Diferentes classificadores na predição de classes de solos em mapeamento digital

Cristiano Cassiano da Silva [1]
Stanley Robson Medeiros Oliveira [2]
Samuel Fernando Adami [1]
Rafael Castro Crivelenti[3]
Ricardo Marques Coelho [1]

[1] Instituto Agronômico de Campinas – IAC
Avenida Barão de Itapura, 1.481, Jardim Guanabara
CEP 13012-970 Campinas, SP.
ccsilva2@yahoo.com.br, samuel@iac.sp.gov.br, rmcoelho@iac.sp.gov.br

[2] Embrapa Informática Agropecuária
Avenida André Tosello, 209, Barão Geraldo
CEP 13083-886 Campinas, SP.
stanley@cnptia.embrapa.br

[3] Coordenadoria de Biodiversidade e Recursos Naturais, SMA-SP
Rua Barão do Bananal, 1950, Jardim Anhanguera
CEP 14092-000 Ribeirão Preto, SP.
grilasso@hotmail.com

ABSTRACT: The study had as objective to develop techniques for digital soil mapping with support of main parameters of relief descriptors, of geologic map and pedological map pre-existing of Dois Córregos (SP, BRAZIL) sheet (1:50,000 scale), using data mining techniques. It was built a database from digital topographic and thematic, and data from soils and geology. Were calculate geomorphometric slope, curvature in plan and in profile and diagonal distance of the drainage area of study. These parameters and the geological map units were crossed through georeferencing the pedological map, enabling the construction of a matrix relating soil mapping units with original caption and simplified legend to the topography and geology parameters of reference areas.. This matrix was analyzed by three different techniques of machine learning, decision trees, k-NN and Naive Bayes, who predicted the soil mapping units. We evaluated soil mapping units accuracy individually and overall maps accuracy. Our results demonstrate that increasing number of records for training the algorithm increased the individual mapping units accuracy and maps. The decision tree algorithms and k-NN had the highest accuracy in both types of legend, but low in relation to training maps.

Palavras-chave: pedologia, mineração de dados, sistemas de informação geográfica

1. INTRODUÇÃO

As informações de um levantamento pedológico são importantes para o planejamento urbano e rural. A escassez de recursos e de tempo leva à necessidade de adoção de novos métodos que tornem os levantamentos de solos mais ágeis e menos onerosos (McBratney et al., 2003).

O mapeamento digital de solos, alternativa rápida e econômica em relação ao método tradicional de mapeamento (McBratney et al, 2003), pode ser definido como a criação de sistemas espaciais de informação, utilizando modelos numéricos para a inferência das variações espaciais dos tipos de solos, a partir de observações e conhecimento dos solos e de variáveis ambientais correlacionadas, como as variáveis geomorfométricas declividade e curvaturas, dentre outras (Moore, 1993).

Uma das vantagens do mapeamento digital com base no conhecimento dos padrões regionais de solos é a predição da ocorrência de tipos de solos em áreas não

mapeadas, com uso de informações geradas previamente em áreas de referência (Lagacherie & Voltz, 2000). Algoritmos de mineração de dados são técnicas que podem ser usadas para essa predição de padrões de solos e geração de conhecimento a partir de conjuntos de dados (Han e Kamber, 2006)

Entre os algoritmos de mineração de dados, o de árvore de decisão se destaca. Ele classifica e prediz amostras desconhecidas por meio de aprendizado de máquina, ou seja, com base em registros conhecidos desenvolve-se um conjunto de treinamento, do qual então uma árvore é montada e, a partir desta árvore, pode-se classificar a amostra desconhecida sem necessariamente testar todos os valores dos seus atributos. A árvore de decisão consiste de uma hierarquia de nós internos e externos que são conectados por ramos. O nó interno, também conhecido como decisório ou nó intermediário, é a unidade de tomada de decisão que avalia através de teste lógico qual será o próximo nó descendente ou filho. Em contrapartida, um nó externo, aquele que não tem nó descendente, também conhecido como folha ou nó terminal, está associado a um rótulo ou valor. Assim, apresenta-se um conjunto de dados ao nó inicial da árvore; dependendo do resultado do teste lógico usado pelo nó, a árvore ramifica-se para um dos nós filhos e este procedimento é repetido até que um nó terminal é alcançado. A repetição deste procedimento caracteriza a recursividade da árvore de decisão (Breiman et al., 1984).

