Algoritmos de *clustering* para separação de culturas agrícolas e tipos de uso e cobertura da Terra utilizando dados de sensoriamento remoto

Adeline Marinho Maciel[1]
Lúbia Vinhas[1]
Gilberto Câmara[1]

[1] Instituto Nacional de Pesquisas Espaciais – INPE
Caixa Postal 515 – 12227-010 – São José dos Campos - SP, Brasil
adelsud6@gmail.com, lubia@dpi.inpe.br, gilberto.camara@inpe.br

Abstract. Remote sensing data are useful in different areas of research and application, among them agriculture, which can be used for monitoring crops and even of the support the productivity prediction of certain crops. For this, one of the most desired features is the ability to separate, or classify, in remote sensing images, the different crops observed in a given region. In order to obtain a good classification is common that are used multiple radiometric attributes available in remote sensing data. Among the various techniques and algorithms for classification are those based in clustering. However, due the high correlation among radiometric attributes and even the difficult to implement classifiers based in multiple attributes is necessary to study how reduce the dimensionality of the attributes used in the data classification. This is a work in progress that aims to exercise the use of feature selection algorithms, for reduce the dimensionality of attributes checking which attributes are more correlated with a class of interest, and of clustering algorithms in the separation of crops from other types of land use and cover, using remote sensing data. The results show that some data are easily separated by the clustering algorithms, because they have a high similarity between its individuals, but other elements require more attributes that can add more information to discriminate them from others.

Keywords: feature selection, clustering, Land cover/use, seleção de atributos, agrupamento, uso e cobertura da Terra

1. Introdução

O sensoriamento remoto permite observar grandes extensões geográficas sistematicamente ao longo do tempo. Usando técnicas de classificação de imagens digitais é possível usar dados de sensoriamento remoto para criar mapeamentos temáticos, onde regiões, ou objetos, são classificados em classes de interesse de acordo com algum critério de similaridade entre os elementos fundamentais da imagem, ou *pixels*, que os formam.

A classificação de um conjunto de dados pode ser feita por métodos supervisionados onde amostras das classes buscadas são passadas para o método. Existem também os métodos não supervisionados, ou seja, onde nenhum conhecimento a priori é passado para o método. Os métodos de classificação não supervisionada baseados em *clustering* tem por objetivo dividir um conjunto de dados em grupos, ou *clusters*, de forma que objetos dentro de um *cluster* possuem alta similaridade entre si, mas são dissimilares dos objetos em outros *clusters*. A similaridade é dada por uma função de distância (HAN; KAMBER; PEI, 2011) considerando um ou mais atributos, como por exemplo proximidade espacial, ou características radiométricas de *pixels* de uma imagem. Nesse caso, os grupos, ou classes encontrados em geral representam classes de uso e ou cobertura da Terra (OHATA; QUINTANILHA, 2005).

Esse trabalho trata em particular de classes de uso do solo relacionadas com agricultura. Dados de sensoriamento remoto podem ser usados por exemplo para discriminar diferentes culturas agrícolas (GLERIANI; EPIPHANIO; SILVA, 2005) de uma região. A possibilidade de se acompanhar o mesmo ponto da superfície da Terra por ao longo do tempo por

sensoriamento remoto também permite por exemplo, caracterizar e acompanhar padrões sazonais característicos de uma dada cultura.

A fim de se obter uma boa classificação é comum que sejam usados vários atributos radiométricos disponíveis nos dados de sensoriamento remoto. Diversos sensores fornecem dados em faixas do espectro radiométrico que são de interesse para o acompanhamento de culturas agrícolas ou ainda produtos derivados de medidas básicas, como por exemplo os produtos do sensor MODIS, presente na plataforma TERRA (RUDORFF; SHIMABUKURO; CEBALLOS, 2007). No entanto, devido a alta correlação entre atributos radiométricos e até mesmo a dificuldade de se implementar classificadores baseados em múltiplos atributos (JAIN; DUIN; MAO, 2000) é necessário estudar quais são os atributos mais adequados para serem usados em uma dada classificação, ou seja, como reduzir a dimensionalidade dos atributos usados na classificação de dados por *clustering*. Essa é uma tarefa difícil, pois depende tanto do domínio de aplicação como do tipo de classificador utilizado (DUTRA, 1999).

