TRATAMENTO DE DADOS OPENSOURCE PARA CLASSIFICAÇÃO DE IMAGENS DE SATÉLITE

GIRÃO[1], Inês; VIANA[2], Cláudia M.; ROCHA[3], Jorge
[1] IGOT-Universidade de Lisboa, inesgirao@campus.ul.pt
[2] IGOT- Universidade de Lisboa, claudiaviana@campus.ul.pt
[3] IGOT- Universidade de Lisboa, jorge.rocha@campus.ul.pt

Resumo: A elaboração, com elevada precisão, de mapas de uso e ocupação do solo, através de imagens de satélite e métodos de classificação supervisionados depende, em grande medida, das amostras. Neste sentido, a disponibilidade de informação aberta e grátis oficial é particularmente relevante, uma vez que possibilita um conhecimento mais aprofundado sobre a paisagem do local em estudo e, também, a aplicação de classificações supervisionadas. Todavia, em ambientes biofísicos de elevada heterogeneidade, a ampla gama de assinaturas espectrais e a pequena variação destas entre as classes de uso e ocupação do solo poderá influenciar, de forma menos positiva, a produção destes mapas (Viana et al., 2019). Além desta questão existem, também, problemas relacionados com a informação de base utilizada para gerar as amostras de treino. No presente estudo, a carta de uso e ocupação do solo portuguesa (COS) de 2015 consistiu na informação de base para elaboração das amostras de treino. Desta forma, considerando que na COS as classes são representadas por polígonos que incluem elementos que, na verdade, não correspondem à classe propriamente dita (i.e. estradas de terra batida em torno de campos de cultivo), torna-se particularmente importante a aplicação de metodologias que permitam, à priori, analisar as amostras e, se necessário, proceder ao tratamento destas. Deste modo, é explorada a técnica de classificação por grupos (k-means) em ambiente R com recurso ao *tclust* package, no sentido de analisar e selecionar as amostras mais representativas de cada classe a classificar (Cuesta-Albertos et al.,1997; Fritz, García-Escudero & Mayo-Iscar, 2012). O presente estudo investiga o potencial desta técnica nas classificações supervisionadas de imagens de satélite numa região predominantemente rural caracterizada por uma mistura de ambientes agro-silvo-pastoris. Assim, realizou-se duas classificações para 2015: i) com as amostras originais ii) com as amostras selecionadas. Por fim, as experiências realizadas resultaram numa melhoria da precisão da classificação, (8%) demonstrando, assim, que a metodologia aplicada evidenciou um impacto positivo nos resultados.

Palavras-chave: Uso e Ocupação do Solo; amostras de treino; Clusters; Landsat; Random Forest

1. Introdução

Nos estudos à escala regional os produtos Landsat TM/ETM + são recorrentemente utilizados pois apresentam, frequentemente, valores de precisão elevados aquando à aplicação de técnicas de classificação baseadas no píxel (Lu & Weng, 2007). Contudo, o desempenho destas técnicas de classificação está altamente dependente da qualidade e quantidade de dados disponíveis utilizados para treinar o modelo de classificação (Lippitt et al., 2008; Brodley & Friedl, 1999). Para a grande maioria dos classificadores supervisionados, como maximum likelihood classification (MLC), multi-layer perceptron (MLP), support vector machine (SVM), or random forest (RF), não ter uma amostra

representativa será prejudicial para os valores de precisão atingidos na classificação realizada (Lu and Weng 2007).

As amostras para treino de um classificador são geralmente adquiridas através de conhecimento empírico e trabalho de campo, ou através de interpretação visual de, por exemplo, fotografias aéreas (Lu & Weng, 2007). No entanto, a aquisição deste tipo de informação acarreta elevados custos do ponto de vista monetário e de tempo, sobretudo em trabalho de campo, sendo que a aquisição de informação através da interpretação visual de outros produtos será sempre subjetiva e por vezes difícil, podendo causar alguns problemas com representatividade da classe (Usman 2013). Além disso, em ambientes biofísicos complexos a representatividade adequada de cada classe pode ser difícil de atingir pela confusão que ocorre entre algumas classes de uso e ocupação do solo (Lu and Weng 2007). Desta forma, torna-se essencial o estudo de alternativas que permitam criar amostras de elevada qualidade. Assim, o presente estudo explora a técnica *clustering k-means* como tratamento para selecionar as amostras que melhor representem cada classe. Este estudo pretende avaliar o potencial desta técnica para a classificação de uso e ocupação do solo numa região predominantemente rural com um ambiente agro-silvo-pastoral – ambiente de elevada heterogeneidade espectral.

