a região 4 utilizando a abordagem regional. As linhas de cada matriz se referem aos valores desejados e as colunas aos valores obtidos. O número de casos para as classes baixa, média e alta são 5, 3 e 1 respectivamente.

As matrizes de confusão para a abordagem global são mostradas na Tabela .8.

Tabela .7 - Matrizes de confusão - Região 4 - Abordagem Regional

W-R	baixa	média	alta	W-R+	baixa	média	alta
baixa	0	5	0	baixa	4	1	0
média	0	3	0	média	1	1	1
alta	0	1	0	alta	1	0	0
WR	baixa	média	alta	WR+	baixa	média	alta
baixa	3	2	0	baixa	5	0	0
média	0	0	3	média	1	2	0
alta	0	1	0	alta	1	0	0
W+R+	baixa	média	alta	W++R+	baixa	média	alta
baixa	5	0	0	baixa	5	0	0
média	0	1	2	média	0	1	2
alta	1	0	0	alta	1	0	0

Tabela .8 - Matrizes de confusão - Região 4 - Abordagem Global

W-R	baixa	média	alta	W-R+	baixa	média	alta
baixa	5	0	0	baixa	5	0	0
média	3	0	0	média	3	0	0
alta	1	0	0	alta	1	0	0
WR	baixa	média	alta	WR+	baixa	média	alta
baixa	0	5	0	baixa	5	0	0
média	0	3	0	média	3	0	0
alta	0	1	0	alta	1	0	0
W+R+	baixa	média	alta	W++R+	baixa	média	alta
baixa	5	0	0	baixa	0	5	0
média	3	0	0	média	0	3	0
alta	1	0	0	alta	0	1	0

ANEXO B - Matrizes de confusão - Estudo de Caso: Padrões de Desmatamento

Neste anexo, são exibidas as matrizes de confusão obtidas para o primeiro estudo de casos.

As Tabelas .1 e .2 mostram as matrizes de confusão obtidas em cada um dos experimentos para a classificação dos casos referentes ao ano de 1997 utilizados nas etapas de treinamento e teste, respectivamente, da metodologia.

Tabela .1 - Matrizes de confusão - Casos de Treinamento - 1997

W-R	C	D	G	M	W-R+	C	D	G	M
C	0	0	2	0	C	2	0	0	0
D	0	209	1	0	D	0	210	0	0
G	0	38	23	0	G	0	0	61	0
M	0	0	44	0	M	0	0	0	44
WR	C	D	G	M	WR+	C	D	G	M
C	0	0	2	0	C	2	0	0	0
D	0	210	0	0	D	0	210	0	0
G	0	57	4	0	G	0	0	61	0
M	0	0	44	0	M	0	0	0	44
W+R+	C	D	G	M	W++R+	C	D	G	M
C	2	0	0	0	C	2	0	0	0
D	0	210	0	0	D	0	210	0	0
G	0	0	61	0	G	0	0	61	0
M	0	0	0	44	M	0	0	0	44

As Tabelas .3 e .4 mostram as matrizes de confusão obtidas em cada um dos experimentos para a classificação dos casos referentes ao ano de 2000 utilizados nas etapas de treinamento e teste, respectivamente, da metodologia.

As Tabelas .5 e .6 mostram as matrizes de confusão obtidas em cada um dos experimentos para a classificação dos casos referentes ao ano de 2003 utilizados nas etapas de treinamento e teste, respectivamente, da metodologia.

As Tabelas .7 e .8 mostram as matrizes de confusão obtidas em cada um dos experimentos para a classificação dos casos referentes ao ano de 2006 utilizados nas etapas de treinamento e teste, respectivamente, da metodologia.

