Università degli studi di
Milano-Bicocca

Machine Learning

Project

Predizione della copertura vegetale a partire da dati cartografici, ecologici

Authors:
Louis Fabrice Tshimanga
September 17, 2019

Abstract

Lo studio presentato mira a predire la classe di copertura vegetale nella Roosevelt National Forest in Colorado, a partire da dati cartografici (geografici, geologici ed ecologici) riferiti a celle di terreno di *30m*x*30m*. Sette tipologie di vegetazione sono state prese in considerazione, con diversa distribuzione e frequenza di rappresentazione nel dataset. Sono stati sviluppati classificatori per il problema multiclasse, senza semplificazioni binarie. Gli obiettivi ulteriori concernono lo sviluppo attributi trasformati ed utili nella riduzione della dimensionalità, nonché una valutazione della generalità dei modelli presentati.

Contents

1 Introduzione **1**

2 Dataset **2**

3 Data Preparation **3**
3.1 Preprocessing 3
3.2 Feature Engineering 4

4 Selezione dei Modelli, Valutazione **4**
4.1 Multilayer Perceptron, Intervallo di Confidenza 5
4.2 Random Forest vs Gradient Boosting Trees 6
4.2.1 Feature Selection 6
4.2.2 Cross-Validation, Comparazione 7
4.3 Generalizzazione 7

5 Conclusioni **8**

1 Introduzione

La copertura vegetale delle aree geografiche è una questione di grande e variegato interesse. La vegetazione che ricopre il suolo è un elemento della biosfera di massimo rilievo per estensione, massa e importanza ecologica. Vi si intende l'insieme delle comunità vegetali e delle piante viventi o fossili che

si ritrovano originariamente in un'area, non piantate o coltivate dall'uomo, ma eventualmente influenzate dall'attività antropica. La vegetazione caratterizza la biogeochimica dell'area di riferimento (cicli dell'acqua, dell'azoto, del carbonio), interagisce con il tipo di suolo su cui si sviluppa, costituisce l'habitat di altri attori della biosfera, nonché una risorsa economica ed energetica. Conoscere la copertura vegetale di un'area è utile per monitorare e conservare la stessa, come avviene nelle riserve naturali istituite dall'uomo. Per tale ragione lo US Forest Service (USFS) mantiene dei Region Resource Information Systems delle aree protette. I dati del servizio forestale e dello US Geological Survey (USGS) sono dunque un ottimo esempio di policy nonché un benchmark ormai acclarato per le analisi via machine learning.

Figure 1: Una vista del Roosevelt National Park (immagine in licenza commons [1])

2 Dataset

Il dataset in oggetto è una aggregazione dei dati botanici e cartografici di 4 aree distinte (Rawah, Neota, Comanche Peak e Cache La Poudre) e pressoché incontaminate della Roosevelt National Forest in Colorado, in cui il genere di vegetazione è perlopiù frutto di processi ecologici, piuttosto che di interventi umani. I record fanno riferimento a celle di terreno di *30m*x*30m*, caratterizzate da 54 attributi cartografici di tipo quantitativo e qualitativo. La maggior parte degli attributi (40) fa riferimento alla tassonomia del suolo, così che

ogni cella presenta il valore booleano di appartenenza ad un solo tipo di suolo, con considerevole dimensionalità e sparsità. Purtroppo non è presente informazione sufficiente per applicare esperienze e conoscenze di dominio, in particolare aggregando i tipi di suolo all'ordine tassonomico superiore, riducendo gli attributi e potendo indagare quale sia il livello più informativo della tipologia dei terreni. Le altre variabili si riferiscono all'elevazione della cella, all'aspetto azimutale (la direzione cardinale in cui la pendenza è massima), alla pendenza media, alla distanza orizzontale da strade, focolai e fonti d'acqua più vicine, nonché la distanza verticale delle stesse fonti d'acqua, e infine 3 indici di ombra registrati alle ore 9:00, 12:00 e 15:00 del solstizio d'estate. La variabile target è il tipo di coltre vegetale della cella, di cui esistono 7 classi disuniformi nel distribuirsi tanto nel dataset complessivo, quanto in ciascuna delle 4 aree di riferimento (a loro volta registrate in 4 variabili booleane di appartenenza).

