Novo algoritmo de classificação automática de dados multidimensionais para identificação de comportamentos, limiares de decisão e *outliers* com potencial utilização para dados de sensores remotos

Lorena Gayarre Peña [1,2]
Liana Oighenstein Anderson [3,4]
Guilherme Conceição Rocha [2]
Luiz E.O.C. Aragão[1]

[1] Instituto Nacional de Pesquisas Espaciais - INPE
Caixa Postal 515 - 12227-010 - São José dos Campos - SP, Brasil
lorenagayarre@gmail.com
{laragao}@dsr.inpe.br

[2] Konatus Soluções Inteligentes
Ed. Vip Center 11 andar, Sala 1108
12245-820 – São José dos Campos – SP, Brasil
guilherme.rocha@konatus.com.br

[3] Centro Nacional de Monitoramento de Desastres Naturais – CEMADEN
Parque Tecnológico de São José dos Campos, Estrada Doutor Altino Bondensan, 500, São José dos Campos - São Paulo, 12247-016
{liana.anderson@cemaden.gov.br}

[4] Environmental Change Institute, ECI, University of Oxford
South Parks Road, Oxford, OX1 3QY, UK
{liana.anderson@ouce.ox.ac.uk}

Abstract. A scientific research starts with a data processing. This process can be divided in three steps which, depending on their characteristics can be applied in a sequential fashion: 1) Outliers research (data that can be considered erroneous), 2) Behaviour identification, and 3) Behaviour threshold definition. Most of the times, the success of a research depends on an adequate accomplishment of the three steps mentioned above, defining accurate thresholds which provide confiability and decreases errors. Many times, this work is carried out by using the manual trial and error methodology until the optimal thresholds are found. Usually, these thresholds must be adjusted, and this process may be repeated many times in case one wants to apply the results to another dataset. This study proposes a new multi-dimensional data automatic classification algorithm which can be used in an exploratory analysis. This algorithm provides a data characterization, pointing out outliers, determining decisions thresholds and identifying data behaviours using less time than the required to do it via the trial and error methodology. In this research, the algorithm is applied to three remote sensing study cases, demonstrating both time and human resources economy and validating the results.

Palavras-chave: Detecção automática multidimensional, limiar, *outliers*, classificação, economia.

1. Introdução

A metodologia de tentativa e erro tem uma grande aplicação em todas as áreas de pesquisa e, cada vez mais, objetiva-se trocar essa metodologia por outras mais eficientes que ofereçam os mesmos resultados ou até mais robustos, visando a economia financeira, computacional e de tempo.

Mesmo assim, existem muitas áreas de pesquisa que ainda não conseguiram automatizar todos os processos e continuam a empregar os métodos de tentativa e erro. Este problema está presente em muitas pesquisas relacionadas ao estudo do meio-ambiente, como aquelas abordadas pelo uso do sensoriamento remoto.

A presença deste problema é devido, em grande parte, à dificuldade de criar um algoritmo automático eficiente que seja genérico o suficiente para abranger um grande número de casos de estudo. Isto ocorre porque a grande maioria dos eventos que acontecem na superfície terrestre são altamente complexos, dependendo de um conjunto grande de parâmetros (multidimensionais) que por sua vez são muito variáveis – ou seja, de difícil identificação de comportamentos bem caracterizados. Isto produz dados muito dispersos que precisam de um extensivo tratamento de dados por parte do pesquisador, feito muitas vezes de forma manual mediante a metodologia de tentativa e erro.

O processo de tratamento de dados pode ser dividido em três passos: 1) definição de *outliers* ou valores que são considerados errados; 2) determinação de grupos de comportamento dos dados; e 3) determinação dos limiares que definem esses grupos de comportamentos.

Considerando estes passos, neste trabalho é apresentado um algoritmo de classificação de dados multidimensionais que realiza 1) apontamento de possíveis *outliers*, 2) determinação de grupos de comportamentos dos dados e 3) identificação de limiares desses comportamentos.

O objetivo deste trabalho é explorar a utilização deste algoritmo que evita o uso da metodologia tentativa e erro, aplicado a dois casos de estudo com dados oriundos de sensores remotos orbitais. O primeiro caso de estudo refere-se ao uso do algoritmo para identificar corpos d´água em uma imagem Landsat, com resolução espacial de 30 m. O Segundo caso de estudo refere-se à identificação de padrões espaço-temporais da ocorrência de pixels de calor, também denominados *hot pixels*.

