Definição de critérios a partir da mineração de dados para a extração automática de redes de drenagem

Lise Christine Banon
Rafael Duarte Coelho dos Santos
Nandamudi Lankalapalli Vijaykumar
Camilo Daleles Rennó

Instituto Nacional de Pesquisas Espaciais - INPE
Caixa Postal 515 - 12227-010 - São José dos Campos - SP, Brasil
{lise, camilo}@dpi.inpe.br, rafael.santos@inpe.br, vijay@lac.inpe.br

Abstract. Conventional methods for automatic extraction of drainage networks usually use a threshold based on the minimum area of contribution criteria. However, these methods rarely present realistic results and cumbersome manual editing is required. This paper proposes selection of a new set of attributes using supervised learning methods (decision trees) to develop a classification methodology for the automatic extraction of drainage networks. Among the methods evaluated, the J48 algorithm showed the best results with accuracy greater than 90%.

Palavras-chave: data mining, decision tree, drainage network, DEM, mineração de dados, árvore de decisão, rede de drenagem, MDE, SRTM.

1. Introdução

A altimetria de uma região pode ser representada por um modelo digital de elevação (MDE), que pode ser estruturado na forma de grades regulares. O valor de cada elemento (célula ou pixel) desta grade indica a altitude média ou a altitude do ponto central desse elemento.

Nos últimos anos, muitos métodos de extração automática de redes de drenagem a partir de uma grade regular têm sido desenvolvidos. Esses métodos podem ser agrupados em dois tipos de algoritmos: um baseado em feições geomorfológicas e outro baseado em características hidrológicas do terreno (Soille et al., 2003).

Em muitos casos, a extração automática de redes de drenagem utiliza limiares para sua definição. Neste processo de classificação, não há garantia da formação de uma rede contínua, demandando a correção manual por um especialista para a conexão de trechos isolados e a eliminação de pontos irrelevantes.

Um dos primeiros trabalhos sobre a extração automática de redes de drenagem, a partir de MDE, foi apresentado por O'Callaghan e Mark (1984). Esta metodologia propõe o uso de um limiar, baseado na área mínima de contribuição, para identificar os pontos onde a rede de drenagem se origina (nascentes). A área de contribuição de um ponto qualquer representa o número de pontos (ou área) que converge àquele determinado ponto. Métodos baseados na definição de um limiar de área de contribuição são muito simples de implementar e portanto muito populares. No entanto, em regiões com diferentes padrões geomorfológicos, a escolha de um único limiar para representar toda a região é extremamente complicada, podendo gerar redes de drenagem com maior ou menor densidade do que a real.

A Figura 1a apresenta o MDE SRTM (*Shuttle Radar Topography Mission*) de uma região com diferentes padrões geomorfológicos. A escolha de um alto limiar para a área de contribuição resultou em uma adequada classificação para as áreas com menor densidade de drenagem, mas em contrapartida não classificou adequadamente as regiões mais densas. Na área ampliada da região mais densa, é possível observar que as extremidades da drenagem não foram classificadas. Por outro lado, a Figura 1b apresenta a mesma região, desta vez usando um baixo limiar, o que resultou em uma adequada classificação para as áreas com maior densidade, mas criou feições inexistentes em áreas com menor densidade.