2. Dados e área de estudo

A área de estudo selecionada diz respeito à Região do Baixo Alentejo, sendo esta caracterizada por uma paisagem de vastas culturas de trigo, sobreiros, azinheiras e oliveiras, onde o uso se releva uma mistura de agro-silvo-pastoral. A organização desta paisagem é bastante complexa devido aos diferentes calendários de cultivo e à geometria dos terrenos agrícolas. Nesta região o tecido urbano é dominantemente disperso.

Como referido, anteriormente, os dados utilizados provêm de imagens de satélite Landsat-8 e da carta de ocupação e uso do solo (COS-2015). A COS é produzida pela Direção-Geral do Território que se encontra disponível para *download* gratuito (http://mapas.dgterritorio.pt/geoportal/catalogo.html). Considerando o nível 1 e 2 deste produto, e com base no conhecimento empírico, selecionou-se as 7 classes com maior representatividade na área em estudo: 1) Superfícies não vegetadas; 2) Herbáceas temporárias; 3) Herbáceas permanentes; 4) Vinhas; 5) Olivais; 6) Florestas e superfícies seminaturais; 7) Corpos de Água. Para gerar o conjunto de amostras o mapa da COS (2015) foi convertido para um ficheiro matricial com uma resolução de 30 metros – a resolução espacial das imagens de satélite Landsat-8. Posteriormente criou-se um ponto para o centróide de cada píxel. Por fim, foram selecionados aleatoriamente 2000 pontos por classe, sendo utilizado metade das amostras para treino do classificador e a outra metade para validação da

classificação. Dado que a metodologia será aplicada no sentido de melhorar a identificação das diferentes classes de uso e ocupação do solo com assinaturas espectrais semelhantes, três índices de vegetação foram calculados: 1) NDVI; 2) NDBI; 3) NDWI.

3. Métodos

A metodologia proposta pode foi aplicada considerando as limitações da origem das amostras (COS2015): 1) representação espacial em polígono o que leva a inclusão de elementos que na verdade não correspondem à classe (e.g. estradas de terra batida em torno de campos de cultivo); 2) elevada probabilidade de existir confusão espectral entre classes pela homogeneidade espectral entre as classes selecionadas, apesar de a área em estudo ser bastante heterogénea do ponto de vista paisagístico.

Os métodos de análise clusters são, normalmente, empregues na tentativa de detetar grupos homogéneos. A lógica utilizada na implementação da análise de clusters para melhorar a qualidade das amostras é a de que cada classe de uso e ocupação do solo deverá ser representada por um (e.g. Floresta) ou no máximo dois clusters (e.g. Herbáceas temporárias). Para verificar esta hipótese foi utilizado o pacote *tclust* do software R (Fritz et al. 2012). Este pacote permite, para além da aplicação diversos métodos de análise de clusters, determinar os parâmetros mais adequados, no que diz respeito ao número ideal de clusters versus o número de pontos mais discordantes a ser eliminados. Neste estudo foi utilizado o método *trimming k-means* introduzido por Cuesta-Albertos et al. (1997). O processo de utilização deste método pode ser dividido em três fases distintas: 1) Aplicação da função *ctlcurves* do pacote *tclust* para calcular o número de clusters a criar versus o número de pontos mais discordantes a ser eliminados (Figura 1); 2) Aplicação do método *k-means* para criação dos clusters; 3) Aplicação da função *DiscrFact* para o cálculo do valor do fator discriminante (Figura 2). A aplicação da função *ctlcurves* a uma sequência de valor k (número de clusters) e α (fração das amostras mais discrepantes) permite visualizar as implicações de aumentar ou diminuir os valores k na fração de pontos mais discrepantes. Para cálculo dos clusters os valores integrados na função *tclust* dependerão dos valores obtidos e selecionados nas funções *ctlcurves* e *DisrFact*. Nos restantes parâmetros da função foram utilizados os valores padrão, com exceção do parâmetro *restr* que foi definido como "eigen" de forma a controlar, simultaneamente, o tamanho de cada cluster bem como a sua esfericidade. O parâmetro *equal.weigths* foi definido como "TRUE" no sentido de evitar a criação de um cluster muito bem