3 Data Preparation

3.1 Preprocessing

Il dataset, di *581 012* record per *54* attributi, è notoriamente e verificatamente privo di missing values, pur contenendo dei valori nulli sospetti per l'indice di ombra nella variabile Hillshade3pm, che comunque sono coerenti con il range della variabile, altresì correlata ai restanti indici di ombra. La preparazione iniziale dei dati è dunque consistita nel trasformare effettivamente la codifica delle variabili binarie in un tipo booleano per l'ambiente KNIME, oltre a rinominare le colonne (salvo se relative al suolo) in modo leggibile e descrittivo, convertendo anche gli interi della variabile target in stringhe mutuate dalla classificazione botanica. Le 7 classi sono dunque:

1	Spruce/Fir
2	LodgepolePine
3	PonderosaPine
4	Cottonwood/Willow
5	Aspen
6	Douglas-fir
7	Krummholz

3.2 Feature Engineering

La natura concreta di alcuni dati (come il distinguo tra distanza verticale e orizzontale dalle risorse idriche) e la natura matematica e statistica di altri (come la misura dell'elevazione, insieme alla sua intensità peculiare per i record dell'area 4, Cache La Poudre) suggeriscono l'opportunità del Feature Engineering, ossia l'elaborazione di nuovi attributi a partire da quelli presenti. L'operazione complessiva è stata applicata con un metanodo invero molto semplice, composto da un nodo *Math Formula* per calcolare e allegare il logaritmo della variabile *Elevation*, e da un nodo *R Snippet* con cui sono stati calcolati e aggiunti gli attributi per la distanza euclidea dalle risorse idriche e la pendenza di tale distanza. Quanto alle variabili *Hillshade*, si è calcolato l'indice di ombra medio e l'interazione (come prodotto fratto la somma) fra gli indici alle 9:00 e alle 12:00, e fra le 9:00 e le 15:00. Da ultimo, sono stati elaborati due range medi, il primo come media delle distanze orizzontali da risorse idriche, principi di incendio e strade, ad indicare un areale approssimativo di vegetazione continua; il secondo come media fra le distanze orizzontali dei principi d'incendio e delle strade, per un areale approssimativo di vegetazione non disturbata o danneggiata.

4 Selezione dei Modelli, Valutazione

Il dataset è stato diviso, anche per ridurre i costi computazionali con la partizione *23* [dal nodo], in un *66%* denominato *23A* (*383 467* elementi) ed un *33%* denominato *23B*, mantenendo la stratificazione della classe target. Successivamente, per bilanciare le classi, *23A* è stato ridotto tramite equal size sampling a *12 691* individui, in modo da migliorare le prestazioni di generalizzazione dei classificatori (possibilmente a discapito della performance di modelli probabilistici sul dataset completo). Si noti quindi che una frazione dei record non è stata utilizzata né in training né in testing o validation.
La selezione dei modelli è stata effettuata inizialmente tra quattro collezioni di modelli probabilistici, (meta)euristici e partizionatori dello spazio di attributi. In prima istanza, per tutti i modelli è stato utilizzato il più ampio sottoinsieme di attributi compatibili, per una prima indagine sulle potenzialità relative all'accuracy. Secondariamente sono stati scelti i tre modelli più promettenti, con i quali indirizzare una selezione ridotta degli attributi ed una valutazione prestazionale di maggior robustezza statistica.

Nella prima fase, 4 metanodi hanno raccoltole famigliei di modelli allenati e testati sul campione già bilanciato, e partizionato in *26A* (training set, *8376* record) e *26B* (test set, *4315*). Il primo metanodo è stato rivolto al *MultiClassClassifier* di WEKA, seguito dal metanodo per *NaiveBayes* e *BayesNet*. Questi tre classificatori si sono attestati con accuratezza registrata inferiore al 71%, ossia in linea con risultati grezzi o da tempo superati in letteratura [2], nonché di gran lunga inferiori a quelli ottenuti dai decision tree esplorativi utilizzati nelle fasi di data preparation. Nel metanodo contenente le foreste e gli alberi decisionali, i migliori risultati in accuracy sono emersi dal *Random Forest* nativo e dal *Gradient Boosted Trees*, con rispettive accuracy di test pari a 83.29% e 84.13%, superando *Random Forest*, *NBTree* e *J48* di WEKA, *Decision Tree* nativo. Infine, un metanodo ha raccolto 3 percettroni, ispirandosi e modificando l'approccio in [2], in particolare con *MLP* a 120 nodi nell'unico hidden layer, mentre gli altri parametri [neuroni]:[layer] implementati sono stati [40]:[2] e [30]:[3]. Anche in questo raggruppamento il percettrone [120]:[1] ha dato la miglior performance sui dati, con un'accuratezza al test set pari a 81.28%, laddove lo score "peggiore" si è verificato per [30]:[3] con il 79.51%. Il confronto, come nei casi precedenti, non è dirimente su quale sia l'implementazione migliore in ciascun modello, ma test statistici sono stati rimandati a dopo la selezione dei 3 classificatori più accurati su questo primo training+test process. Dunque *MLP [120]*, *Gradient Boosted Trees* e *Random Forest* sono stati approfonditi.