2. Algoritmo de classificação automática

A metodologia utilizada neste trabalho para a implementação do algoritmo de classificação automática é a *clusterização*. Clusterização – adaptado do Inglês *Clustering*, - consiste em classificar elementos em agrupados (ou *clusters*) que representam um comportamento no contexto de um problema ou característica particular. O conjunto de elementos deve ser representativo da população sendo estudada; Anil (1988).

A grande maioria dos algoritmos de *clusterização* existentes até o momento estão divididos em dois grupos. Primeiro, existem os que realizam clusterização automática, pois tem uma aplicação específica e, portanto, estão otimizados para essa aplicação sem poder utilizá-lo em outra diferente, por exemplo Arifin (2006). O segundo tipo de algoritmo refere-se aos de uso genérico, mas que precisam de conhecimento a priori dos resultados esperados para determinar os parâmetros necessários para aplicação do algoritmo (por exemplo, número de agrupados esperados ou tamanho máximo dos agrupados finais); Chandola (2009), Anil (1988).

O algoritmo apresentado neste trabalho para uso em sistemas de sensoriamento remoto é caracterizado por juntar as duas características anteriores, clusterização automática e uso genérico.

Para conseguir este resultado foi utilizada a metodologia dos histogramas, Anil (1988), que consiste em construir o histograma para cada dimensão e identificar os comportamentos dessa dimensão em função dos vales do histograma Chhikara (1979).

Para formalizar matematicamente o algoritmo, considera-se um conjunto de i=1..N amostras X={x1..xN} multidimensional com j=1..M dimensões, sendo xi={xi1, xi2 ... xiM}, xN = {xN1, xN2, XNM}. O parâmetro 'j' da amostra 'i' seria representado como 'xij'.

Para ilustrar a explicação define-se um conjunto X com cinco amostras (N=5) bidimensionais (M=2), sendo X={x1,.., xi,..,x5} e cada amostra xi = {xi1, xi2}.

O algoritmo de clusterização realiza os seguintes passos:

1. Estudar o histograma de cada dimensão 'j' para as N amostras. O estudo é realizado identificando os vales do histograma e utilizando eles para criar os clusters unidimensionais.
2. Identificar as amostras pertencentes à interseção interdimensional dos clusters obtidos.
3. Definir um cluster final para cada uma das interseções cujo número de amostras seja diferente de zero.

Como saída, o algoritmo cria dois arquivos chamados "clustersFormados.txt" (identificador do cluster (1..N), número de amostras no cluster e valores mínimo e máximo de cada parâmetro que definem cada cluster) e "dadosClassificados.txt" (relação de parâmetros de cada amostra adicionando uma coluna com o número de cluster ao que pertence essa amostra).

3. Estudos de caso

São apresentados três estudos de caso para testar a adequação do algoritmo de clusterização em diferentes aplicações do sensoriamento remoto.

3.1 Detecção de corpo d´agua

O objetivo deste estudo é identificar os comportamentos dos dados da imagem em função do nível de cinza (espera-se encontrar dois comportamentos; terra e água) e identificar os limiares em rango de cinza para cada comportamento.

3.2 Focos de Calor

Os focos de calor utilizados nesse estudo correspondem ao produto MCD14ML derivados do sensor MODIS a bordo das plataformas orbitais Terra e Aqua. A série temporal dos dados compreendem de Janeiro de 2001 a Dezembro de 2010.

O estudo de caso utilizando-se dos dados de focos de calor objetivou avaliar a existência de correlação nos dados de queimadas levando em conta diferentes parâmetros disponibilizados no produto MCD14ML (Tabela 1). Para este estudo, não existe conhecimento a priori do número de clusters esperado.

Tabela 1. Identificação das informações disponibilizadas no produto MODIS MCD14ML.