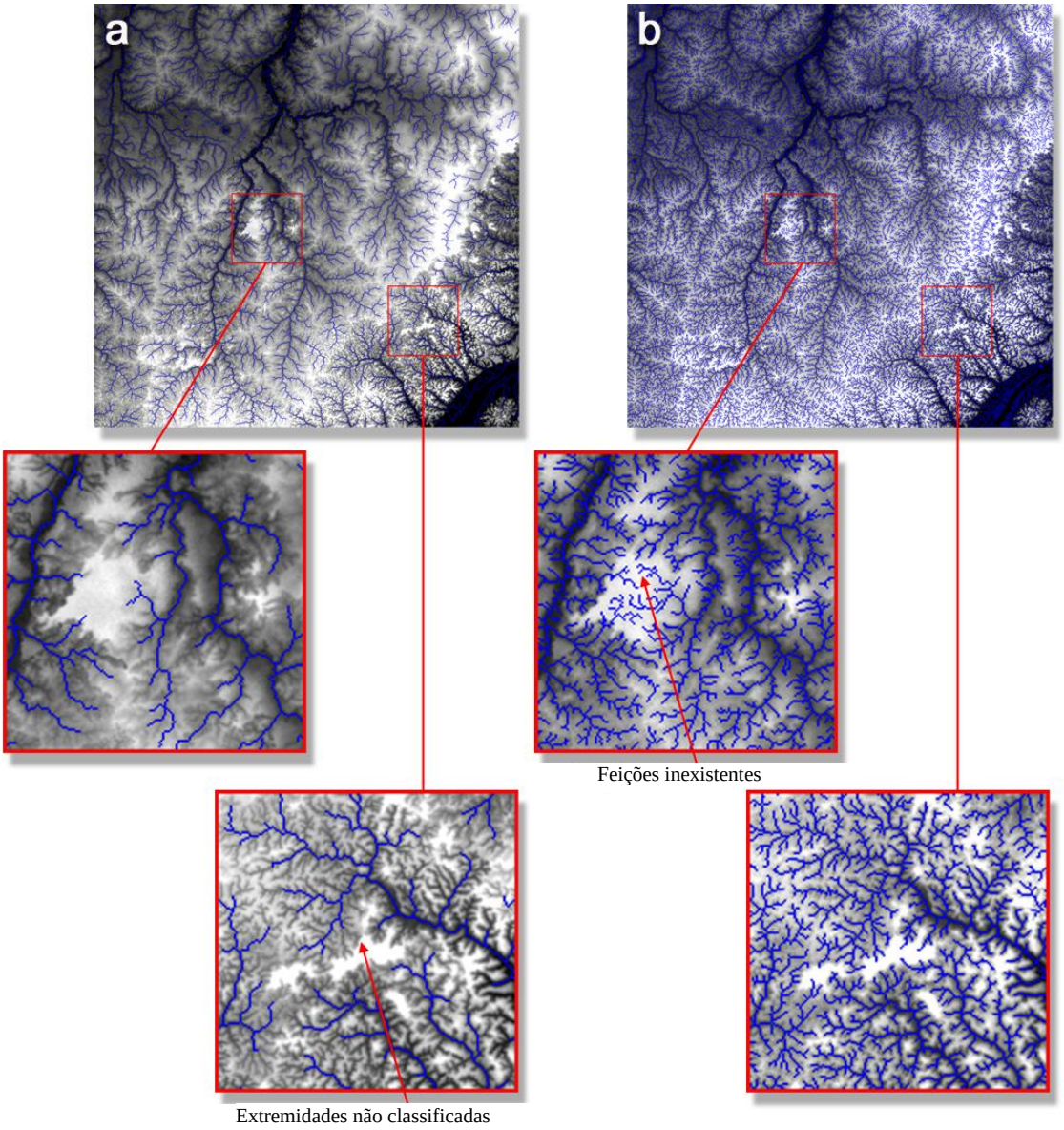

Figura 1. MDE SRTM com drenagem (em azul) resultante da classificação considerando área de contribuição igual ou superior a: (a) 1.287.900m^2 e (b) 85.860m^2.

Além da área de contribuição, existem outros critérios que poderiam ser utilizados na definição da rede de drenagem. Por exemplo, a declividade e as curvaturas vertical e horizontal podem indicar áreas de convergência ou divergência dos fluxos de água na superfície, auxiliando na identificação de regiões potencialmente associadas às nascentes.

Desta forma, baseando-se em atributos extraídos do MDE SRTM, o presente trabalho propôs uma metodologia para a extração automática de uma rede de drenagem capaz de representar áreas com diferentes padrões geomorfológicos. Com o emprego de técnicas de Mineração de Dados pretendeu-se definir o conjunto de atributos mais representativo desta rede.

2. Material e Métodos

2.1 Área de Estudo

A área de estudo localiza-se numa região coberta pela floresta amazônica, próxima ao município de Santarém/PA, entre as coordenadas 3°50’33’’S, 56°10’27’’W e 3°5’36’’S, 55°25’33’’W, conforme pode ser observado na Figura 2. Esta região foi escolhida como área de estudo por apresentar diferentes padrões de drenagem.