4.1 Multilayer Perceptron, Intervallo di Confidenza

Per il MLP è stato calcolato l'intervallo di confidenza al 95% per l'accuracy, così da valutare l'affidabilità della performance con cui è stato selezionato. Ripartendo *26A* si possono misurare due valori-limite dell'intervallo, in questo caso *[0.78, 0.81]*, laddove il MLP otteneva un valore al limite significativamente più alto per il training sull'intero *26A*, e l'estremo opposto testato su *26B* con la riduzione del training set (vd. Figura 2). La differenza di training set spiega i valori differenti, mentre la "significatività" di tale differenza è in realtà decisa dalla confidenza scelta (aumentarne le percentuali significa aumentare l'ampienza dell'intervallo).

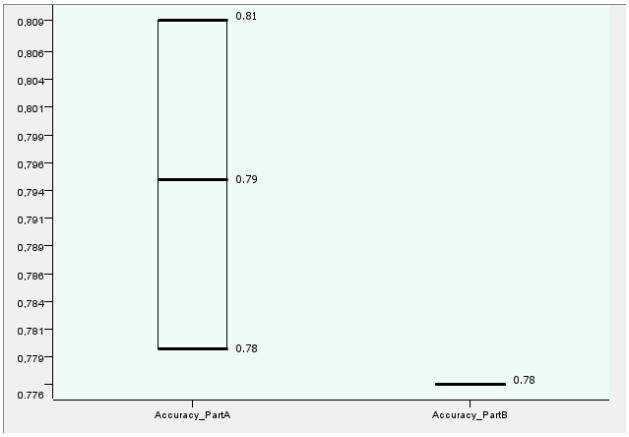

Figure 2: Intervallo di confidenza al 95% e risultato effettivo

4.2 Random Forest vs Gradient Boosting Trees

Gli alberi decisionali si sono dimostrati particolarmente efficaci ed efficienti nelle diverse fasi esplorative, e le tecniche collettive di Random Forest e della variante Gradient Boosting [2] sono risultate le migliori in termini di accuracy. Data la facilità di interpretazione degli alberi decisionali meno profondi, la soluzione è stata considerata particolarmente adatta anche a valutare l'importanza delle features, con attenzione a quelle sviluppate ad hoc nel workflow.

4.2.1 Feature Selection

La feature selection è operata attraverso loop che richiedono costi computazionali e temporali troppo ingenti se usati con reti neurali molto ampie, ma sono stati gestibili per i metodi ad albero. Si è deciso di valutare l'importanza tramite selection "forward", ossia recuperando un massimo di 20 attributi che accrescessero progressivamente l'accuracy dei due learner specifici, separatamente. Dopo l'ispezione delle 20 variabili dal maggior contributo all'accuratezza, si è scelto il miglior sottoinsieme di 10 variabili, ottenendo per il *Random Forest* un'accuratezza dell'*83.4%* e per il *Gradient Boosting Trees* l'*81.3%*. Per entrambi i classificatori le variabili prodotte *ln.Elevation, EucliDistHydro, MeanDisturbanceRange* sono entrate tra le 10 più significative per la bontà della performance, e nel caso dell'elevazione persino a discapito della forma originale.

4.2.2 Cross-Validation, Comparazione

Con l'utilizzo dei due filtri ottenuti alle colonne, i due modelli sono stati resettati al fine di ottenere l'accuracy per 10-folds cross-validation, più robusta e adatta a comparare la significatività della differenza tra le frequenze di misclassificazione dei due modelli. L'errore del *Gradient Boosting Trees* è risultato significativamente (al *95%*, per quanto leggermente, maggiore di quello da *Random Forest*. Per tale ragione, ad evitare la ridondanza tra le

Figure 3: Intervallo di confidenza al 95% della differenza fra errori. Sempre negativa, perché sempre maggiore l'errore GB

tecniche, solo il *Random Forest* è stato utilizzato nell'ultimo test di classificazione, confrontato al percettrone.