Coluna	*Nome*	*Unidade*	*Descrição*
1	YYYYMMDD	-	Ano (YYYY), Mês (MM) e Dia (DD)
2	HHMM	-	Hora (HH) e Minuto (MM)
3	Sat	-	Satélite: Terra (T) ou Aqua (A)
4	Lat	Graus	Latitude no centro do pixel de calor
5	Lon	Graus	Longitude no centro do pixel de calor
6	T 21	Kelvin	Temperatura de brilho do pixel de calor na banda 21
7	T 31	Kelvin	Temperatura de brilho do pixel de calor na banda 31
8	Sample	-	Número da amostra (entre 0-1353)

9	FRP	MW	Força radiativa do fogo (Fire Radiative Power)
10	conf	%	Grau de confiança (entre 0 e 100)

3.2.1 Brasil

O conjunto de dados estudados recobrindo o Brasil foi formado por um total de 614.500 amostras aproximadamente. Cada amostra do conjunto de está composta pelos parâmetros definidos na Tabela 1.

3.2.2 Mato Grosso

Este caso de estudo é um derivado do caso de estudo anterior. Neste caso foram determinadas as amostras de queimadas pertencentes ao Estado do Mato Grosso com intenção de estudar essa área com maior detalhamento devido ao conhecimento de campo desta região pelos autores do estudo. O numero total de amostras para o Estado do Mato Grosso corresponde aproximadamente a 177.000. Assim como no caso anterior, não existe conhecimento *a priori* dos resultados esperados.

4. Resultados e discussão

Nesta seção se mostram os resultados após aplicar o algoritmo comparando os dados brutos com os dados classificados. Para cada caso de estudo explicado na secção 3 se realiza uma interpretação dos resultados.

4.1 Resultados

4.1.1 Estudo de caso: corpos d´Água

Neste caso de estudo tinha-se um conhecimento prévio dos resultados esperados, pois ao tratar-se de uma imagem bidimensional, o ser humano tem a capacidade de visualizar os dados e prever o resultado. Foram identificados somente duas feições, a água e a superfície terrestre.

Aplicou-se o algoritmo ao conjunto de dados unidimensional caracterizado pelo parâmetro 'nível de cinza' e fez-se a plotagem da nova figura em diferentes cores em função dos clusters obtidos.

- **Comparação de dados brutos e classificados**

Apesar da análise visual indicar a presença somente de dois tipos de alvos (água e superfície terrestre), ao executar o algoritmo foram detectados três clusters; água, superfície terrestre e nuvens. Portanto, conclui-se que mesmo feições que não são significativas na imagem ou não são observadas pelo interprete são reconhecidas automaticamente pelo algoritmo.

A Figura 1a mostra uma imagem derivada do satélite Landsat TM, em nível de cinza – ou seja, as mesmas informações fornecidas ao algoritmo. A Figura 1b mostra a imagem reproduzida em três níveis de cinza; branco, cinza e preto, diferenciando os três clusters achados pelo algoritmo. Foram identificados também os valores máximo e mínimo de nível de cinza que caracterizam estes três comportamentos (limiares). Com base nestes resultados, o pesquisador pode ajustar os limiares e fazer um estudo da confiabilidade dos resultados.

Os resultados obtidos neste estudo de caso podem ser aplicados a outras imagens sendo que, se os limiares não funcionarem, não é necessário que o pesquisador realize um estudo para ajustar eles manualmente, pois pode ser aplicado o algoritmo sobre a nova imagem, consumindo não mais do que poucos segundos.

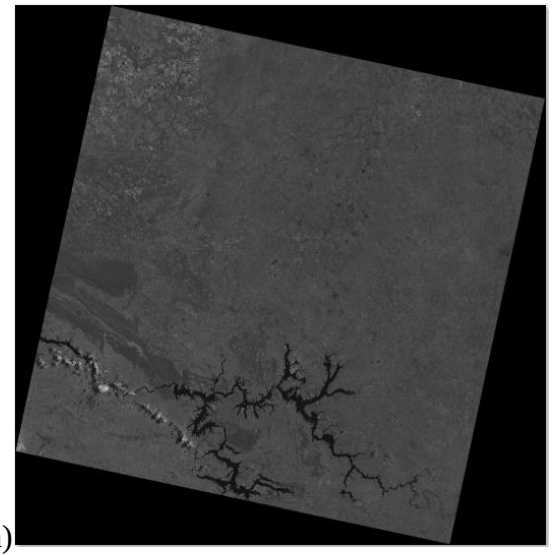
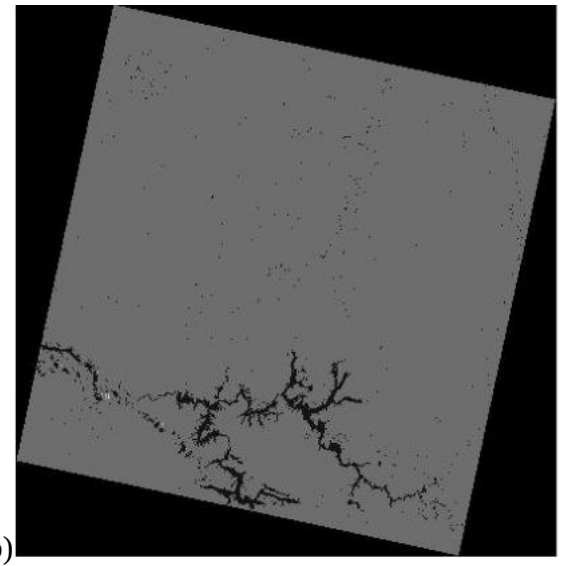
a)  b) 

