USO DA REDE NEURAL SOM PARA MELHORIA DA QUALIDADE DE AMOSTRAS DE DADOS SATELITAIS

Alex. P. A. de Oliveira[1,3]*, Marcos A. S. da Silva*[1]*, Vinícius L. Xavier*[2] *e Júlio A. Navoni*[3]

[1]Embrapa Tabuleiros Costeiros, Av. Gov. Paulo B. de Menezes, 3250, 49025-040, Aracaju, SE
{alex.oliveira,marcos.santos-silva}@embrapa.br, [2]Universidade Federal de Sergipe, São Cristovão, SE
vinicius.xavier@dcomp.ufs.br, [3]Universidade Federal do Rio Grande do Norte, Natal, RN navoni.julio@gmail.com

RESUMO

O objetivo deste trabalho foi desenvolver um metodo de classificação o qual permita a eliminação amostras mal classificadas aumentando a acurácia de classificadores de imagens de satélite voltados para a identificação das mudanças de uso e cobertura da terra (LUCC). A proposta baseia-se num modelo que combina a rede neural não-supervisionada Mapa Auto Organizável para ordenamento dos dados e a teoria bayesiana para determinar e seleccionar os dados utilizados baseados na qualidade da informação sem a necessidade de passos ulteriores de validação por um especialista humano. O método proposto prevê a aplicação do algoritmo de eliminação de dados atípicos baseado na densidade dos pontos amostrais antes do treinamento da rede neural, a aplicação de um algoritmo de clusterização sobre a mesma rede neural antes de aplicar a teoria bayesiana. A proposta baseia-se em proposta anterior que combina a rede neural não-supervisionada Mapa Auto Organizável para ordenamento dos dados e a teoria bayesiana para determinar quais observações devem ser mantidas, removidas ou marcadas para avaliação por especialistas.

Palavras-chave – Cerrado, Inferência Bayesiana, Mapa Auto-Organizável, MODIS.

ABSTRACT

This work aimed to develop a classification method that eliminates misclassified samples, increasing the accuracy of satellite image classifiers aimed at identifying land use and land cover changes (LUCC). The proposal is based on a model that combines the unsupervised Self-Organizing Map neural network for data ordering and Bayesian theory to determine and select the data used based on the quality of the information without the need for further validation steps by a human expert. The proposed method involves the application of the outlier elimination algorithm based on the density of the sample points before training the neural network and the application of a clustering algorithm on the same neural network before applying Bayesian theory. The proposal builds on a previous proposal that combines the unsupervised neural network Self-Organizing Map for data ordering and Bayesian theory to determine which observations should be kept, removed, or marked for expert evaluation.

Key words – *Bayesian Inference, Cerrado, MODIS, Self-Organizing Map.*

1. INTRODUÇÃO

O processo de classificação supervisionada de imagens de satélites, por meio de rotulação dos *pixels*, depende fortemente da qualidade dos dados amostrais. Os ruídos na amostra, rotulação incorreta de alvos, prejudicam a acurácia da classificação, mesmo quando usamos classificadores menos sensíveis a ruídos como o *Random Forest* [1].

Os ruídos na base amostral podem ter origem na coleta manual, geralmente realizadas em atividades de campo e levantamentos via inspeção visual, no uso de bases de dados referenciais com erros ou desatualizadas, ou via processos automatizados e não-supervisionados de geração de amostras. A qualidade da amostra depende, evidentemente, do processo de construção das amostras, incluindo o projeto amostral e de tratamento e limpeza desses dados [2].

Para o caso de termos uma base amostral rotulada disponível, a estratégia mais usada para melhorar sua qualidade é por meio de remoção de observações com ruídos. Esta remoção pode levar em consideração a detecção de mudanças [3], o ranqueamento das incertezas da classificação [4] e o uso de medidas de distância entre as observações, objetivando identificar amostras díspares dos conjuntos de classes [5].