As Tabelas .9 e .10 mostram as matrizes de confusão obtidas em cada um dos

Tabela .2 - Matrizes de confusão - Casos de Teste - 1997

W-R	C	D	G	M	W-R+	C	D	G	M
C	0	0	1	0	C	1	0	0	0
D	0	85	6	0	D	0	83	7	0
G	0	10	16	0	G	0	2	24	0
M	0	0	19	0	M	0	0	2	17
WR	C	D	G	M	WR+	C	D	G	M
C	1	0	1	0	C	1	0	0	0
D	0	90	1	0	D	0	70	21	0
G	0	24	2	0	G	0	5	21	0
M	0	1	18	0	M	0	0	3	16
W+R+	C	D	G	M	W++R+	C	D	G	M
C	1	0	0	0	C	1	0	0	0
D	0	73	13	5	D	0	72	14	5
G	0	8	13	5	G	0	6	16	4
M	0	1	5	13	M	0	0	3	16

Tabela .3 - Matrizes de confusão - Casos de Treinamento - 2000

W-R	C	D	G	M	W-R+	C	D	G	M
C	0	0	12	0	C	12	0	0	0
D	0	161	17	0	D	0	178	0	0
G	0	15	48	0	G	0	0	63	0
M	0	0	64	0	M	0	0	0	64
WR	C	D	G	M	WR+	C	D	G	M
C	0	0	12	0	C	12	0	0	0
D	0	173	5	0	D	0	178	0	0
G	0	32	31	0	G	0	0	63	0
M	0	0	64	0	M	0	0	0	64
W+R+	C	D	G	M	W++R+	C	D	G	M
C	12	0	0	0	C	12	0	0	0
D	0	178	0	0	D	0	178	0	0
G	0	0	63	0	G	0	0	63	0
M	0	0	0	64	M	0	0	0	64

experimentos para a classificação dos casos referentes ao ano de 2009 utilizados nas etapas de treinamento e teste, respectivamente, da metodologia.

Tabela .4 - Matrizes de confusão - Casos de Teste - 2000

W-R	C	D	G	M	W-R+	C	D	G	M
C	0	0	6	0	C	6	0	0	0
D	0	69	7	0	D	0	76	0	0
G	0	8	19	0	G	0	4	23	0
M	0	0	26	3	M	0	0	0	29
WR	C	D	G	M	WR+	C	D	G	M
C	0	0	6	0	C	6	0	0	0
D	0	74	2	0	D	0	76	0	0
G	0	14	13	0	G	0	4	23	0
M	0	0	25	4	M	3	0	2	24
W+R+	C	D	G	M	W++R+	C	D	G	M
C	5	1	0	0	C	5	1	0	0
D	0	76	0	0	D	0	76	0	0
G	0	6	21	0	G	0	5	22	0
M	2	0	2	25	M	3	0	0	26

Tabela .5 - Matrizes de confusão - Casos de Treinamento - 2003

W-R	C	D	G	M	W-R+	C	D	G	M
C	0	0	2	28	C	30	0	0	0
D	0	171	21	0	D	0	192	0	0
G	0	6	93	0	G	0	0	104	0
M	0	0	85	9	M	0	0	0	94
WR	C	D	G	M	WR+	C	D	G	M
C	4	0	1	25	C	30	0	0	0
D	0	192	0	0	D	0	192	0	0
G	0	57	46	1	G	0	0	104	0
M	0	18	67	9	M	0	0	0	94
W+R+	C	D	G	M	W++R+	C	D	G	M
C	30	0	0	0	C	30	0	0	0
D	0	192	0	0	D	0	192	0	0
G	0	0	104	0	G	0	0	104	0
M	0	0	0	94	M	0	0	0	94

Tabela .6 - Matrizes de confusão - Casos de Teste - 2003

W-R	C	D	G	M	W-R+	C	D	G	M
C	0	0	0	11	C	9	0	2	0
D	0	69	14	0	D	0	81	2	0
G	0	8	34	2	G	0	2	42	0
M	0	0	24	16	M	4	0	1	35
WR	C	D	G	M	WR+	C	D	G	M
C	1	0	0	10	C	8	0	3	0
D	0	83	0	0	D	0	83	0	0
G	0	25	18	1	G	0	0	43	1
M	0	3	21	16	M	0	1	10	29
W+R+	C	D	G	M	W++R+	C	D	G	M
C	7	0	4	0	C	8	0	3	0
D	0	83	0	0	D	0	83	0	0
G	0	0	44	0	G	0	0	43	1
M	1	2	10	27	M	1	1	12	26