Este trabalho apresenta o uso de algoritmos de *clustering* na separação de culturas agrícolas de tipos de uso e cobertura da Terra, considerando diferentes conjuntos de atributos radiométricos. Para isso, toma-se um conjunto de amostras pontuais obtidas em campo, com a anotação de um tipo de cultura observada em uma certa data. Para esses pontos foram extraídos valores de atributos radiométricos de alguns produtos de sensoriamento. Em seguida, alguns experimentos de classificação baseados em *clustering* são mostrados.

2. Materiais e Métodos

Para este trabalho foram utilizados um conjunto de amostras pontuais de talhões[1] em diferentes localidades do Brasil. Para cada amostra sabe-se o tipo de culturas agrícolas. A Tabela 1 mostra um extrato desse conjunto de dados:

Tabela 1: Conjunto de dados analisado

class	**class2**	**intensity_blue**	**maxval_blue**	**minval_blue**	...
Agua-1	Agua	0.0719	0.1479	0.076	...
Agua-2	Agua	0.1007	0.2167	0.116	...
algodao/pasto-1	algodao/pasto	0.0602	0.0927	0.0325	...
algodao/pasto-2	algodao/pasto	0.0447	0.0734	0.0287	...
algodao/pasto-3	algodao/pasto	0.0527	0.0851	0.0324	...
algodao/pasto-4	algodao/pasto	0.0555	0.0857	0.0302	...
algodao/pousio-1	algodao/pousio	0.047	0.0736	0.0266	...
algodao/pousio-2	algodao/pousio	0.0669	0.0811	0.0142	...

Além do conjunto de amostras de culturas agrícolas também foi utilizado um conjunto com amostras de uso e cobertura da Terra, por exemplo, água, área urbana, reflorestamento ou floresta. Com base nesses dois conjuntos de amostras foi gerado um arquivo de assinaturas com métricas (valor de intensidade, mínimo, máximo, média e desvio padrão) para diferentes bandas (banda *BLUE*, índices de *Normalized Difference Vegetation Index* (NDVI) e *Enhanced Vegetation Index 2* (EVI2)) visto na Tabela 1.

[1]Talhão tem como finalidade representar a divisão real ou imaginária de uma propriedade (fazenda), o que possibilita o controle mais apurado dos custos de produção, individualizados por safra (ciclo produtivo). Fonte: (TOTVS, 2014)

2.1. Seleção de atributos

Com o objetivo de reduzir a dimensionalidade do conjunto de atributos, foi utilizado o módulo de seleção de atributos do ambiente *Waikato Environment for Knowledge Analysis* (WEKA) que permite que métodos para seleção de atributos possam ser aplicados sobre uma determinada base de dados. Neste trabalho foi utilizado a técnica *CfsSubsetEval* (*Correlation based Feature Selection* - CFS) e o método de busca *BestFirst*, que busca pelo melhor conjunto de atributos.

O algoritmo de seleção de atributos *Correlation based Feature Selection* proposto em (HALL, 1998), considera que um bom subconjunto de atributos é aquele que contém atributos altamente correlacionados com a classe, porém com baixa correlação entre si. Com isso subconjuntos de atributos muito correlacionados com a classe e com baixa correlação entre si são escolhidos para a seleção.

Para selecionar os atributos mais correlacionado com a classe alvo, o algoritmo CFS foi aplicado sobre o conjunto de amostras com 17 atributos (Tabela 2), que referem-se aos valores de atributos radiométricos dos tipos de culturas agrícolas e de uso e cobertura da Terra, resultando em 7 atributos como sendo correlacionados com a classe alvo (*class2*): *intensity_blue*, *minval_blue*, *stddev_blue*, *maxval_evi2*, *stddev_evi2*, *intensity_ndvi* e *stddev_ndvi*.