A depender do número de classes, do tamanho da amostra e da não-linearidade do conjunto de dados, pode ser necessário o uso de técnicas mais elaboradas para a detecção de ruídos como na proposta do uso do conceito de *ensemble* e *Random Forest* proposto por Feng et al. [1], ou por meio de redes neurais profundas pré-treinadas [6].

O método proposto por Santos et al. [5] usa a rede neural não-supervisionada SOM [7] para fazer um mapeamento topológico dos dados amostrais na grade neural de duas dimensões, agregando os dados por proximidade em termos de distância euclidiana, e depois aplica a teoria de *Bayes* para identificar observações que merecem ser excluídas, ou reavaliadas por um especialista. Embora este método apresente as limitações inerentes ao uso de medidas de distância (dimensionalidade e custo computacional) como alertado por [1], se mostrou eficiente na identificação de amostras com ruídos, e duvidosas, para um conjunto de dados amostrais com classes apresentando alto grau de confundimento [5]. Neste caso é necessário usar outro mecanismo para decidir se as amostras duvidosas devem ser removidas ou rotuladas corretamente.

Neste artigo propomos um método baseado na proposta de Santos et al. [5], de forma a diminuir, ou até mesmo eliminar, o número de amostras consideradas duvidosas, suprimindo, assim, esse segunda etapa necessária na proposta original, tendo como referência a mesma base de dados amostrais

avaliada em [5].

2. MATERIAL E MÉTODOS

2.1. Dados amostrais do Cerrado

A base de dados amostrais é composta de séries temporais extraídas do sensor *MODIS* (produto *MOD13Q1*, coleção 6) do satélite Terra da NASA, com resolução espacial de 250 metros e intervalo de 16 dias e contém quatro índices: o Índice de Vegetação Normalizada (*NDVI*), o Índice de Vegetação Melhorado (*EVI*), a banda de infravermelho próximo (*NIR*) e a banda de infravermelho médio (*MIR*). As amostras de treinamento possuem as seguintes componentes: *id, latitude, longitude, start_date, end_date, label, ndvi01, ndvi02, ..., ndvi23, evi01, evi02, ..., evi23, nir01, nir02, ..., nir23, mir01, mir02, ..., mir23*. Esses valores representam medições periódicas dos índices para diferentes localizações. As referidas amostras foram coletadas por meio de levantamentos de campo e interpretação de imagens de alta resolução por especialistas do Instituto Nacional de Pesquisas Espaciais (*INPE*), *Embrapa* e parceiros. Este conjunto de dados cobre o período de 2000 a 2017 e inclui 50.160 amostras de uso e cobertura do solo, divididas em 12 classes, desbalanceadas conforme a Tabela 1.

Tabela 1: Quantitativo e percentual das mostras da área do Cerrado por classe de uso e cobertura. Fonte: Santos et al. [5].

Classe	Quantitativo	Percentual (%)
Dense Woodland	9966	19.87
Dunes	550	1.10
Fallow-Cotton	630	1.26
Millet-Cotton	316	0.63
Pasture	7206	14.37
Rocky Savanna	8005	15.96
Savanna	9172	18.29
Savanna Parkland	2699	5.38
Silviculture	423	0.84
Soy-Corn	4971	9.91
Soy-Cotton	4124	8.22
Soy-Fallow	2098	4.18

2.2. Proposta de melhoria da qualidade de dados amostrais [5]

Santos et al. [5] propuseram um método de melhoria da qualidade de dados amostrais rotulados para fins de classificação por *pixel* de imagens de satélite. Este método pode ser dividido em três etapas. Na primeira, os dados são mapeados numa grade bidimensional a partir da rede neural não-supetrvisionada Mapa Auto-Organizável [7]. Este mapeamento topológico visa ordenar as observações de forma que a proximidade no espaço de entrada *d-dimensional* seja respeitada no espaço bidimensional da rede neural. Assim, observações próximas segundo a distância euclideana no espaço de entrada tenderão a se associar a neurônios próximos no espaço bidimensional neural. Na segunda etapa, cada neurônio será rotulado com a classe mais frequente das observações associados a ele a partir do voto majoritário. A partir dessa informação são calculadas as probabilidades a posteriori ($\mu_{\text{posterior}}$) e a condicional (μ_{prior}) de uma certa classe estar associada a um determinado neurônio, considerando a vizinha de cada neurônio. Na última etapa é definido um limiar (τ_c) para cada classe, de forma que, se a probabilidade condicional estiver abaixo desse valor a observação é eliminada (**Remover**), e, caso seja superior, observa-se a probabilidade a *posteriori*, se for maior ou igual que esse mesmo limiar mantem a observação (**Manter**), caso contrário essa observação é considerada um dado atípico que deverá ser analisado (**Marcar**), Eq. 1.