Também ganha destaque o algorítmo k-NN (K Nearest Neigbours), que possui uma forma de aprendizado baseado em instâncias, ou seja, apenas armazena os exemplos de treinamento e quando um novo exemplo precisa ser classificado, ele é comparado com os dados armazenados. É um método que classifica objetos com base em exemplos mais próximos, um objeto é classificado pelo voto da maioria de seus vizinhos, com o objeto que está sendo atribuído à classe mais comum entre os seus k vizinhos mais próximos (k é um inteiro positivo). Se k = 1, então o objeto é simplesmente atribuído à classe de seu vizinho mais próximo (Batista et al, 2003). Na fase de classificação, k é uma constante definida pelo usuário. Os vizinhos são tomados a partir de um conjunto de objetos para os quais a classificação correta é conhecida.

Outro classificador comumente utilizado é o algoritmo Naïve Bayes que é um dos mais simples classificadores probabilísticos. O modelo que é construído por este algoritmo representa um grupo de probabilidades, que por sua vez são estimadas pelo cálculo da freqüência de cada valor de característica para as instâncias dos dados de treinamento. Dada uma nova instância, o classificador estima a probabilidade de essa instância pertencer a uma classe específica, baseada no produto das probabilidades condicionais individuais para os valores característicos da instância. O cálculo exato utiliza o teorema de Bayes e é por essa razão que o algoritmo é denominado um classificador de Bayes (Martins et al, 2009). O algoritmo é também denominado de Naïve, uma vez que considera todos os atributos independentes entre si dado o valor da variável da classe. Estudos experimentais sugerem que este algoritmo tende a aprender mais rapidamente que a maioria dos algoritmos de indução e daí o seu uso na nossa análise (Witten e Frank, 2005).

Crivelenti et al, (2009) utilizaram o algoritmo de árvores de decisão para predição de solos duas áreas do estado de São Paulo (Dois Córregos e São Pedro) e obtiveram acurácias gerais de 61% e 51% respectivamente. Com a finalidade de possibilitar o delineamento de unidades homogêneas de solos, Bui et al. (2008) testaram várias metodologias de predição de mapas de solos a partir da relação destes com a posição topográfica na paisagem, geologia, grupo de vegetação e uso do solo. Dentre as metodologias testadas estão a das árvores de decisão e a Expector. Os resultados obtidos pelos autores com as duas metodologias foram relativamente próximos, mas com

desempenho melhor para as árvores de decisão, que proporcionou acurácia geral de 69 %, indicando substancial concordância do mapa predito com o mapa tradicional.

Mucherino et al., (2009) aplicaram o algoritmo k-NN a um conjunto de amostras com valores de textura de solo conhecidos para estimar parâmetros do solo como capacidade de campo e ponto de murcha permanente. Skidmore et al (1996) realizaram integração entre sistemas de informação geográfica (SIG) e sistemas bayesianos no mapeamento de cinco classes de solos florestais, em que foram utilizando modelo digital de elevação, mapa de vegetação e mapa pedológico produzido por métodos tradicionais. Os autores conseguiram uma acurácia geral de 69,8% .

O objetivo deste trabalho foi avaliar três algoritmos de classificação de dados para predição de unidades de mapeamento de solos no mapeamento digital da folha Dois Córregos (escala 1:50.000).

2.METODOLOGIA DE TRABALHO

A folha Dois Córregos na escala 1:50.000 (SF-22-Z-B-III-3) localiza-se na região central do estado de São Paulo e caracteriza-se por dois tipos climáticos predominantes: um Aw, tropical chuvoso com inverno seco e mês mais frio com temperatura média superior a 18ºC; e outro Cwa, subtropical, com inverno seco e mês mais quente com temperatura média superior a 22ºC.