4.3 Generalizzazione

Nonostante i test statistici ed i procedimenti di validazione ottenuti sui dati a disposizione del Machine Learning Engineer, c'è sempre una componente imprevista della performance di un algoritmo o di un classificatore su dati completamente nuovi, specialmente se riferiti ad insiemi di campionamento diversi (come potrebbero esserlo in questo caso per geografia o fase temporale). In questo studio è stato ancora una volta simulato il contesto di ignoranza e novità dei dati, recuperando la partizione *23B*, relativamente minoritaria ma composta da *197 545* record. Su di essa sono stati definitivamente testati il Multilayer Perceptron [120]:[1] e gli alberi da *Random Forest*,

così come sviluppati e validati in precedenza, su *5528* e *8376* record rispettivamente. Data la differente dimensione del training set, il confronto non è da farsi tra i due modelli, bensì rispetto alle relative precedenti performances. La figura 4 mostra, specialmente rispetto al *Random Forest*, una buona generalizzazione ottenuta con dati di circa 3 ordini di grandezza meno numerosi rispetto al test "definitivo". Considerata anche la forte disuniformità nelle frequenze delle classi, e dunque il peso modificato degli indici di buona o cattiva classificazione di ciascuna tipologia, si può ritenere la generalità dei modelli piuttosto elevata.

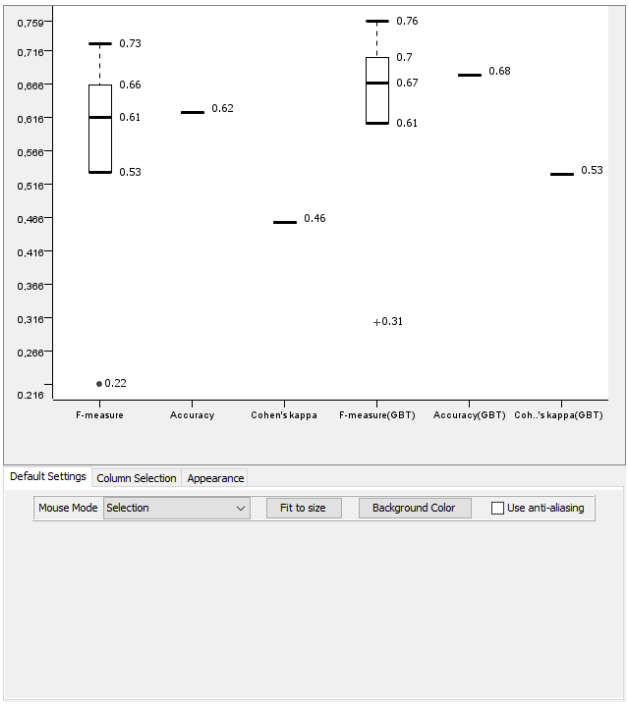

Figure 4: Intervallo di confidenza al 95% della differenza fra errori. Sempre negativa, perché sempre maggiore l'errore GB

5 Conclusioni

La ricchezza in termini di attributi e record permette di testare un'ampia varietà di classificatori per il problema della copertura vegetale. Fra questi, i migliori per interpretabilità ed efficacia risultano essere i metodi basati su

alberi decisionali, in particolare modificati con Gradient Boosting e Random Forest. I risultati ottenuti con bilanciamento delle classi per un training set contenente ¡1% dei record, e 16% degli attributi totali (6/53 originali degli originali) dimostrano una capacità di generalizzazione promettente per la classe di problemi affrontata, dando contemporaneamente giustificazione alla trasformazione e creazione di attributi eseguita nel flusso di lavoro. Per migliorare ulteriormente le prestazioni del classificatore e testarne definitivamente la generalità (anche al fine di affiancare e catalizzare le conoscenze di dominio ecologico) si attende una raccolta analoga per tipo di informazioni, su record di terreni ed ecosistemi possibilmente comparabili.

Referimenti e Note

[1] commons.wikimedia.com

[2] Jock A. Blackard , Denis J. Dean, *Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables*, 1999

[3] Leo Breiman, *Arcing the Edge*, 1997