$$\text{Ação} = \begin{cases} \text{Remover}, & \text{se } \mu_{\text{prior}} < \tau_c \\ \text{Manter}, & \text{se } \mu_{\text{prior}} \geq \tau_c \text{ e } \mu_{\text{posterior}} \geq \tau_p \\ \text{Marcar}, & \text{se } \mu_{\text{prior}} \geq \tau_c \text{ e } \mu_{\text{posterior}} < \tau_p \end{cases} \quad (1)$$

Com base nas probabilidades *a priori* (τ_c), e *a posteriore* (τ_p) os neurônios são marcados para serem mantidos, excluídos, ou analisados por um analista. Os autores atingiram uma acurácia de classificação de 98.4% para os dados amostrais do cerrado usando o *Random Forest*, definido por [8]. Tendo sido necessária a análise por um especialista humano, das observações "marcadas para análise".

2.3. Método proposto nesta pesquisa

A proposta de Santos et al. [5] parte do princípio tácito de que os ruídos nas amostras são de natureza única e, portanto, passíveis de serem detectados pelo método proposto. Ademais, os autores utilizam a propriedade da rede neural *SOM* de ordenar os dados na sua grade neural bidimensional segmentando-a a partir do método de classe majoritária. No entanto, no processo de análise das observações consideram-se todos os vizinhos dos neurônios, desconsiderando que é possível que hajam neurônios vizinhos que estejam associados a classes com padrões distintos.

Desta forma propomos três melhorias na proposta de [5]. A primeira é a aplicação do algoritmo *Local Outlier Factor* para detecção de dados atípicos, e eliminação inicial de observações efetivamente fora dos padrões do conjunto de dados [9]. A segunda é a aplicação de uma algoritmo de clusterização sobre os pesos da rede *SOM*, de forma a segmentar a rede de acordo com a distância euclideana, identificando neurônios que são vizinhos, mesmo perencendo a classes distintas. Por último, sugerimos que as probabilidades condicionais e a *prosteriori* sejam calculadas considerando neurônios vizinhos que pertencem ao mesmo cluster, Fig. 1, e não a toda a vizinha, como proposto por [5].

O aprendizado sequencial ou estocástico da rede *SOM* é sensível aos seus hiperparâmetros, sobretudo às dimensões (NxM), o formato da grade (retangular ou hexagonal), ao raio inicial da vizinhança (sigma) e ao valor da taxa de aprendizagem. Assim, para determinar qual o melhor conjunto de hiperparâmetros, avaliamos uma combinação de deles e usamos o erro de quantização vetorial para decidir qual a melhor rede *SOM* para o conjunto de dados do Cerrado.

Definiu-se que as redes SOM avaliadas teriam grade haxagonal com vizinhança Gaussiana. O teste considerou uma combinação das seguintes variações de hiperparâmetros: foram avaliadas onze configuração distintas para o tamanho

NxM da *SOM* bidimensional ($15x15$, $18x18$, $15x25$, $20x20$, $25x25$, $30x30$, $25x50$, $40x40$, $50x50$, $60x60$, $50x75$), quatro valores para o raio inicial de aprendizagem (1, 1.5, 2, 2.5, 3) e dez valores para a taxas de aprendizagem (0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5). A combinação de todas essas parametrizações resultaram em 550 testes, que revelaram o valor 0.449 como sendo o menor erro de quantização do *SOM*, que serviu para eleger a seguinte parametrização, para ser utilizada nesta pesquisa: topology='hexagonal'; neighborhood_function='gaussian'; dimension='60x60'; sigma=1.5; learning_rate=0.45.