O relevo é representativo de três províncias geomorfológicas, que também delimitam formações geológicas distintas (IPT, 1981a; IPT 1981b): a) Planalto Ocidental, com arenitos da formação Itaqueri; b) Cuestas Basálticas, com basaltos da formação Serra Geral e arenitos da formação Botucatu; e c) Depressão Periférica, com arenitos e folhelhos da formação Pirambóia.

Foram obtidas do trabalho de Crivelenti et al (2009) a carta topográfica 1:50.000, a carta geológica 1:1.000.000 (IPT 1981b) e a carta pedológica da quadrícula Brotas (escala 1:100.000) (Almeida, 1981) todas escaneadas, georreferenciadas e vetorizadas em formato raster. Foi utilizada a legenda de solos do mapa pedológico original, bem como uma legenda simplificada para a folha Dois Córregos, obtida pela unificação das unidades de mapeamento originais pelo 3º nível categórico do Sistema Brasileiro de Classificação de Solos (EMBRAPA, 2006) com base em conceitos geomorfológicos e pedológicos. Com base na carta topográfica foi gerado o modelo digital de elevação (MDE) com 30 m de resolução no software Ilwis (FACULTY FOR GEO-INFORMATION SCIENCE AND EARTH OBSERVATION, 2001) e a partir desse MDE foram obtidos os parâmetros geomorfométricos declividade, curvaturas em planta e perfil e distância diagonal da drenagem (Valeriano, 1999).

Após a geração desses parâmetros de relevo, eles foram cruzados com os mapas de geologia e solos da área, o que permitiu obtenção de uma matriz de dados com 770.992 linhas, em que cada linha representou um pixel (30 x 30 m) do mapa 1:50.000, contendo informações discretas de cada parâmetro.

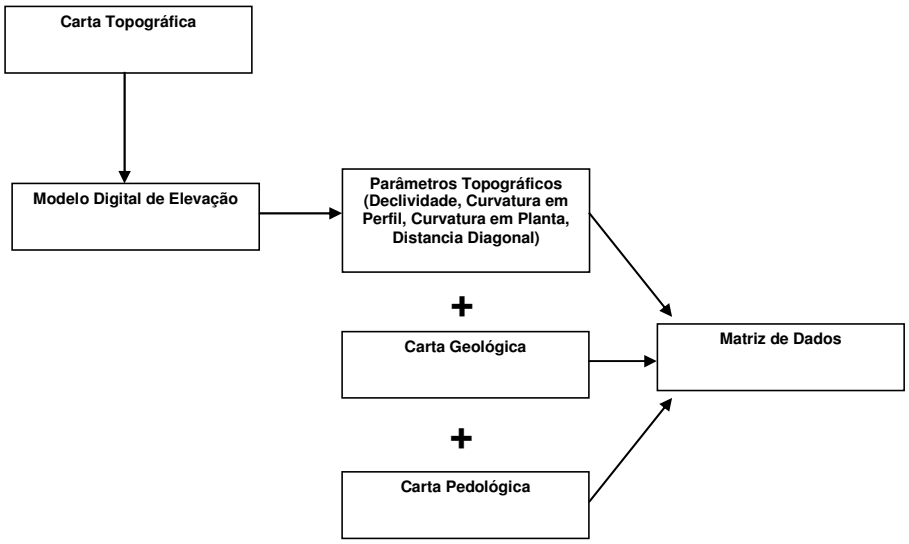

Figura 1. Esquema da obtenção da matriz de dados

No software Weka 3.5.6 (Witten e Frank, 2005), após ser realizado o pré-processamento da matriz de dados (retirada de inconsistências e padronização de dados), foram usados 90 % dos registros (linhas) da matriz de dados, escolhidos aleatoriamente, para treinamento pelos algorítmos e 10 % dos registros foram usados para validar o modelo gerado. Foram realizados treinamentos por três algoritmos distintos: árvores de clasificação, k-NN e Naive Bayes. Esses treinamentos foram realizados em três diferentes balanceamentos de classes, recurso usado para não favorecer, na geração do modelo de aprendizado, as unidades de mapeamento com maior área de ocorrência. Os balanceamentos utilizados foram 0, 0,5 e 1, que representam, respectivamente, a distribuição original dos dados, a distribuição com subamostragem das classes (unidades de mapeamento) com maior ocorrência e a distribuição com igual proporção de ocorrência para todas as classes. A validação foi feita por meio das acurácias geral do modelo e individual de cada unidade de mapeamento.