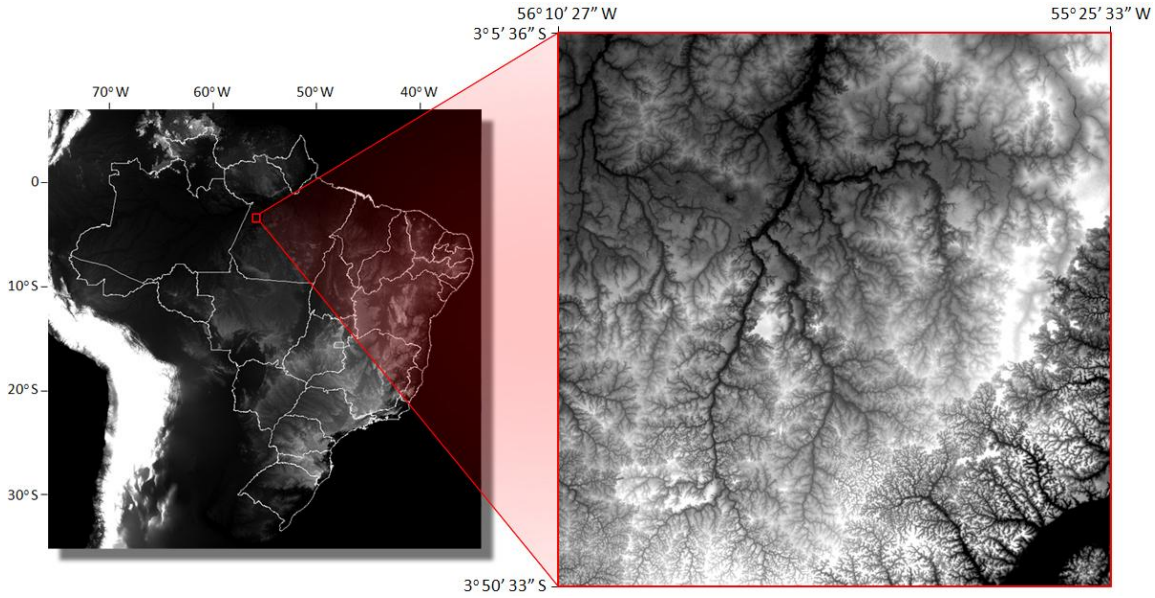

Figura 2. MDE SRTM da área de estudo próxima ao município de Santarém/PA.

2.2 Extração de Atributos

Para este trabalho, foram calculados vinte e três atributos agrupados em duas categorias, sendo dezoito morfométricos e cinco baseados na direção de fluxo. A Tabela 1 apresenta a lista de atributos. Os atributos morfométricos descrevem feições do terreno e são obtidos a partir da topografia local através do ajuste de uma função bivariada quadrática utilizando os valores do MDE SRTM de uma janela 3x3.

Tabela 1. Lista de Atributos extraídos a partir do MDE SRTM.

	Atributos	Unidade	Fonte
Morfométricos	Declividade	%	Wood (1996); Shary et al. (2002)
	Fator de Gradiente	$10^2 m.m^{-1}$	Shary et al. (2002)
	Curvatura Horizontal	$10^3 m^{-1}$	Shary et al. (2002)
	Curvatura Plana	$10^3 m^{-1}$	Wood (1996); Shary et al. (2002)
	Curvatura Vertical	$10^3 m^{-1}$	Wood (1996), Shary et al. (2002)
	Curvatura Média	$10^3 m^{-1}$	Wood (1996), Shary et al. (2002)
	Não Esfericidade	$10^3 m^{-1}$	Shary et al. (2002)
	Diferença de curvatura	$10^3 m^{-1}$	Shary et al. (2002)
	Rotor	$10^3 m^{-1}$	Shary et al. (2002)
	Excesso de Curvatura Horizontal	$10^3 m^{-1}$	Shary et al. (2002)
	Excesso de Curvatura Vertical	$10^3 m^{-1}$	Shary et al. (2002)
	Curvatura Mínima	$10^3 m^{-1}$	Wood (1996); Shary et al. (2002)
	Curvatura Máxima	$10^3 m^{-1}$	Wood (1996); Shary et al. (2002)
	Curvatura Gaussiana Total	$10^6 m^{-2}$	Shary et al. (2002)
	Curvatura Circular Total	$10^3 m^{-2}$	Shary et al. (2002)
	Curvatura de Acumulação Total	$10^6 m^{-2}$	Shary et al. (2002)
	Curvatura Longitudinal	$10^3 m^{-1}$	Wood (1996)
	Curvatura Transversal	$10^3 m^{-1}$	Wood (1996)
Baseados na Direção de Fluxo	Área de Contribuição	m^2	Rennó e Banon (2012a)
	Desnível ao Topo	m	Rennó e Banon (2012a)
	Desnível Vertical	m	Rennó e Banon (2012a)
	Declividade à Jusante	%	Rennó e Banon (2012a)
	Ordem Máxima de Sthraler	adimensional	Rennó e Banon (2012a)