Tabela 2: Atributos radiométricos considerados

Atributos	**Bandas**			**Número de atributos**
Valor de intensidade (*intensity*)	*BLUE*	NDVI	EVI2	15
Valor máximo (*maxval*)				
Valor mínimo (*minval*)				
Valor médio (*meanval*)				
Desvio padrão (*stddev*)				
Tipos de culturas agrícola (*class*)				2
Junção de tipos similares de culturas agrícola (*class2*)				

2.2. *Clustering*

As técnicas de *clustering* desempenham uma função importante na identificação de relações existentes entre dados de um domínio, que não são identificados de modo eficiente por uma pessoa, e possibilitam diferenciar classes de dados ao caracterizar o padrão considerando os conjuntos.

Os métodos de *clustering* podem ser divididos em cinco categorias (HAN; KAMBER; PEI, 2011): Métodos de particionamento, em que os dados são divididos em subconjuntos (*clusters*). Dados n objetos, esse método constrói k partições do dado, onde cada partição representa um *cluster*, com $k \leq n$; Métodos hierárquicos, criam uma decomposição hierárquica de um dado conjunto de objetos, em que os *clusters* são aninhados e organizados em uma árvore de dendograma; Métodos baseados em densidade, que constrói *clusters* com base na densidade das regiões, ou seja, no número de objetos ou pontos; Métodos baseados em grade, que visa encontrar *clusters* com base no número de pontos em cada célula, que forma uma estrutura de grade; e Métodos com base em modelo, para cada um dos *clusters*, tenta encontrar o melhor ajuste dos dados ao modelo dado.

Neste trabalho foram utilizados três algoritmos de *clustering*, o algoritmo *K-Means* (MACQUEEN, 1967) e o algoritmo *k-medoid Partitioning Around Medoids* (PAM) (KAUFMAN; ROUSSEEUW, 1990), ambos baseados em particionamento. E o algoritmo hierárquico *hierarchical clustering*, utilizando o método aglomerativo *complete linkage* (MURTÁGH, 1985), disponíveis em R (ZHAO, 2012).

Apesar dos algoritmos *K-Means* e PAM serem similares por dividirem o conjunto de dados em grupos tentando minimizar o erro quadrado. Eles possuem características particulares. O *K-Means* é sensível a *outliers* e ruídos, e seu desempenho depende da posição dos centroides iniciais. Por sua vez, o algoritmo PAM trabalha com o conceito de *medoid*, que são objetos que representam o grupo em que estão contidos, sendo tolerante a *outliers* e adequado para atributos categóricos.

2.3. Método Silhueta

Dentre as abordagens existentes para auxiliar na decisão do número de grupos, foi utilizado o método Silhueta, em inglês *silhouette*, proposto por (ROUSSEEUW, 1987), que subsidia na escolha de um número ótimo de grupos, avaliando os particionamentos encontrados, e permite visualizar graficamente os agrupamentos.

A silhueta é um gráfico do *cluster* C composto por um valor de silhueta $s(i)$, $i = 1, ..., n$, que reflete a qualidade da alocação dos objetos no grupos. Cada objeto (indivíduo) do *cluster* é representado por i. E para cada objeto i o valor $s(i)$ é calculado (Equação 1):

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{1}$$

Onde $a(i)$ é a dissimilaridade média do objeto i em relação a todos os objetos do mesmo grupo C, e $b(i)$ é a dissimilaridade média entre o objeto i em relação a todos os objetos do grupo vizinho mais próximo a ele, grupo X.

O valor de $s(i)$ varia no intervalo entre -1 e 1, sendo adimensional. Quando um valor de $s(i) \approx 1$, significa que o objeto i foi bem classificado no grupo C, pois $a(i) < b(i)$. Se o valor de $s(i) \approx -1$, significa que o objeto foi mal classificado, pois $a(i) > b(i)$, ou seja, o objeto i, em média, está mais distante dos objetos do seu próprio grupo, isto é, o objeto do grupo C está mais próximo dos objetos do grupo X. Por sua vez, se $s(i) \approx 0$, o objeto i está entre os grupos C e X, isso ocorre quando $a(i)$ = $b(i)$, indicando que o objeto está num ponto intermediário a dois grupos. Logo, quanto mais próximo a 1, melhor será a qualidade do agrupamento (SOUZA, 2007).