3.RESULTADOS E DISCUSSÃO

Foram obtidas 38 classes do mapa pedológico original, sendo 14 classes simples e 24 classes com associações. As acurácias gerais obtidas pelos algoritmos testados, com diferentes balanceamentos de classe são mostradas na tabela 1.

Tabela 1: Acurácia geral da predição com os três algoritmos, em três balanceamentos, do mapa pedológico com a legenda original.

Acurácia (%)			
Algorítmos	Balanceamento		
	0	0,5	1
Árvore de Decisão	44.1627	42.889	14.1234
k-NN	44.186	42.7956	14.0936
Naive Bayes	43.2652	41.8397	10.4632

Apesar de o algoritmo Naive Bayes ter tido pior desempenho nos três balanceamentos de classes, a acurácia geral nos balanceamentos 0 e 0,5 é praticamente

igual para os três algoritmos testados, com o algoritmo Naive Bayes mostrando ligeira inferioridade em relação aos demais algoritmos. Essa pequena diferença entre a acurácia nos três algoritmos se deve ao grande número de registros (pixels) que compõe o banco de dados, atenuando pequenas diferenças entre as classes.

A acurácia geral no balanceamento 1 caiu significativamente. Isto ocorreu devido a este balanceamento tratar todas as classes existentes como de igual representatividade, assim diminuindo a importância de classes muito representativas (extensas) na mesma proporção que aumenta a importância de classes pouco representativas. No balanceamento de classes 1, o Naive Bayes (10,5 %) tem desempenho nitidamente inferior aos demais algoritmos (14,1 %).

Tabela 2: Acurácia da predição por classe do mapa pedológico com a legenda original, com os três algoritmos e em três balanceamentos.