Enquanto os atributos morfométricos são extraídos diretamente do MDE, os atributos baseados na direção de fluxo necessitam que o MDE seja hidrologicamente consistido, ou

seja, os fluxos que determinam o caminho da água devem ser contínuos. O MDE SRTM possui algumas limitações devido à geração de seus dados, não representando fielmente os fluxos locais. Para minimizar este problema foi elaborada uma fase de pré-processamento dos dados. Nesta fase, o MDE SRTM foi previamente corrigido a fim de eliminar os sumidouros (mínimos locais), resultando na grade de fluxos locais (LDD, *Local Drainage Direction*) que representa os caminhos contínuos da água de um ponto até uma das bordas da grade. Esta correção é imprescindível, pois os atributos baseados na direção de fluxo são fundamentados nas informações contidas no LDD e para implementação desta etapa foi utilizado um aplicativo desenvolvido na plataforma ENVI/IDL, o HAND_GRID (Rennó et al., 2008).

2.3 Escolha das Amostras

A partir de uma grade de 900x900 pixels, foram escolhidos 660 pontos dos quais 160 representam nascentes, 300 representam pontos que não pertencem à rede de drenagem e 200 representam drenagens perenes. Estes pontos foram escolhidos por um especialista usando critérios de interpretação visual (Figura 3).

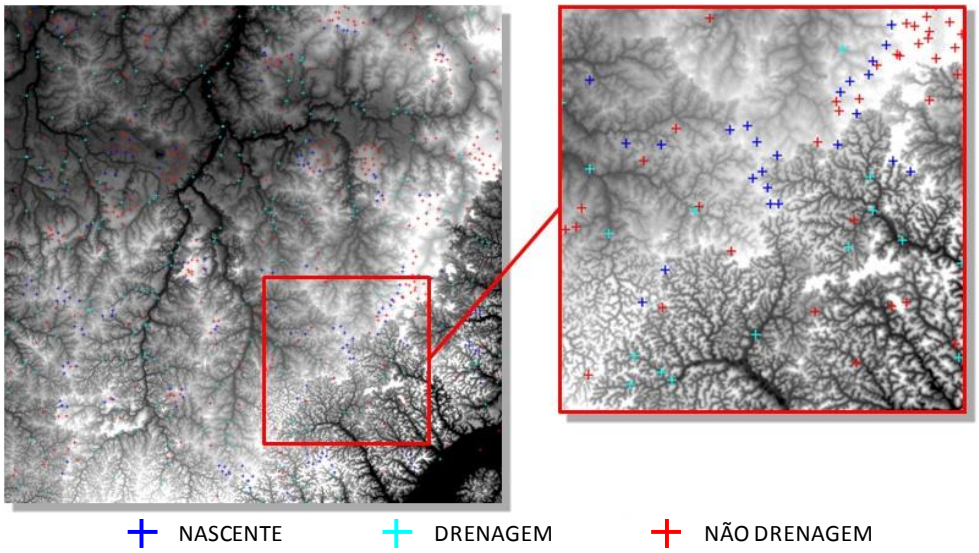

Figura 3. Distribuição dos pontos selecionados.

Os valores referentes a cada ponto da amostra foram exportados, gerando uma tabela onde cada linha representa um ponto e as colunas representam os valores de cada atributo para esse ponto.

2.4 Mineração de Dados

Diante de um volume considerável de dados, é possível definir técnicas que permitam identificar informações relevantes, entender melhor o significado destes dados e estipular procedimentos de tomada de decisão. Para esta finalidade surgiu a Descoberta do Conhecimento em Banco de Dados (DCBD) - termo derivado do inglês *Knowledge Discovery in Databases* (KDD), que segundo Fayyad et al. (1996), pode ser definida como o processo não trivial de identificar em dados, padrões que sejam válidos, novos (agreguem conhecimento), potencialmente úteis e compreensíveis. A Mineração de Dados é um passo fundamental dentro deste processo maior conhecido como DCBD e compreende a seleção dos métodos que serão utilizados para identificar padrões de interesse nos dados em análise, sempre buscando o melhor ajuste dos parâmetros no algoritmo escolhido.