Figura 1. a) Banda 5 (infravermelho próximo) oriunda do satélite Landsat-TM. b) Imagem classificada depois de aplicar o algoritmo de análise de comportamentos.

O tempo utilizado para organizar os dados, aplicar o algoritmo, identificar os três clusters e gerar a figura resultante (Figura 1b) foi de 20 minutos. Dentre eles, 15 minutos foram dedicados a formatar os dados segundo o formato aceito pelo algoritmo, e aproximadamente 1 minuto para aplicar o algoritmo e obter os clusters, e finalmente 4 minutos para gerar figura classificada.

Tabela 2. Classificação das amostras segundo nível de cinzas

ID clusters	Número de amostras	Nível de cinza (Valor mínimo)	Nível de cinza (Valor máximo)
1	11114	3	10907.2
2	407632	10907.2	45800.64
3	27	45800.64	54524

4.1.2 Focos de calor

4.1.2.1 Brasil

Devido ao grande número de amostras e desconhecimento do comportamento intrínseco delas, aplicou-se o algoritmo múltiplas vezes filtrando os dados em função dos parâmetros discretizados na Tabela 1, até obter uma classificação onde pudessem ser reconhecidos diferentes comportamentos dos dados. Essa classificação foi estudada tentando reconhecer se, efetivamente, esses clusters identificam comportamentos diferentes do padrão espaço-temporal de queimada. Finalmente, selecionaram-se as amostras considerando: filtro de confiabilidade igual a 100%, força radiativa do fogo maior que 1.000 e somente dados oriundos do satélite Terra.

O tempo de processamento do algoritmo variou entre 15 segundos e 3 minutos dependendo da quantidade de amostras utilizadas.

- **Comparação entre dados brutos e dados classificados**

Observa-se que a nuvem de pontos associada aos três parâmetros não apresentam informações palpáveis em relação à localidade, data de ocorrência ou nenhum outro padrão emergente facilmente identificável (Figura 2a). No entanto, após a utilização do algoritmo de classificação, que levou em consideração simultaneamente os dados de dia, latitude e

longitude (Figura 2b), observam-se onze clusters, em que as datas de maior ocorrência por região do país fica evidenciado. É interessante observar a presença de dois grupos de *outliers* no cluster azul e no cluster cyan, que representam locais em que somente ocorreu fogo naquele momento nas determinadas latitudes e longitudes.