Acurácia (%)										
Símbolos da legenda	**Área**	**Árvore de Decisão**			**k-NN**			**Naive Bayes**		
		0	**0,5**	**1**	**0**	**0,5**	**1**	**0**	**0,5**	**1**
LE-3_+_LRd_+_TE-2	0,30%	0,0%	2,1%	10,2%	0,0%	10,2%	2,1%	0,0%	0,0%	0,8%
LE-1_+_LE-2	7,95%	0,0%	0,0%	0,3%	0,0%	0,3%	0,1%	0,0%	0,0%	0,7%
LV-2_+_LV-3	31,19%	90,8%	90,0%	12,9%	90,8%	12,9%	89,6%	93,3%	88,2%	0,1%
PV-2	0,68%	0,0%	1,2%	3,9%	0,2%	4,0%	1,5%	0,0%	0,2%	0,4%
Li-1+_Li-2	4,32%	83,5%	81,2%	62,1%	83,4%	61,1%	80,3%	88,5%	83,9%	78,4%
PV-1_+_LV-1	1,76%	4,9%	22,3%	15,0%	4,0%	15,0%	22,4%	8,1%	20,8%	20,1%
AQ	1,08%	18,5%	27,3%	43,5%	18,8%	43,5%	27,1%	0,0%	19,0%	19,0%
PV-1	9,29%	90,6%	67,4%	11,9%	90,7%	11,9%	67,4%	74,4%	55,8%	7,8%
PV-2_+_PV-3	5,66%	0,7%	0,5%	2,1%	0,8%	2,3%	0,6%	2,5%	1,3%	0,1%
LE-3_+_LRd_+_LRe	0,31%	0,0%	5,8%	13,6%	0,0%	13,6%	5,8%	0,0%	0,0%	0,0%
LE-2	1,03%	0,0%	0,0%	7,4%	0,0%	7,4%	0,0%	0,0%	0,0%	0,3%
LRd_+_TE-2	0,15%	0,8%	16,5%	24,0%	0,8%	24,0%	16,5%	0,0%	0,8%	20,7%
TE-2	0,02%	0,0%	38,5%	38,5%	0,0%	38,5%	38,5%	0,0%	23,1%	46,2%
LRe_+_LRd_+_LE-3	2,51%	43,4%	38,1%	33,8%	39,1%	33,8%	38,1%	2,5%	39,3%	33,4%
TE-2_+_LRe	0,42%	0,0%	1,2%	1,2%	0,0%	1,2%	1,2%	0,0%	0,0%	0,0%
PV-3_+_PV-2	1,84%	0,0%	0,2%	12,1%	0,3%	12,0%	3,1%	0,0%	2,7%	5,0%
LE-2_+_LE-1	3,59%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,5%
TE-1_+__Li-2	0,25%	0,0%	0,5%	21,8%	0,0%	21,8%	0,5%	0,0%	0,5%	4,0%
TE-2_+_LRd	2,82%	22,2%	27,7%	13,1%	26,4%	13,1%	27,7%	12,7%	31,9%	27,7%
LRd	0,81%	0,6%	1,4%	3,0%	0,8%	3,6%	1,4%	0,0%	0,0%	0,0%
LE-3	2,40%	0,1%	0,1%	0,0%	0,1%	0,0%	0,1%	0,0%	0,0%	0,0%
TE-1_+_TE-2_+_Li2	2,78%	56,0%	35,8%	3,9%	54,3%	3,9%	35,7%	45,6%	36,3%	0,1%
LV-2_+_LE-1_+_LE-2	3,20%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,8%
LRe_+_LRd	1,95%	0,0%	2,5%	0,0%	2,5%	0,0%	2,5%	0,0%	3,9%	2,7%
LRd_+_LE-3_+_TE-2	2,38%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
PV-3	1,97%	0,0%	0,0%	0,9%	0,0%	0,9%	0,0%	0,0%	0,0%	0,0%
LV-3	0,28%	0,0%	0,0%	31,4%	0,0%	31,4%	0,0%	0,0%	0,0%	0,0%
LV-4	0,71%	0,0%	0,0%	43,0%	0,0%	43,0%	0,0%	0,0%	0,0%	43,0%
LV-2_+_LE-1	1,35%	0,0%	0,0%	51,9%	0,0%	51,9%	0,0%	0,0%	0,0%	61,1%
LE-3_+_TE-2	0,21%	0,0%	0,0%	3,5%	0,0%	3,5%	0,0%	0,0%	0,0%	12,2%
LV-1	1,41%	0,0%	0,3%	34,0%	0,0%	34,0%	0,4%	0,0%	0,0%	29,6%
LV-1_+__AQ	2,42%	0,0%	45,7%	67,2%	0,0%	67,2%	45,7%	45,7%	60,4%	78,9%
TE-1_+_TE-2	1,81%	2,8%	4,6%	4,0%	2,9%	4,0%	4,6%	10,8%	6,7%	5,1%
LRe	0,09%	0,0%	1,4%	2,7%	0,0%	2,7%	1,4%	0,0%	0,0%	37,8%
TE-1	0,20%	0,0%	10,7%	12,1%	0,0%	12,9%	10,7%	0,0%	2,9%	2,1%
Li-2	0,22%	0,6%	27,8%	37,7%	0,6%	37,7%	28,4%	0,0%	17,9%	24,7%
LV-2_+_AQ	0,28%	0,0%	9,7%	26,9%	0,5%	26,4%	9,7%	0,0%	0,0%	4,6%
LRd_+_LRe	0,38%	0,0%	0,3%	4,2%	0,0%	4,2%	0,3%	0,0%	13,4%	13,4%