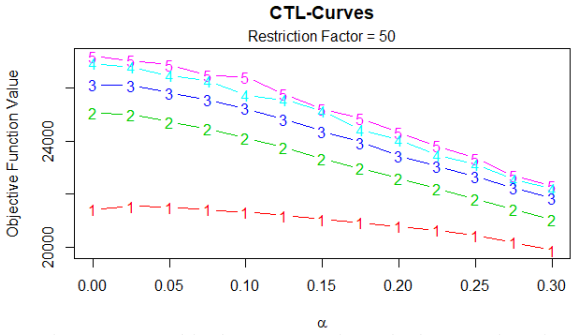

Figura 1: Graphical representation of ctlcurves function

definido mas que na verdade é composto por pontos que não representam a classe. Assim, evitamos a criação de grupos de elevada homogeneidade uma vez que as classes de uso e ocupação do solo não são caracterizadas por assinaturas espectrais únicas. Desta forma, alguma variabilidade espectral será assumida na criação dos clusters.

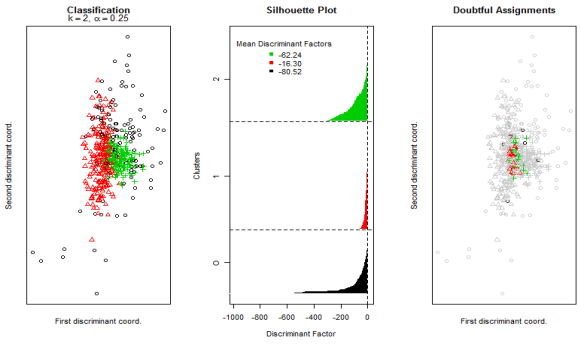

Figura 2: Representação gráfica da função DisrFact baseada nos valores dos fatores discriminantes (Df) – clusters com um elevado número de amostras com valores discriminantes elevados (e.g. próximos de zero) traduzem um cluster mal definido.

4. Resultados

O método utilizado para realizar a classificação foi o *RandomForest* dado este ser aceite pela comunidade cientifica aquando a classificação de imagens de satélite (Belgiu and Drăguţ 2016). A aplicação do método foi realizada em ambiente R utilizando o pacote *RandomForest*. Como referido, foram realizadas duas classificações sendo que a sua validação foi realizada com base no cálculo de uma matriz de confusão sendo que avaliamos a precisão das classificações com base nas métricas de precisão do usuário e produtor (Congalton 1991). A Figura 2 apresenta os resultados obtidos para ambas as classificações. Os resultados demonstram uma precisão global, para a classificação que utilizou as amostras originais, de 65% verificando-se alguma variação nas métricas de precisão relativas às classes (e.g. a classe corpos de água e florestas e superfícies seminaturais obtiveram elevada precisão sendo que a classes representativas de estrato herbáceo obtiveram valores relativamente baixos). Apesar dos valores do produtor, para estas mesmas classes, serem ligeiramente mais baixos estes continuam a sugerir a correta classificação das mesmas. Os resultados para a classificação que utilizou as amostras tratadas refletem uma elevada concordância (73,3%).

COS 2015 – AMOSTRAS ORIGINAIS

Classificação	Superfícies não vegetadas	Herbáceas temporárias	Herbáceas Permanentes	Vinhas	Olivais	Florestas e superfícies seminaturais	Corpos de água	Usuário
Superfícies não vegetadas	30	3	1	13	2	0	1	60
Herbáceas temporárias	5	28	11	4	0	1	1	56
Herbáceas Permanentes	4	3	29	0	3	9	2	58
Vinhas	12	2	0	31	4	0	1	62
Olivais	4	1	3	5	30	5	2	60
Florestas e superfícies seminaturais	0	1	3	2	1	37	6	74
Corpos de água	1	0	2	1	0	1	43	90
Produtor	54	74	59	55	75	70	77	65.5