O método escolhido para este trabalho utilizou os classificadores da categoria Árvore de Decisão, por serem representações simples do conhecimento e um meio eficiente de construir classificadores que predizem ou revelem classes ou informações úteis baseadas nos valores de atributos de um conjunto de dados (De'ath e Fabricius, 2000). Para a execução desta etapa de

Mineração de Dados, optou-se pelo pacote de aplicativos WEKA *(Waikato Environment for Knowledge Analysis*) (Witten e Frank, 2005), por ser um *software* gratuito e de código aberto (*opensource*). Na aplicação do método escolhido, foi necessário definir previamente as classes a serem utilizadas no conjunto de treinamento, motivo pelo qual a Árvore de Decisão é conhecida como um método de classificação supervisionada.

A partir de um conjunto de treinamento, a Árvore de Decisão é organizada em uma hierarquia de nós internos e externos que são interligados por ramos. O nó interno (não-folha) recebe como rótulo um dos atributos previsores e é considerado uma unidade de tomada de decisão pois avalia através de teste lógico qual será o próximo nó descendente. Dependendo do resultado do teste lógico, a árvore se ramifica e este procedimento é repetido até que um nó externo seja alcançado. A repetição deste procedimento caracteriza a recursividade da árvore de decisão (Breiman et al., 1984). O teste lógico que determina a ramificação da árvore é baseado na entropia (medida da falta de homogeneidade dos dados de entrada em relação à sua classificação). O valor do limiar utilizado como critério para ramificação é associado a cada ramo que conecta um nó interno ao seu descendente. Por fim, o nó externo (sem descendentes), também conhecido como folha, é rotulado com uma classe.

Neste trabalho, foram analisados todos os algoritmos oferecidos pelo WEKA na categoria Árvore de Decisão: J48, BFTree, Decision Stump, FT, LADTree, LMT, NBTree, Random Forest, Random Tree, REPTree e Simple Cart.

2.5 Classificação da Rede de Drenagem e Pós-Processamento

Para obter a classificação hierárquica da rede de drenagem no ambiente IDL/ENVI, a partir da árvore de decisão fornecida pelo WEKA, foi desenvolvido um programa denominado TREE_CLASSIFICATION em IDL/ENVI (Rennó e Banon, 2012d). Este programa reproduz as etapas definidas na árvore de decisão, ou seja, os atributos-chave são selecionados e aplica-se a regra definida para cada nó, de modo que cada ponto da grade (pixel) será rotulado para uma das classes pré-definidas.

Normalmente, a rede de drenagem resultante deste processo de classificação apresenta descontinuidades que devem ser corrigidas. Para este processo, foi desenvolvido outro programa em IDL/ENVI denominado DRAINAGE_CORRECTION (Rennó e Banon, 2012b), que utiliza a informação de conectividade presente na grade LDD para completar a rede de drenagem.

A rede de drenagem resultante foi posteriormente processada a fim de eliminar pequenos trechos de drenagem de primeira ordem (aqueles associados diretamente às nascentes). Este procedimento utilizou um terceiro programa também desenvolvido em IDL/ENVI denominado DRAINAGE_SIMPLIFICATION (Rennó e Banon, 2012c) para eliminação automática dos trechos de drenagem de primeira ordem com apenas 1 pixel de comprimento.

2.6 Avaliação da Classificação

A avaliação do resultado da classificação da rede de drenagem foi realizada em duas etapas: Análise da Validação Cruzada e Avaliação Qualitativa da Rede de Drenagem gerada a partir do Pós-processamento.

A partir da utilização do recurso de Validação Cruzada Estratificada com 10 partições do WEKA (*Stratified Ten-Fold Cross Validation*), as amostras são divididas automaticamente em 10 conjuntos, sendo 9 utilizados para treinamento do classificador e o conjunto restante para testá-lo. Na Validação Cruzada, o conjunto de teste é alternado por um dos conjuntos de treinamento e o processo é repetido até que todos os conjuntos tenham sido utilizados como teste. Como resultado desta análise, o WEKA fornece, além da estrutura da árvore de decisão, a matriz de confusão e a porcentagem de acerto.

A segunda etapa se refere à avaliação qualitativa da Rede de Drenagem gerada a partir do Pós-processamento. O mesmo especialista que selecionou as amostras, avaliou

qualitativamente a rede de drenagem usando critérios de interpretação visual, a fim de verificar incoerências na definição desta rede. Foram considerados dois tipos de erro nessa avaliação: a falta de trechos de drenagem em regiões cujo padrão geomorfológico no MDE SRTM original indicava presença de canais e a presença de falsos canais de drenagem em regiões sem evidências fortes da presença dos mesmos.