Os resultados da acurácia por classe (unidade de mapeamento) estão mostrados na tabela 2. De maneira geral, nos três algoritmos testados, as classes que possuem maior acurácia são as classes mais simples (com menor número de associações) e com maior área. No balanceamento de classes 0 a maior classe existente LV2+LV3 apresentou acurácias de 90,8%, 90,8% e 93,3% (árvore de decisão, k-NN e naive bayes, respectivamente) e diminuiu na medida que se aumentaram os balanceamentos, devido à menor representatividade de amostragens nos balanceamentos maiores. A classe simples PV-1, a segunda de maior extensão, apresentou acurácias de 90,6%, 90,7% e 74,4%

(árvore de decisão, k-NN e naive bayes, respectivamente), que também diminuiram na medida que se aumentou os balanceamentos, comportamento que se manteve com os três algorítmos testados. A associação de neossolos litólicos Li1+Li2 obteve acurácias de 83,5%, 83,4% e 88,5% (árvore de decisão, k-NN e naive bayes, respectivamente), porém ao contrário das outras classes elas não diminuiram significativamente com o aumento dos balanceamentos, provavelmente devido a esta classe apresentar características singulares como ocorrência em relevo acidentado, o que a torna mais facilmente caracterizável por parâmetros geomorfométricos. As classes com maior número de associações e menor número de pixels (classes menores) possuem acurácias muito baixas, ou mesmo acurácias nulas, como a classes associada LRd+LE3+TE2 que apresenta acurácia nula nos três balanceamentos e nos três algoritmos, assim como a associação LV2+LE1+LE2.

Foi também foi elaborada uma legenda simplificada para a folha Dois Córregos, unificando-se as unidades de mapeamento pelo 3º nível categórico do Sistema Brasileiro de Classificação de Solos (EMBRAPA, 2006), utilizando-se complementarmente conceitos geomorfológicos e pedológicos para unificação. Nessa nova legenda foram obtidas 16 classes, sendo 3 classes de unidades de mapeamento simples e 13 classes de unidades de mapeamento compostas (com associações).

Tabela 3: Acurácia geral da predição do mapa pedológico com a legenda simplificada, com os três algoritmos e em três balanceamentos.

Acurácia (%)			
Algoritmos	Balanceamento		
	0	0,5	1
Árvore de Decisão	51.9851 %	51.0551 %	43.3663 %
k-NN	51.989 %	50.8826 %	43.3495 %
Naive Bayes	51.1861 %	49.7114 %	40.2781 %

A acurácia geral obtida pelos algoritmos testados, com diferentes balanceamentos de classe é mostrada na tabela 3. Os balanceamentos 0 e 0,5 são praticamente iguais para os três algoritmos testados. O algoritmo Naive Bayes foi o pior nos três balanceamentos de classes testados.

Comparando-se os dois tipos de legenda (completa e simplificada), nos balanceamentos 0 e 0,5 houve ganhos de cerca de 8 % na acurácia geral nos três algoritmos na legenda simplificada. Já no balanceamento de classes 1 houve ganhos cerca de 29 % (Tabelas 1 e 3) na acurácia nos três algoritmos na legenda simplificada. O mesmo comportamento apresentado no conjunto de dados com a legenda original nos balanceamento 0 e 0,5 se repete com a legenda simplificada, ou seja, é necessário uma grande diferença entre as classes para que se reflita na acurácia geral.

Por outro lado, a acurácia geral no balanceamento 1 não caiu tão acentuadamente como na legenda original. Acredita-se que isto tenha ocorrido devido à junção de classes aumentar o número de pixels por classe e, assim, o número de registros para treinamento do algoritmo, nos três balanceamentos de classe.

Tabela 4: Acurácia da predição por classe do mapa pedológico com a legenda simplificada, com os três algoritmos e em três balanceamentos.