Uma interpretação subjetiva para este método foi proposta por (KAUFMAN; ROUSSEEUW, 1990), que subsidia na avaliação do agrupamento encontrado (Tabela 3). O coeficiente de silhueta ($CS(i)$) é uma medida de qualidade para toda estrutura de agrupamento que foi descoberta pelo algoritmo de classificação.

3. Resultados e Discussão

Nesta sessão apresentaremos alguns dos resultados obtidos com a aplicação dos três algoritmos de *clustering*, *K-Means*, PAM e *hierarchical clustering* sobre os valores de atributos.

3.1. Experimento 1: Algoritmo *K-Means*

Neste experimento foi aplicado o algoritmo *K-Means* sobre os dados, com valor de k=4, em que k representa o número de *clusters*. O método silhueta foi utilizado para subsidiar na escolha do número adequado de *clusters*. O valor ótimo encontrado foi de quatro grupos para separação do conjunto de valores de atributos relacionados ao uso e cobertura da Terra

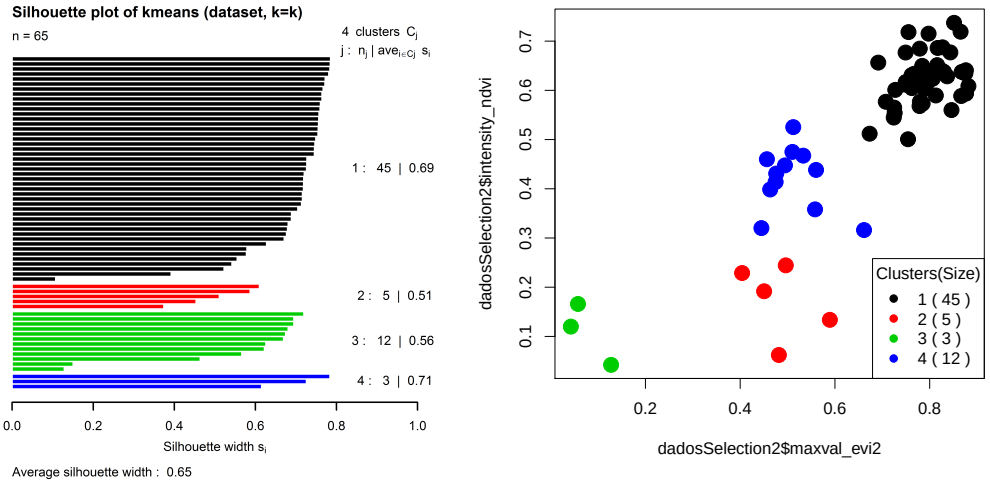
Tabela 3: Interpretação de agrupamentos segundo coeficiente de silhueta ($CS(i)$)

$CS(i)$	**Interpretação sugerida**
$0.71 - 1.00$	Grupos descobertos possuem uma estrutura muito robusta
$0.51 - 0.70$	Grupos possuem uma estrutura razoável
$0.26 - 0.50$	Os grupos encontrados possuem uma estrutura fraca e pode ser artificial. É aconselhável tentar outros métodos sobre o conjunto de dados
≤ 0.25	Nenhuma estrutura foi descoberta

Fonte: Adaptada de (KAUFMAN; ROUSSEEUW, 1990)

e culturas agrícolas. No gráfico de silhueta o eixo vertical representa os n objetos, enquanto o eixo horizontal representa o valor da silhueta para cada indivíduo. A Figura 1(a) mostra o gráfico de silhueta para um conjunto de valores de atributos que foi subdividido em quatro *clusters*. A avaliação do agrupamento foi realizada conforme Tabela 3, em que foi verificado que o agrupamento encontrado possui uma estrutura razoável, com média 0.65.