Baseando-se nas etapas anteriores, foram selecionados os algoritmos do WEKA que obtiveram os melhores resultados. Para o algoritmo com o melhor resultado, os pontos amostrais foram sobrepostos à sua classificação e, no ambiente ENVI, os pontos classificados erroneamente foram identificados e reposicionados para garantir uma melhor caracterização de cada classe.

Constatando-se a necessidade de readequação das amostras, toda a etapa de Mineração de Dados foi refeita até que o resultado fosse considerado satisfatório pelo especialista.

3. Resultados

Inicialmente, a avaliação da classificação da rede de drenagem apresentou o melhor resultado para o algoritmo J48. Utilizando a Validação Cruzada com 10 partições, este algoritmo apresentou apenas 85% de acerto e gerou uma Árvore de Decisão contendo 16 atributos, o que representou uma árvore extensa e complexa. A maior confusão ocorreu entre as classes "Não Drenagem" e "Nascente" totalizando 79 erros. As confusões entre as classes "Nascente" e "Drenagem" não foram consideradas, pois ambas representam a rede de drenagem.

Uma vez que os resultados do classificador J48 não foram satisfatórios, optou-se por reavaliar os pontos amostrais classificados erroneamente. Desta etapa, cerca de 50 pontos foram reposicionados gerando um novo conjunto amostral. Consequentemente, as etapas de Mineração de Dados (classificação, análise e avaliação) foram refeitas, sendo escolhidos os quatro algoritmos com a maior porcentagem de acerto: BFTree (93,64%), simpleCart (93,33%), REPTree (93,03%) e J48 (92,58%).

Apesar dos outros algoritmos apresentarem porcentagens de acerto superiores ao J48, qualitativamente, este algoritmo obteve um resultado muito superior em relação aos demais (Figura 4). A Validação Cruzada com 10 partições deste algoritmo gerou uma matriz indicando que a confusão entre as classes "Não Drenagem" e "Nascente" reduziu para 23 erros (Tabela 2).

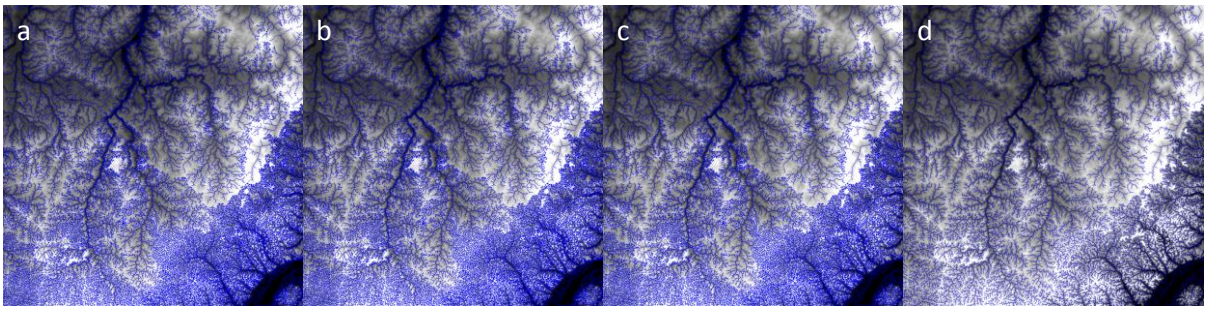

Figura 4. Resultado final obtido a partir dos algoritmos: a) BFTree, b) simpleCart, c) REPTree e d) J48.

Tabela 2. Matriz de Confusão relativa à Validação Cruzada para a amostra reformulada usando J48.

		CLASSIFICAÇÃO CORRETA		
		Não Drenagem	Nascente	Drenagem
PREVISÃO	Não Drenagem	**283**	9	8
	Nascente	14	**139**	7
	Drenagem	6	5	**189**

O algoritmo J48 apresentou uma Árvore de Decisão com nove atributos, mais simples em relação a anterior (Figura 5). Este resultado indicou a importância dos atributos morfométricos, uma vez que várias representações de curvaturas foram selecionadas. Além disso, dos cinco atributos baseados na direção de fluxo, três foram selecionados, destacando-se a Área de Contribuição utilizada na metodologia clássica.