Para a tarefa de clusterização avaliamos quatro métodos comumente usados na clusterização dos pesos da rede *SOM*, k-médias, hierárquico aglomerativo, *DBSCAN* e *HDBSCAN*. Para decidir sobre a melhor partição do conjunto de dados rotulados usamos os índices *ACC* para análise de clusters, *NMI* e *ARI*.

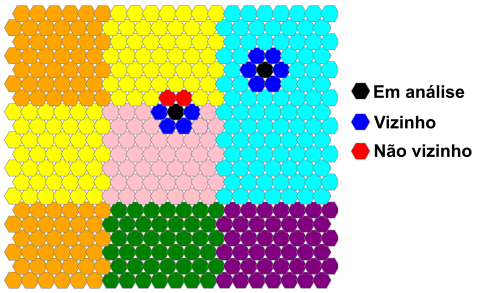

Figura 1: Ilustração de uma rede neural SOM 2D 20x20 segmentada por algum algoritmo de clusterização. Na proposta de classidicação das observações baseado na inferência *bayesiana* de [5] todos os vizinhos do neurônio em análise são considerados. Na nossa proposta serão considerados apenas os vizinhos que pertecentrem ao mesmo cluster. Fonte: elaborado pelos autores.

Para que os resultado fossem comparáveis com a proposta de Santos et al. [5] usamos o classificador *Random Forest*, antes e após a eliminação de observações, e, também, na fase inicial de centralização dos dados com o $StandardScaler$, e de remoção inicial dos *outliers* com o $LocalOutlierFactor$.

É possível descrever o método proposto em duas etapas. Na primeira etapa o objetivo é melhorar a qualidade das amostras satiletais, utilizando uma rede *SOM*, e uma inferência *bayesiana* personalizada, conforme descrito no Algoritmo 1. Na segunda etapa, o objetivo é identificar, empiricamente, uma parametrização mais performática para o classificador (*Random Forests*), e, para isso, foi considerada a combinação dos seguintes hiperparâmetros: duas possibilidades de *bootstrap* (*true* e false), duas variações de *criterion* (*gini* e *entropy*), quatro variações de *max_depth* (10, 20, 30, none), três variações de *min_samples_leaf* (1, 2 e 4), três variações de *min_samples_split* (2, 5 e 10) e três variações de *n_estimators* (100, 200 e 500). Foram realizados 1853 testes. E, considerando os testes realizados, a melhor configuração percebida foi: bootstrap=false; criterion='entropy'; max_depth=30; max_features='sqrt'; min_samples_leaf=1; min_samples_split=2; e, n_estimators=200.

Algoritmo 1 Remoção de ruído em amostras satelitais, por clusterização de *SOM* e inferência *bayesiana* personalizada

Requer: X – Dados satelitais rotulados
Requer: ACC_{meta} – Meta de acurácia para o classificador
Procedimento REMOVERRUIDOS(X, ACC_{meta})
$X^{'} \leftarrow SS(X)$ // Centralizar X com *StandardScaler*
$X^{''} \leftarrow LOF(X')$ // Aplicar *Local Outlier Factor*
$H \leftarrow MH(SOM(X^{''}))$ // Melhores hiperparâmetros
$S \leftarrow SOM((X^{''}, H))$ // Criar *SOM* com $X^{''}$ e H
$A \leftarrow AC(S))$ // Definir algoritmo do clusterizador
$X^{i} \leftarrow X^{''}$ // Inicializar dados para loop de remoção
Repita
$S \leftarrow SOM((X^{i}, H))$ // Criar *SOM* com X^{i} e H
$C \leftarrow A(S)$ // Criar cluster com algoritmo definido
$E \leftarrow IB(S, C)$ // Definir mapa de exclusões
$P \leftarrow KF(X^{i})$ // Aplicar *k-Fold Cross Validation*
$Res \leftarrow RF(P)$ // Resultado com *Random Forests*
$ACC_{i} \leftarrow MC(Res)$ // Avaliar acurácia
Se Então $ACC_{i} < ACC_{meta}$
$X^{i} \leftarrow REM(X^{i}, E)$ // Remover *outliers*
end Se
Até $ACC_{i} \geq ACC_{meta}$ // Condição de parada
Retorna X^{i} // Retorno
end Procedimento