Tabela .7 - Matrizes de confusão - Casos de Treinamento - 2006

W-R	C	D	G	M	W-R+	C	D	G	M
C	0	0	3	28	C	31	0	0	0
D	0	175	27	0	D	0	202	0	0
G	0	2	107	9	G	0	0	118	0
M	0	0	104	16	M	0	0	0	120
WR	C	D	G	M	WR+	C	D	G	M
C	0	0	2	29	C	31	0	0	0
D	0	200	2	0	D	0	202	0	0
G	1	41	76	1	G	0	0	118	0
M	0	6	99	15	M	0	0	0	120
W+R+	C	D	G	M	W++R+	C	D	G	M
C	31	0	0	0	C	31	0	0	0
D	0	202	0	0	D	0	202	0	0
G	0	0	118	0	G	0	0	118	0
M	0	0	0	120	M	0	0	0	120

Tabela .8 - Matrizes de confusão - Casos de Teste - 2006

W-R	C	D	G	M	W-R+	C	D	G	M
C	0	0	0	13	C	13	0	0	0
D	0	76	10	0	D	0	84	2	0
G	0	5	40	5	G	0	3	47	0
M	0	0	37	14	M	0	0	3	48
WR	C	D	G	M	WR+	C	D	G	M
C	0	0	0	13	C	11	0	2	0
D	0	86	0	0	D	0	86	0	0
G	0	20	27	3	G	0	0	48	2
M	0	0	37	14	M	0	2	6	43
W+R+	C	D	G	M	W++R+	C	D	G	M
C	12	0	1	0	C	11	0	2	0
D	0	86	0	0	D	0	86	0	0
G	0	0	49	1	G	0	0	48	2
M	1	2	14	34	M	1	0	15	35

Tabela .9 - Matrizes de confusão - Casos de Treinamento - 2009

W-R	C	D	G	M	W-R+	C	D	G	M
C	0	0	4	35	C	39	0	0	0
D	0	157	34	0	D	0	191	0	0
G	0	2	104	7	G	0	0	113	0
M	0	0	117	24	M	0	0	0	141
WR	C	D	G	M	WR+	C	D	G	M
C	25	0	0	14	C	39	0	0	0
D	0	190	1	0	D	0	191	0	0
G	0	33	76	4	G	0	0	113	0
M	0	5	114	22	M	0	0	0	141
W+R+	C	D	G	M	W++R+	C	D	G	M
C	39	0	0	0	C	39	0	0	0
D	0	191	0	0	D	0	191	0	0
G	0	0	113	0	G	0	0	113	0
M	0	0	0	141	M	0	0	0	141

Tabela .10 - Matrizes de confusão - Casos de Teste - 2009

W-R	C	D	G	M	W-R+	C	D	G	M
C	0	0	2	14	C	15	0	0	1
D	0	68	14	0	D	0	82	0	0
G	0	0	39	9	G	1	0	46	1
M	0	0	47	13	M	0	0	1	59
WR	C	D	G	M	WR+	C	D	G	M
C	9	0	2	5	C	14	0	2	0
D	0	82	0	0	D	0	82	0	0
G	0	8	35	5	G	0	0	48	0
M	0	1	50	9	M	0	7	25	28
W+R+	C	D	G	M	W++R+	C	D	G	M
C	13	0	3	0	C	14	0	1	1
D	0	82	0	0	D	0	82	0	0
G	0	0	48	0	G	0	0	45	3
M	1	7	12	40	M	0	10	22	28