Na Figura 1(b) é apresentado a relação entre dois atributos *maxval_evi2* X *intensity_ndvi* com o posicionamento dos *clusters*, para visualização dos grupos.



(a) Gráfico de silhueta

(b) Relação entre os atributos *maxval_evi2* x *intensity_ndvi*

Figura 1: Visualização dos *clusters* criados pelo algoritmo *K-Means*

Analisando cada *cluster* (Figura 1(b)), podemos verificar que os agrupamentos possuem algumas características relevantes, tais como:

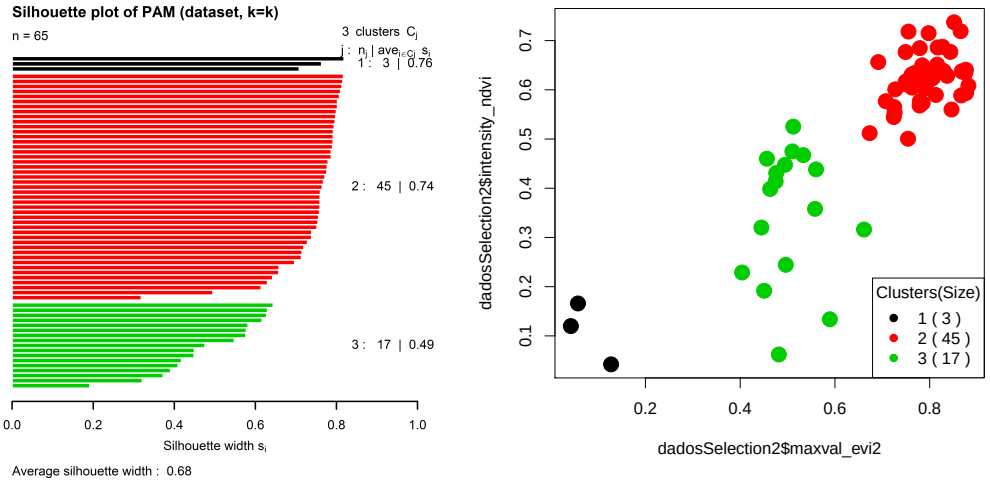
- Grupo 1 (45 indivíduos) é composto na maioria apenas por tipos de culturas agrícolas (soja, algodão ou feijão);
- Grupo 2 (5) formado apenas por tipos de uso e cobertura da Terra (reflorestamento, florestas ou mata);
- Grupo 3 é composto por 3 tipos de uso e cobertura da Terra (água e urbano);

- Grupo 4 (com 12 objetos), sendo heterogêneo em relação aos outros grupos, com dados sobre uso e cobertura da Terra e culturas agrícolas (mata, reflorestamento, algodão, pasto ou soja).

3.2. Experimento 2: Algoritmo *Partitioning Around Medoids* (PAM)

Para este experimento foi aplicado o algoritmo PAM sobre o conjunto de atributos com k=3. Por meio do método silhueta foi descoberto o número ideal de três *clusters* para separação dos dados de uso e cobertura da Terra e culturas agrícolas. A Figura 2(a) mostra o gráfico de silhueta do conjunto de dados que foi subdividido em três grupos. Na avaliação realizada sobre todo o agrupamento foi verificado que a estrutura encontrada é razoável, com média 0.68.

A Figura 2(b) apresenta uma visualização para a relação entre dois valores de atributos, *maxval_evi2* X *intensity_ndvi*, com a disposição dos *clusters*.



(a) Silhueta para os dados com PAM

(b) Relação entre os atributos *maxval_evi2* x *intensity_ndvi*

Figura 2: Visualização dos *clusters* criados pelo algoritmo PAM

Observando cada *cluster* (Figura 2(b)), é possível inferir algumas relações entre os agrupamentos:

- Grupo 1, com 3 indivíduos, é formado apenas por tipos de uso e cobertura da Terra (água e urbano);
- Grupo 2 é semelhante ao grupo 1 do algoritmo *K-Means*;
- Grupo 3 é composto por tipos de uso e cobertura da Terra e culturas agrícolas, sendo um grupo heterogêneo.