3. RESULTADOS E DISCUSSÃO

A combinação total de 11 dimensões, 5 sigmas e 10 taxas de aprendizados diferentes resultou numa bateria de 550 testes, considerando dois hiperparâmetros fixos, conforme explanado nesta seção. Esses testes revelaram o menor erro de quantização na dimensão 60 x 60, conforme Figura 2.

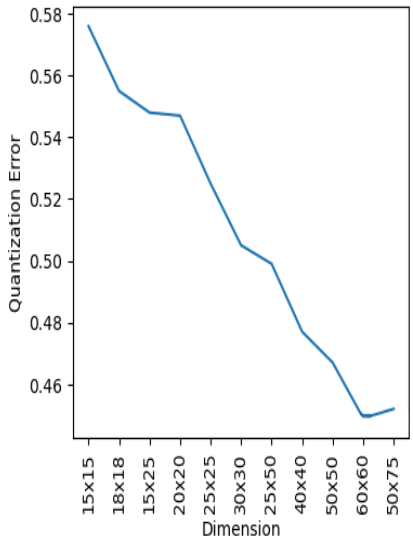

Figura 2: Erro de Quantização x Dimensão NxM das redes neurais avaliadas. Fonte: resultados da pesquisa.

Na Tabela 2 é possível verificar um comparativo entre os algoritmos analisados. Para todos os índices de análise da qualidade da clusterização para dados rotulados (ACC, NMI e ARI) o algoritmo k-médias se mostrou o mais eficiente, tendo sido adotado para as etapas posteriores do processo.

Tabela 2: Resultados para a eficiência dos clusterizadores considerando os índices ACC, NMI e ARI. Fonte: resultados da pesquisa.

Algoritmo de clusterização	ACC	NMI	ARI
HDBSCAN	0.2121	0.0616	0.0093
DBSCAN	0.2660	0.3374	0.1394
Hierárquico aglomerativo	0.6760	0.6317	0.5149
K-médias	**0.6790**	**0.6559**	**0.5610**

Na Fig. 3 temos o resultado das acurácias médias obtidas por meio de validação cruzada k-fold ($k = 10$) para todas as iterações que incluem o treinamento dos dados sem outliers à rede SOM 2D 60x60, a clusterização dessa rede SOM por meio do k-médias, a aplicação do método de inferência bayesiana considerando somente os vizinhos do mesmo cluster, a eliminação dos dados rotulados como ruídos e cálculo da acurácia para o classificador *Random Forest*. Importante destacar o incremento da acurácia a cada iteração, sendo que em nenhum momento qualquer observação da amostra foi rotulada para análise posterior por especialista.

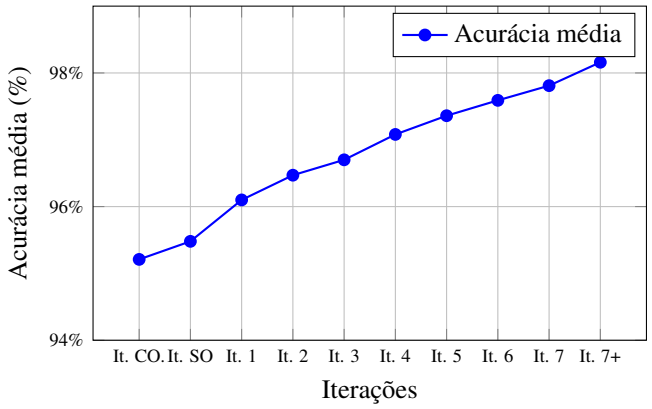

Figura 3: Evolução da acurácia durante as iterações do método proposto. Iteração com outlier (It. CO), iteração sem outliers segundo o algoritmo LOF (It. SO), iterações para eliminação de dados atípicos a partir da inferência baysesiana (It. 1-7) e iteração com ajuste fino dos parâmetros do Random Forest (It. 7+) . Fonte: resultados da pesquisa.