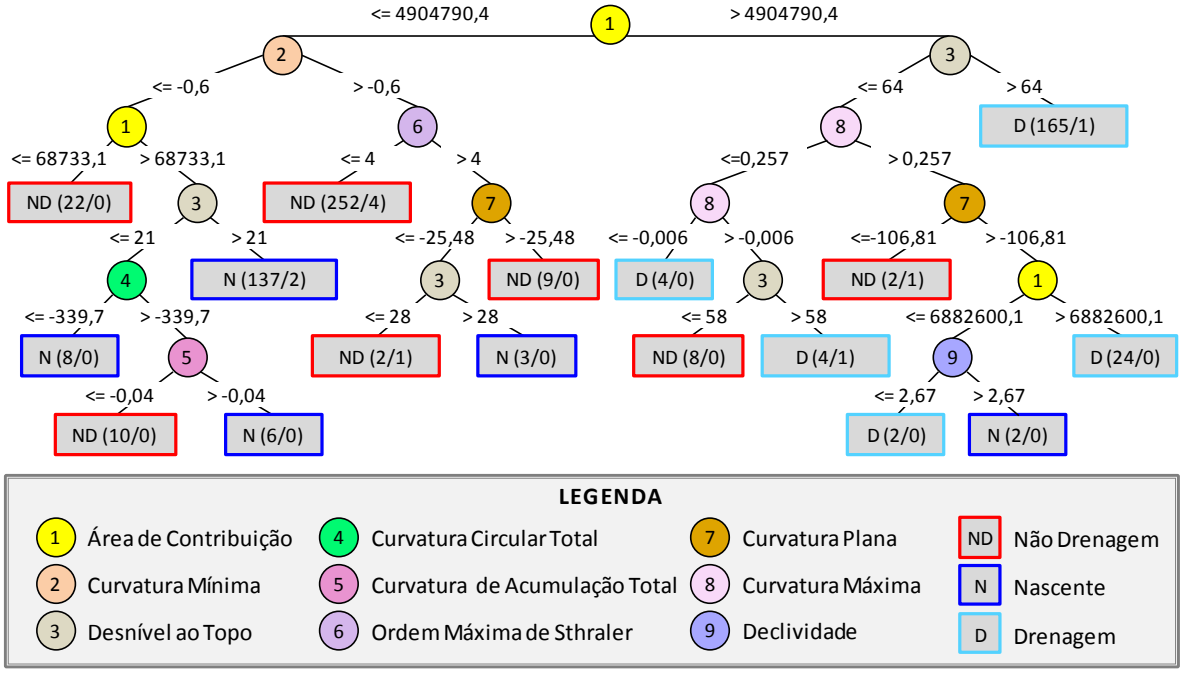

Figura 5. Árvore de Decisão construída pelo algoritmo J48 para a amostra reformulada.

É importante ressaltar que o algoritmo J48 também apresentou um resultado muito superior em relação à metodologia clássica, exemplificada na Figura 1. Ao considerar as mesmas áreas em destaque nesta figura, pode-se observar que o J48 não resultou em tantas feições inexistentes e a classificação das extremidades foi aprimorada, resultando em uma classificação adequada para áreas com menor e maior densidade de drenagem (Figura 6).

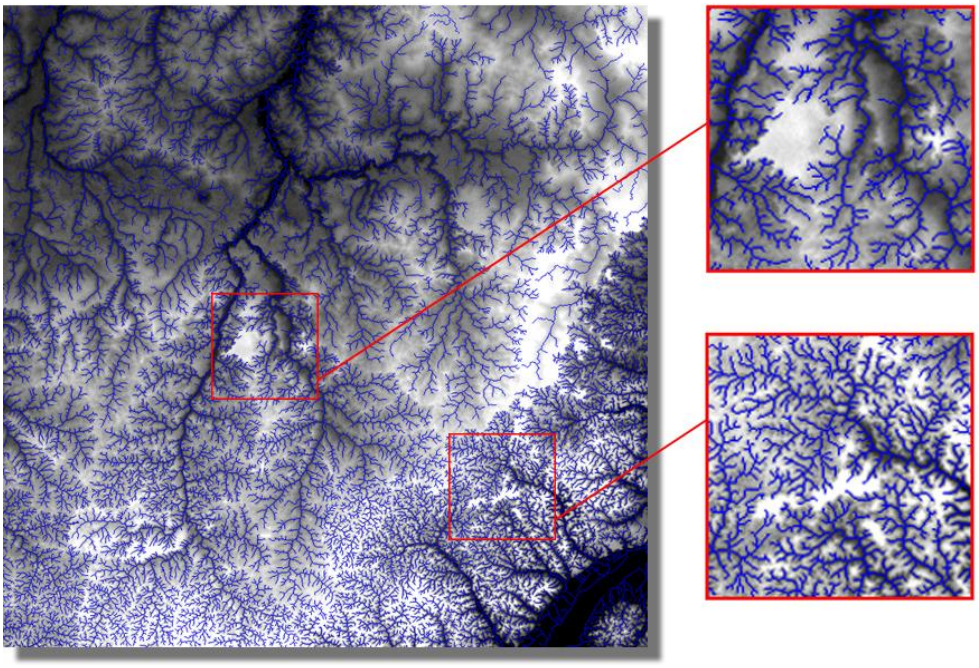

Figura 6. Resultado final obtido para o algoritmo J48.