Acurácia (%)										
Símbolos da legenda	**Área**	**Árvore de Decisão**			**k-NN**			**Naive Bayes**		
		0	**0,5**	**1**	**0**	**0,5**	**1**	**0**	**0,5**	**1**
LE_Argilosa_Assoc	0,82%	0,0%	1,6%	5,2%	0,0%	1,6%	5,2%	0,0%	0,0%	0,0%
LE_MEDIA	12,56%	0,0%	0,0%	8,5%	0,1%	0,1%	8,5%	0,0%	0,0%	6,0%

LVA_+_LE	36,73%	91,0%	89,1%	75,6%	91,0%	88,9%	75,6%	90,1%	87,5%	66,8%
PVA	6,34%	0,6%	0,5%	4,1%	0,1%	1,5%	4,7%	2,5%	3,0%	6,5%
Litólicos	4,53%	81,5%	86,6%	86,4%	81,0%	85,6%	86,0%	84,9%	86,0%	86,3%
PVA_+_LVA	11,05%	88,1%	60,2%	35,2%	88,1%	56,8%	34,6%	69,4%	45,9%	33,5%
AQ	1,08%	1,2%	43,4%	56,4%	1,5%	43,4%	57,0%	0,0%	36,5%	37,3%
LRd_Assoc	0,53%	0,0%	2,8%	2,8%	0,0%	2,8%	2,8%	0,0%	0,0%	15,9%
NV_EUTROF_LATOSS	3,26%	0,0%	25,1%	25,1%	0,0%	25,1%	25,1%	0,0%	25,1%	29,3%
LRe_Assoc	4,46%	40,8%	43,1%	34,4%	43,1%	53,1%	34,4%	44,5%	44,5%	32,6%
PV_MEDIA	3,81%	0,0%	0,8%	5,9%	0,2%	1,3%	5,9%	0,0%	0,0%	10,7%
NV_EUTROF_ASSOC	5,03%	62,4%	54,0%	14,9%	60,2%	53,2%	15,0%	52,9%	52,9%	17,7%
LRd	3,19%	0,0%	0,1%	0,1%	0,0%	0,2%	0,2%	0,0%	0,1%	0,0%
LE_ARGILOSA	2,40%	0,0%	0,0%	0,0%	0,0%	0,1%	0,1%	0,0%	0,0%	0,0%
LVA_ARGISSOLICO	4,11%	0,0%	40,5%	58,9%	0,2%	46,5%	59,1%	40,3%	60,6%	65,9%
LRe	0,09%	0,0%	1,4%	56,8%	0,0%	2,7%	56,8%	0,0%	0,0%	41,9%

Os resultados da avaliação da acurácia por classes do mapa com a legenda simplificada estão mostrados na tabela 4. O mesmo padrão apresentado na acurácia por classe da legenda original se repete na legenda simplificada, ou seja, nos três algoritmos testados, as classes que possuem maior acurácia são as classes de unidades de mapeamento simples (com menor número de associações) e com maior número de pixels. No balanceamento de classes 0, a maior classe existente LVA+LE apresentou acurácia de 91%, 91% e 90,1% (árvore de decisão, k-NN e naive bayes, respectivamente) e diminuiu na medida que aumentaram os balanceamentos. Isto devido à menor representatividade de amostragens nos balanceamentos maiores, o mesmo acontece com a classe PVA+LVA que apresentou acurácia de 88,1%, 88,1 e 69,4% (árvore de decisão, k-NN e naive bayes, respectivamente) e também diminuiu a medida que se aumenta os balanceamentos, comportamento que se manteve com os três algorítmos testados.

4.CONCLUSÕES

a) Os algoritmos de árvore de decisão e k-NN apresentaram a maior acurácia para os dois tipos de legenda (original e simplificada) e não diferiram entre si nos três balanceamentos de classe.

b) Os balanceamentos de classes 0 e 0,5 apresentaram os melhores resultados para todos os algoritmos testados.

c) O número de registros (pixels) para treinamento do algoritmo influenciou sua acurácia. Isso foi notado (i) no balanceamento de classes, que quanto maior, menor a acurácia das classes mais representativas e (ii) na simplificação da legenda, que aumentou a acurácia nos três balanceamentos de classe.

d) As técnicas de árvores de decisão e k-NN apresentam-se como alternativas viáveis na redução de custos para elaboração de mapas digitais pedológicos preliminares, mas sua aplicação necessita de aperfeiçoamento quando o objetivo é fornecer informação de acurácia equivalente à dos mapas tradicionais, já que a acurácia obtida ainda foi baixa em relação aos mapas de treinamento.