COS 2015 – AMOSTRAS TRATADAS

Classificação	Superfícies não vegetadas	Herbáceas temporárias	Herbáceas Permanentes	Vinhas	Olivais	Florestas e superfícies seminaturais	Corpos de água	Usuário
Superfícies não vegetadas	**42**	**1**	**0**	**3**	**3**	**0**	**1**	**84**
Herbáceas temporárias	**0**	**37**	**7**	**3**	**1**	**2**	**0**	**74**
Herbáceas Permanentes	**1**	**7**	**29**	**1**	**3**	**9**	**0**	**58**
Vinhas	**6**	**2**	**1**	**33**	**4**	**4**	**0**	**66**
Olivais	**2**	**3**	**5**	**1**	**34**	**5**	**0**	**68**
Florestas e superfícies seminaturais	**3**	**0**	**3**	**1**	**1**	**42**	**0**	**84**
Corpos de água	**1**	**1**	**2**	**2**	**0**	**4**	**40**	**79**
Produtor	**76**	**73**	**62**	**75**	**74**	**64**	**97**	**73.3**

Figura 3: Matrizes de confusão para as duas classificações.

Note-se o aumento dos valores de precisão do usuário para a maioria das classes em particular, superfícies não vegetadas com taxas de precisão de usuário e produtor de 84% e 76%, respetivamente. As superfícies florestais e seminaturais também apresentaram altos valores de precisão do usuário (84%), enquanto as taxas de precisão do produtor foram menores, de 64%. A precisão do usuário para corpos de água diminuiu para 79%, enquanto a precisão do produtor aumentou para 97%.

5. Conclusão

Para criar um mapa de ocupação e uso do solo a partir de uma imagem de satélite, podemos seguir uma abordagem de classificação supervisionada se soubermos quais classes existentes na área de estudo e se tivermos amostras representativas para cada classe. No entanto, mesmo respeitando estas duas "regras", em ambientes biofísicos complexos, a ampla gama de assinaturas espectrais entre as classes de ocupação e uso do solo terá implicações nos resultados da classificação (Lu and Weng 2007). Assim, a interpretação e seleção da informação com base em conhecimento especializado e análises estatísticas revela-se pertinente (Brodley and Friedl 1999; Lippitt et al. 2008).

O estudo demonstra que o tratamento das amostras teve um impacto positivo na precisão atingida pelas classificações, sobretudo em situações de elevada complexidade espectral em cada classe. Assim, verificou-se uma melhoria da precisão global entre as duas classificações de 8%. A análise de clusters realizada em ambiente R revelou-se como eficiente e direta, sendo assim uma abordagem promissora.

6. Bibliografia

Belgiu M, Drăguţ L (2016) Random forest in remote sensing: A review of applications and future directions. ISPRS J Photogramm Remote Sens 114:24–31. doi: https://doi.org/10.1016/j.isprsjprs.2016.01.011

Brodley CE, Friedl MA (1999) Identifying Mislabeled Training Data. J Artif Intell Res 131–167. doi: 10.1136/ard.2003.010348

Congalton RG (1991) A review of assessing the accuracy of classifications of remotely sensed data. Remote Sens Environ 37:35–46. doi: 10.1016/0034-4257(91)90048-B

Cuesta-Albertos JA, Gordaliza A, Matrán C (1997) Trimmed k-means: An attempt to robustify quantizers. Ann Stat 25:553–576. doi: 10.1214/aos/1031833664

Fritz H, García-Escudero LA, Mayo-Iscar A (2012) **tclust** : An *R* Package for a Trimming Approach to Cluster Analysis. J Stat Softw. doi: 10.18637/jss.v047.i12

Lippitt CD, Rogan J, Li Z, et al (2008) Mapping Selective Logging in Mixed Deciduous Forest: A Comparison of Machine Learning Algorithms. Photogramm Eng Remote Sens 74:1201–1211.

Lu D, Weng Q (2007) A survey of image classification methods and techniques for improving classification performance. Int J Remote Sens 28:823–870. doi: 10.1080/01431160600746456

Usman B (2013) Satellite Imagery Land Cover Classification using K-Means Clustering Algorithm Computer Vision for Environmental Information Extraction. Sci Engg 63:18671–18675. doi: 10.1145/1837853.1693485

Viana CM, Girão I, Rocha J (2019) Long-Term Satellite Image Time-Series for Land Use/Land Cover Change Detection Using Refined Open Source Data in a Rural Region. Remote Sens 11:1104. doi: 10.3390/rs11091104