3.3. Experimento 3: Algoritmo *Hierarchical Clustering* (HC)

Nesta etapa foi utilizado o algoritmo *Hierarchical Clustering* com o método aglomerativo *complete linkage*, para verificar o comportamento dos dados utilizando uma abordagem hierárquica. Para isto, foi aplicado o método silhueta com o objetivo de verificar um número ótimo de *clusters*. Como resultado, foi retornado um valor de quatro grupos como sendo ideal. Semelhante resultado obtido pelo Experimento 1.

Com base na quantidade de *clusters* obtidos pelo método silhueta, aplicou-se o algoritmo HC sobre os dados, utilizando k=4. O resultado pode ser visto na Figura 3, que apresenta uma árvore de dendograma com os elementos de cada grupo separados por uma linha vermelha. A disposição dos elementos de cada grupo foi similar ao resultado obtido pelo algoritmo *K-Means*.

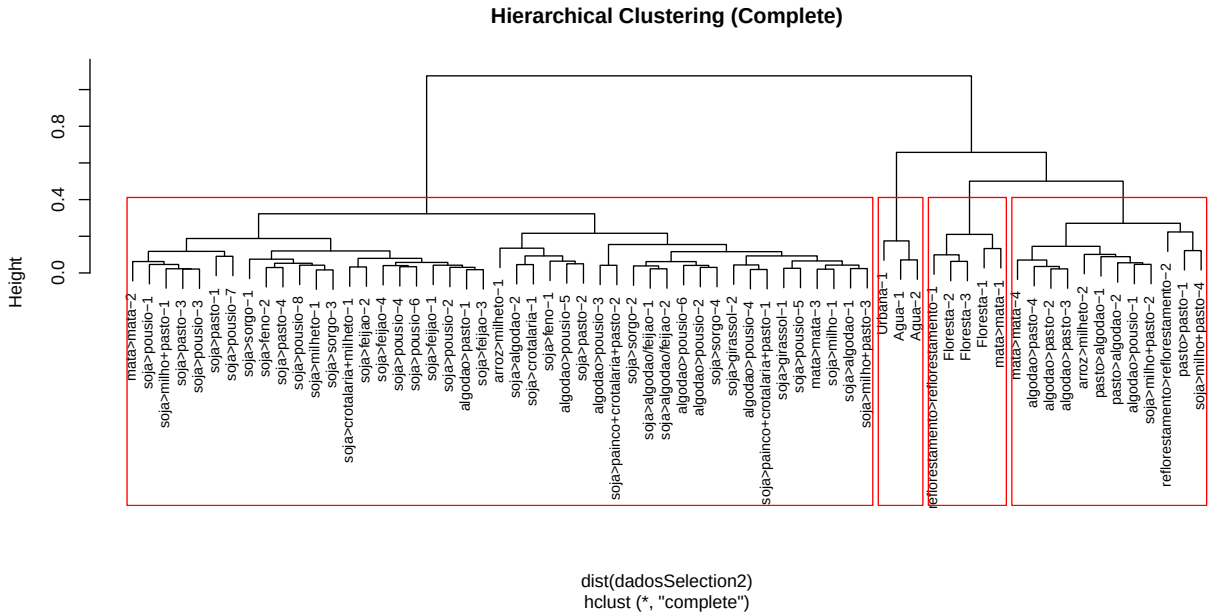


Figura 3: Árvore de Dendograma do algoritmo HC com separação em 4 grupos distintos

4. Conclusões

Neste trabalho foi realizado alguns experimentos utilizando duas categorias de *clustering*, método particionado e hierárquico, para verificar a separabilidade de culturas agrícolas e tipos de uso e cobertura da Terra, utilizando dados de sensoriamento remoto.

Foi verificado que apesar de alguns *clusters* possuírem homogeneidade entre seus elementos, apresentando apenas dados de uso e cobertura da Terra ou de tipos de culturas agrícolas, foi observado que em outros grupos ocorreu uma mistura entre diferentes classes. Fato que pode ter ocorrido por erros nos dados coletados, pela necessidade de mais amostras para compor as métricas ou por uma seleção de atributos mais abrangente, que possa agregar mais informação aos dados.