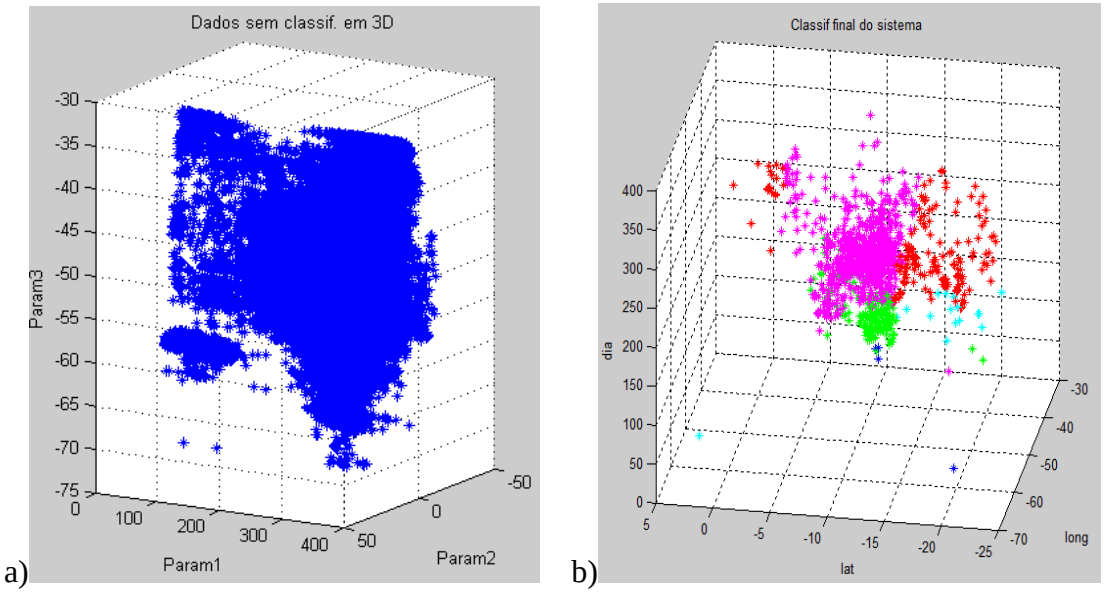


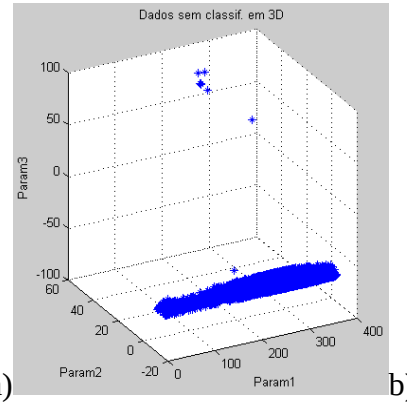
Figura 2. Distribuição de dados sem classificar (a) e classificados depois da aplicação do algoritmo (b)

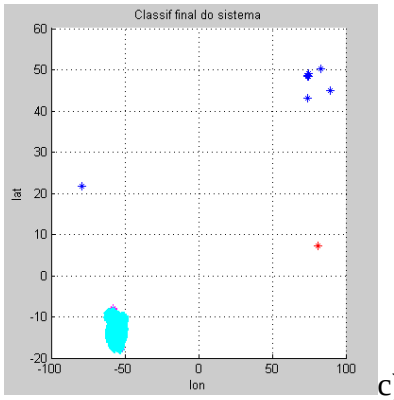
4.1.2.2 Mato Grosso

- **Comparação de dados brutos e classificados**

A Figura 3a mostra os dados brutos, e observa-se a existência de valores *outlie*rs. O sistema classificou os dados identificando os *outliers* em clusters diferentes do que as amostras localizadas no Estado do Mato Grosso (Figura 3b). Isso possibilita uma rápida identificação numérica destes valores, permitindo a eliminação deles, a partir da utilização do arquivo "dadosClassificados.txt". Já na Figura 3c mostra-se a classificação dos dados uma vez eliminados os *outliers*. É interessante notar que o cluster rosa (Figura 3c) localiza-se na região de nova fronteira do desmatamento dentro do Estado do Mato Grosso, com incidência de focos de calor em um período anterior aos focos localizados na região central do estado (cluster cyan), em áreas que a agricultura e pastagens já estão bem estabelecidas.

A utilidade de identificação de *outliers* pode ser considerada prescindível quando se trata de conjunto de dados uni/bi/tridimensionais, pois o ser humano tem a capacidade de enxergar tais elementos, No entanto, quando trata-se de casos multidimensionais, necessita-se de um aporte computacional, provido por este algoritmo.

a) 

b) 

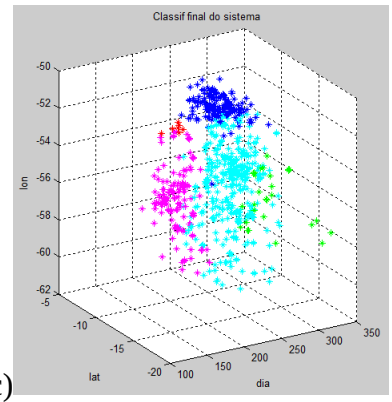
c) 

Figura 3. a) Distribuição de dados sem classificar no Mato Grosso. b) Dados classificados do Mato Grosso com *outliers*. c) Dados classificados do Mato Grosso depois da eliminação dos *outliers*.