Tabela 3: Número de amostras mantidas a partir do método proposto neste artigo e comparativo entre esse percentual de eliminção de amostras (A) e o percentual obtigo por [5] (B). Fonte: resultados da pesquisa.

Rótulo	Original	Mantido	A	B
Dense Woodland	9966	8104	18,68%	9,00%
Dunes	550	534	2,91%	0,00%
Fallow-Cotton	630	181	71,27%	81,00%
Millet Cotton	316	219	30,70%	67,00%
Pasture	7206	5655	21,52%	13,60%
Rocky-Savanna	8005	6660	16,80%	11,50%
Savanna	9172	7709	15,95%	9,30%
Savanna Parkland	2699	2375	12,00%	4,60%
Silviculture	423	323	23,64%	19,90%
Soy-Corn	4971	4328	12,94%	9,40%
Soy-Cotton	4124	3510	14,89%	4,60%
Soy-Fallow	2098	1448	30,98%	41,80%

Na Tabela 3 temos o total de amostras mantidas e o percentual de eliminação para o método proposto por [5] e o proposto neste artigo. Observa-se que o método proposto nesta pesquisa eliminou menos amostras dos rótulos "Soy-Fallow", "Fallow-Cotton"e "Millet Cotton"onde reduzimos em apenas $30,70\%$ enquanto o método proposto por Santos et al. [5] eliminou $67,00\%$.

4. CONCLUSÕES

O método proposto nesta pesquisa revelou resultados satisfatórios para a eliminação de ruídos em séries temporais de imagens de satélite. Foi possível alcançar acurácia de 98.15% ao final do processo para a base de dados avaliada, sendo que nenhuma observação tenha sido rotulada para análise posterior. Apesar do custo computacional o método proposto elimina os dados ruidosos sem a necessidade de ter de recorrer a um especialista do domínio para avaliação de observações duvidosas.

Trabalhos futuros incluirão o uso de métodos mais robustos para determinação dos hiperparãmetros da rede neural SOM e avaliação do método proposto sobre outras bases de dados amostrais rotulados.

5. REFERÊNCIAS

[1] W. Feng, X. Gao, S. Boukir, Z. Xie, Y. Quan, W. Huang, and M. Xing. Hypothesis margin-based ensemble method for the classification of noisy remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–21, 2023.

[2] D. Moraes, M. L. Campagnolo, and M. Caetano. Training data in satellite image classification for land cover mapping: a review. *European Journal of Remote Sensing*, 57(1):2341414, 2024.

[3] K. J. Wessels, F. Van den Bergh, D. P. Roy, B. P. Salmon, K. C. Steenkamp, B. MacAlister, D. Swanepoel, and D. Jewitt. Rapid land cover map updates using change detection and robust random forest classifiers. *Remote Sensing*, 8(11), 2016.

[4] Z. S. Venter and M. A. K. Sydenham. Continental-scale land cover mapping at 10 m resolution over europe (elc10). *Remote Sensing*, 13(12), 2021.

[5] L. A. Santos, K. R. Ferreira, G. Camara, M.C.A. Picoli, and R.E. Simoes. Quality control and class noise reduction of satellite image time series. *ISPRS Journal of Photogrammetry and Remote Sensing*, 177:75–88, 2021.

[6] X. Zhao, D. Hong, L. Gao, B. Zhang, and J. Chanussot. Transferable deep learning from time series of landsat data for national land-cover mapping with noisy labels: A case study of china. *Remote Sensing*, 13(21), 2021.

[7] Teuvo Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69, 1982.

[8] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

[9] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. Lof: Identifying density-based local outliers. *ACM SIGMOD Record*, 29(2):93–104, 2000.