4. Conclusões

Este trabalho apresentou uma metodologia para a extração automática de redes de drenagem, que possibilitou a representação de áreas com diferentes padrões geomorfológicos.

Com o emprego de técnicas de Mineração de Dados, foi definido o conjunto de atributos mais representativo da rede de drenagem em estudo. Este objetivo foi atingido utilizando um classificador baseado em Árvore de Decisão, o algoritmo J48, que na fase de avaliação do pós-processamento, apresentou um resultado qualitativo muito superior aos demais classificadores da mesma categoria.

Na metodologia deste trabalho foram abordados apenas os classificadores da categoria Árvore de Decisão, por serem um meio simples e eficiente de revelar informações baseadas nos valores dos atributos de um conjunto de dados.

Para trabalhos futuros, é sugerida a aplicação desta metodologia em novas áreas de estudo e a avaliação de outras técnicas de Mineração de Dados para a extração automática de redes de drenagem.

5. Referências

Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. **Classification and regression trees**. Monterey: Wadsworth and Brooks, 1984. 358p.

De'ath, G.; Fabricius, K. E. Classification and regression trees: A powerful yet simple technique for ecological data analysis. **Ecology**, v. 81, n. 11, p. 3178−3192, 2000.

Fayyad, U.; Piatesky-Shapiro, G.; Smyth, P.; Uthurusamy, R. **Advances in knowledge discovery and data mining.** Cambridge: MIT Press, 1996.560 p.

O'Callaghan, J. F.; Mark, D. M. The extraction of drainage networks from digital elevation data. **Computer Vision, Graphics and Image Processing**, v. 28, p. 323−344, 1984.

Rennó, C. D.; Banon, L. C. **DEM_ATTRIBUTES**. 1.0. São José dos Campos: INPE, 2012a. Disponível em: <http://urlib.net/sid.inpe.br/mtc-m19/2012/09.03.01.24>. Acesso em: 02 set. 2012.

Rennó, C. D.; Banon, L. C. **DRAINAGE_CORRECTION**. 1.0. São José dos Campos: INPE, 2012b. Disponível em: <http://urlib.net/sid.inpe.br/mtc-m19/2012/09.04.02.09>. Acesso em: 02 set. 2012.

Rennó, C. D.; Banon, L. C. **DRAINAGE_SIMPLIFICATION**. 1.0. São José dos Campos: INPE, 2012c. Disponível em: <http://urlib.net/sid.inpe.br/mtc-m19/2012/09.04.02.11>. Acesso em: 02 set. 2012.

Rennó, C. D.; Banon, L. C. **TREE_CLASSIFICATION**. 1.0. São José dos Campos: INPE, 2012d. Disponível em: <http://urlib.net/sid.inpe.br/mtc-m19/2012/09.04.02.06>. Acesso em: 02 set. 2012.

Rennó, C. D.; Nobre, A. D.; Cuartas, L. A.; Soares, J. V.; Hodnett, M. G.; Tomasella, J.; Waterloo, M. J. HAND, a new terrain descriptor using SRTM-DEM: Mapping terra-firme rainforest environments in Amazonia. **Remote Sensing of Environment**, v. 112, p. 3469-3481, 2008. doi: <10.1016/j.rse.2008.03.018>.

Shary, P.A., Sharaya, L.S., Mitusov, A.V., Fundamental quantitative methods of land surface analysis. **Geoderma**, v. 107 (1–2), p. 1–32, 2002.

Soille, P.; Vogt, J.; Colombo, R. Carving and adaptive drainage enforcement of grid digital elevation models. **Water Resources Research**, v. 39(12), p. 1366−1375, 2003.

Witten, I.; Frank, E. **Data mining:** pratical machine learnig tools and techniques.2. Ed. San Francisco, CA: Morgam Kaufmann Publishers, 2005. 524 p.

Wood, J.D. **The geomorphological characterisation of digital elevation models.** 1996. Thesis - University of Leicester, UK, 1996. Disponível em: <http://www.soi.city.ac.uk/~jwo/phd>. Acesso em: 02 set. 2012.