5. Vantagens e desvantagens de utilização do algoritmo

Em base aos resultados obtidos na secção 4, nesta secção é feito um levantamento das vantagens e desvantagens da utilização do algoritmo para tratamento de dados de sistemas de sensoriamento remoto.

Tabela 3. Vantagens e desvantagens de utilização do algoritmo no tratamento de dados

Vantagens	Desvantagens
Resultados obtidos em segundos ou minutos de processamento.	Não identifica limiares quando estes não dependem do comportamento dos dados e sim da necessidade do pesquisador de definir um valor limite.
Possibilidade de realizar múltiplos estudos com diferentes parâmetros e dados até obter os resultados mais interessantes.	Um conjunto de dados de entrada muito grande pode saturar o algoritmo produzindo a obtenção de um único cluster de comportamento.
Capacidade de identificar, num mesmo processamento, erros, comportamentos e limiares de decisão.	A implementação em Matlab reduz a quantidade de dados de entrada e aumenta o tempo de execução do algoritmo devido à quantidade de memória disponível.
Descobrimento de clusters não esperados segundo o conhecimento prévio dos dados.	Criação de agrupados onde não existe uma classificação natural dos dados.
Aplicação em conjunto de dados unidimensional ou multidimensional.	
Filtro ajustável para aplicações específicas do algoritmo	
Possibilidade de estudar os dados sem saber os resultados esperados (não é necessário conhecimento a priori dos resultados esperados)	
Sem reajuste manual de limiares, o algoritmo pode ser aplicado rapidamente a cada conjunto específico dos dados	

É importante destacar que o objetivo deste algoritmo não é fazer o trabalho minucioso do pesquisador e sim assistir ao pesquisador na realização desse trabalho. Por isso este algoritmo deve ser visto como uma ferramenta de caracterização de dados e não como um resultado final.

É importante também clarificar que este algoritmo é de caráter genérico, i.e., pode ser aplicado a qualquer tipo de dado numérico, tanto unidimensional quanto multidimensional. Por este motivo, os resultados do algoritmo não são considerados de uma confiabilidade definida, pois o estudo de confiabilidade deve ser feito em base a uns dados específicos de entrada e uma classificação resultante como saída. O estudo da confiabilidade deverá ser feito após aplicação em um caso de estudo particular.

6. Conclusões

Em todos os estudos de caso, os resultados advindos da execução do algoritmo permitiram ao pesquisador realizar a análise dos dados de forma mais rápida focando o estudo nas anomalias já pré-determinadas pelo algoritmo.

O tempo de processamento do algoritmo sobre os dados demonstra a grande potencialidade exploratória da utilização deste método para grande número de amostras.

A utilização de algoritmo como ferramenta de ajuda para realizar o processamento de dados em sistemas da área de sensoriamento remoto diminui o tempo e trabalho e aumenta a confiabilidade dos resultados.

A utilização do algoritmo permite a rápida localização de dados *outliers*. Graças aos arquivos produzidos pelo algoritmo é possível eliminar essas amostras rapidamente e aplicar o algoritmo de novo nas amostras já sem erros.

Finalmente, o algoritmo apresentado encontra-se em fase final de desenvolvimento. A primeira autora tem interesse em futuras explorações com outros dados, ficando a disposição para colaborações.

7. Citações e Referências

Jain, Anil K. e Richard C. Dubes. **Algorithms for clustering data**. Prentice-Hall, Inc., 1988. ISBN:0-13-022278-X.

Chhikara, R. S., e D. T. Register. A numerical classification method for partitioning of a large multidimensional mixed data set. **Technometrics** 21.4:531-537, 1979.

Chandola, V.; Banerjee, A.; Kumar, V. Anomaly detection: A survey. **ACM Computing Surveys** (CSUR), v. 41, n. 3, p. 15, 2009.

Arifin, A. Z.; Asano, A. Image segmentation by histogram thresholding using hierarchical cluster analysis. **Pattern Recognition Letters**, v. 27, n. 13, p. 1515-1521, 2006.