Nota-se que os algoritmos *K-Means* e *Hierarchical Clustering* apresentaram o mesmo resultado com *clusters* iguais. Entretanto, o algoritmo PAM, apresentou divergência quanto ao número de *clusters* ideais, sendo um resultado interessante e que deve ser levado em consideração em trabalhos futuros para verificar a separabilidade entre uso e cobertura da Terra e tipos de culturas agrícolas.

Agradecimentos

Os autores agradecem a CAPES pelo apoio financeiro.

Referências

DUTRA, L. V. Feature extraction and selection for ers-1/2 insar classification. *International Journal of Remote Sensing*, v. 20, n. 5, p. 993–1016, 1999. Disponível em: <http://dx.doi.org/10.1080/014311699213046>.

GLERIANI, J. M.; EPIPHANIO, J. C. N.; SILVA, J. D. S. d. Classificação espectro-temporal de culturas agrícolas tropicais: tolerância de dois modelos de redes neurais a dados falhos. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 12. (SBSR), 16-21 abr. 2005, Goiânia. *Anais...* São José dos Campos: INPE, 2005. p. 151–158. ISBN 85-17-00018-8. Disponível em: <http://urlib.net/ltid.inpe.br/sbsr/2004/11.20.12.00>. Acesso em: 02 set. 2014.

HALL, M. A. *Correlation-based Feature Subset Selection for Machine Learning*. Tese (Doutorado) — University of Waikato, Hamilton, New Zealand, 1998.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining: Concepts and Techniques*. 3rd. ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011. 744 p. ISBN 0123814790, 9780123814791.

JAIN, A.; DUIN, R.; MAO, J. Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 22, n. 1, p. 4–37, Jan 2000. ISSN 0162-8828.

KAUFMAN, L.; ROUSSEEUW, P. J. *Finding groups in data: an introduction to cluster analysis*. New York: John Wiley and Sons, 1990.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability - Vol. 1*. [S.l.]: University of California Press, Berkeley, CA, USA, 1967. p. 281–297.

MURTÁGH, F. *Multidimensional clustering algorithms*. [S.l.]: Physica-Verlag, 1985. (Compstat lectures, v. 4). ISBN 9783705100084.

OHATA, A. T.; QUINTANILHA, J. A. O uso de algoritmos de clustering na mensuração da expansão urbana e detecção de alterações na região metropolitana de são paulo. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 12. (SBSR), 16-21 abr. 2005, Goiânia. *Anais...* São José dos Campos: INPE, 2005. p. 647–656. ISBN 85-17-00018-8. Disponível em: <http://urlib.net/ltid.inpe.br/sbsr/2004/11.21.15.43>. Acesso em: 02 set. 2014.

ROUSSEEUW, P. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 20, n. 1, p. 53–65, nov. 1987. ISSN 0377-0427.

RUDORFF, B. F. T.; SHIMABUKURO, Y. E.; CEBALLOS, J. C. (Ed.). *O sensor MODIS e suas aplicações ambientais no Brasil*. 1. ed. São José dos Campos: Parêntese, 2007. 425 p. ISBN 978-85-60507-00-9.

SOUZA, E. F. *Comparação e escolha de agrupamentos: uma proposta utilizando a entropia*. Dissertação (Mestrado) — Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2007. Disponível em: <http://www.teses.usp.br/teses/disponiveis/45/45133/tde-13092007-145328/>.

TOTVS. *Talhões*. 2014. Disponível em: `http://www.totvs.com/mktfiles/tdiportais/helponlineprotheus/portuguese/agra010_talhoes.htm`. Acesso em: 03 setembro 2014.

ZHAO, Y. *R and Data Mining: Examples and Case Studies*. Elsevier Science, 2012. ISBN 9780123972712. Disponível em: <http://www.rdatamining.com/books/rdm>.