REFERÊNCIAS BIBLIOGRÁFICAS

ALMEIDA, C.L.F. de; OLIVEIRA, J.B. de; PRADO, H. do. Levantamento pedológico semidetalhado do Estado de São Paulo: quadrícula de Brotas. I. Mapas de solos. Campinas: Instituto Agronômico, 1981. Mapa. Escala 1:100.000.

BATISTA, G.A.P.A. Pré-processamento de dados em aprendizado de máquina supervisionado.Tese (Doutorado) Universidade de São Paulo, 2003.

BREIMAN, L.; FRIEDMAN, J.H.; OLSHEN, R.A. Stone, Classification And Regression Trees. Wadsworth, 1984.

BUI, E.N.; LOUGHHEAD, A.; CORNER, R. Extracting soil-landscape rules from previous soil surveys. Australian Journal of Soil Research. 37.3 (May 1999): 495. General OneFile. Gale. CAPES, 2008.

CRIVELENTI, R. C. ; Coelho, Ricardo Marques ; ADAMI, Samuel Fernando ; Oliveira, Stanley Robson de Medeiros . Mineração de dados para inferência de relações solo-paisagem em mapeamentos digitais de solo. Pesquisa Agropecuária Brasileira (Online), v. 44, p. 1-9, 2009

EMBRAPA – EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA. Sistema Brasileiro de Classificação de Solos (SiBCS), Rio de janeiro: Embrapa Solos. 2ºEd. 316p. 2006.

FACULTY FOR GEO-INFORMATION SCIENCE AND EARTH OBSERVATION. ILWIS 3.3: user's guide. Enschede: ITC, 2001. 530p

HAN, J.; KAMBER, M. Data Mining - Concepts and Techniques. 2a edição. Nova York: Morgan Kaufmann, 2006.

IPT - INSTITUTO DE PESQUISAS TECNOLÓGICAS DO ESTADO DE SÃO PAULO. Mapa Geomorfológico do estado de São Paulo (Série Monografias, 5). São Paulo. v. 1 (Nota Explicativa) e 2 (Mapa), 1981a.

IPT (INSTITUTO DE PESQUISAS TECNOLÓGICAS DO ESTADO DE SÃO PAULO). Mapa geológico do estado de São Paulo (Série Monografias, 6). São Paulo. v. 1 (Nota Explicativa) e 2 (Mapa), 1981b

LAGACHERIE, P.; VOLTZ, M. Predicting soil properties over a region using sample information from a mapped reference area and digital elevation data: a conditional probability approach. Geoderma. v.97, p. 187–208, 2000.

MARTINS, A.C.; MARQUES, M. J.; COSTA, P.D. Estudo comparativo de três algoritmos de machine learning na classificação de dados electrocardiográficos. Tese (Mestrado) Universidade do Porto, 2009

McBRATNEY, A.B.; MENDONÇA SANTOS, M.L.; MINASNY, B. On digital soil mapping. Geoderma. v.117, p. 3-52, 2003.

MOORE, I. D.; GESSLER, P.E.; NIELSEN, G.A.; PETERSON, G.A. Soil attribute prediction using terrain analysis. Soil Science Society of America Journal, v.57, p.443-452, 1993.

MUCHERINO, A.; PAPAJORGJI, P.J.; PARDALOS, P.M. A Survey of Data Mining Techniques Applied to Agriculture, Operational Research: An International Journal 9 (2), 121–140, 2009

SKIDMORE, A.K., WATFORD, F., LUCKANANURUG, P., RYAN, P.J. An operational GIS expert system for mapping forest soils. Photogrammetric Engineering and Remote Sensing 62, 501–511, 1996

VALERIANO, M.M. Estimativa de variáveis topográficas para modelagem da perda de solos por geoprocessamento. 1999. 172p. Tese (Doutorado). Universidade Estadual Paulista, "Júlio de Mesquita Filho", Rio Claro, 1999.

WITTEN, I.H., FRANK, E. Data mining: practical machine learning tools and techniques. 2nd edition. San Francisco: Morgan Kaufmann; 2005.