Classificazione tipologia di foresta in base a dati satellitari
Report per l'Esame di Fondamenti di Machine Learning

Francesco Baboni

128860
Corso di Laurea in Ingegneria Informatica (sede di Mantova)
254252@studenti.unimore.it

26/07/2023

Abstract

In questo report verrà analizzato un algoritmo di machine learning il cui scopo è quello di prevedere la tipologia di una foresta in base a dati spettrali forniti da un satellite. Verranno successivamente effettuate valutazioni sulle caratteristiche dei modelli implementati e sulle metriche di valutazione ottenute in fase di testing del modello finale.

1 Struttura e Composizione Dataset

Il dataset contiene informazioni spettrali di immagini satellitari ottenute attraverso il sensore ASTER (Advanced Spaceborne Thermal Emission and Reflection Radiometer) raccolte in tre diverse date (riferite a stagioni dell'anno differenti, in cui è presente un fogliame diverso per ognuno dei tre casi):

- 26 / 09 / 2010 INIZIO AUTUNNO
- 19 / 03 / 2011 FINE INVERNO
- 08 / 05 / 2011 PRIMAVERA INOLTRATA

riferito a quattro diverse classi di foreste:

- **'s'** ('Sugi' forest) foreste di cedro del Giappone.
- **'h'** ('Hinoki' forest) foreste di cipresso giapponese.
- **'d'** ('Mixed deciduous' forest) foreste miste decidue.
- **'o'** ('Other' non-forest land) riferito ad altre tipologie di terreno non forestale.

Una volta definite le 4 classi di foreste (che saranno poi la variabile target) sono incluse le seguenti features:

- **b1** $--> $ **b9**: Features che rappresentano le bande spettrali dell'immagine ASTER che contengono informazioni spettrali nei canali verde, rosso e infrarosso vicino, per ciascuna delle tre date.
- **PredMinusObsSb1** $--> $ **PredMinusObsSb9**: Rappresentanti i valori spettrali previsti (basati sull'interpolazione spaziale) sottratti dai valori spettrali effettivi per la classe 's' (b1 $-->$ b9).
- **PredMinusObsHb1** $--> $ **PredMinusObsHb9**: Riferite ai valori spettrali previsti (basati sull'interpolazione spaziale) sottratti dai valori spettrali effettivi per la classe 'h' (b1 $-->$ b9).

2 Exploratory Data Analysis

Durante l'Esplorazione dei Dati (EDA) sul dataset sono state svolte diverse operazioni al fine di creare un set di dati pronto per il modello di classificazione. In primo luogo, la variabile target "class" è stata trasformata da categorica a discreta, al fine di convertire le etichette testuali delle classi in valori numerici, passaggio cruciale poiché i modelli che verranno successivamente utilizzati richiedono dati numerici per un funzionamento ottimale. Si segnala inoltre l'assenza di campi delle feature non definiti, non è risultato quindi necessario aggiungerli inserendo valori di moda o media. Successivamente, sono state estratte le variabili indipendenti (features) e la variabile dipendente (target) dal DataFrame, che rappresentano rispettivamente le colonne del dataset escludendo le classi di foreste, e i valori numerici associati alle classi di tipologie di foresta. E' stata inoltre effettuata un'analisi delle distribuzioni delle classi al fine di valutare la presenza di uno sbilanciamento tra di esse. Dopo aver implementato alcune funzioni per il conteggio delle istanze per ciascuna classe nel dataset, i risultati ottenuti hanno fornito una panoramica della distribuzione delle classi di foreste e della loro rappresentanza nel dataset di training.

Dai conteggi, si può osservare che il numero di istanze per ciascuna classe è il seguente:

- 'Mixed decidual' (d) 54 istanze
- 'Hinoki' (h) 48 istanze
- 'Sugi' (s) 59 istanze
- 'Other' (o) 37 istanze

Questi valori indicano che le classi presentano una frequenza relativamente equilibrata, con un numero di istanze simile tra loro, è quindi possibile considerare il dataset come relativamente bilanciato.

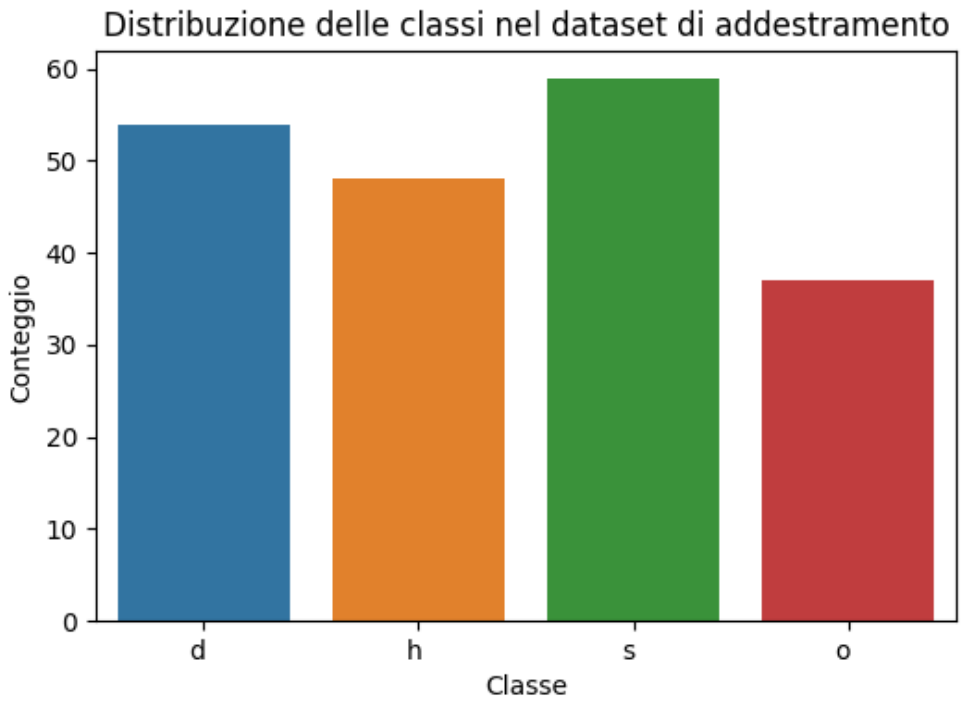

Figure 1: Distribuzione Classi

3 Scelta Modelli di Classificazione e Iperparametri

- K-Nearest Neighbors (K-NN): Il modello K-NN è stato scelto poiché è un classificatore basato sull'assunzione che istanze simili siano vicine. Per questo dataset, in cui la classificazione delle tipologie di foresta risulta essere influenzata da fattori spaziali, K-NN potrebbe essere una buona scelta. Le possibilità di iperparametri sono state definite con Nneighbors che rappresenta il numero di vicini da considerare. È stata esaminata una lista di valori dispari da 1 a 9, in quanto un valore dispari evita indecisioni nei casi di parità durante la classificazione.

- Softmax Regression: La regressione softmax è stata scelta perché è un modello adatto per la classificazione multiclasse, essendo una generalizzazione della regressione logistica binaria. Le possibilità di iperparametri includono penalty che rappresenta la norma della regolarizzazione (l1 o l2) e C che è il parametro di regolarizzazione. È stata esaminata una lista di valori (0.00001, 0.00004, 0.0001, 0.0004, 1) per C in modo da valutare diverse forze di regolarizzazione.

- Support Vector Machine (SVM): Le macchine a vettori di supporto sono state scelte perché sono un potente algoritmo di classificazione capace di gestire anche dati non linearmente separabili attraverso l'utilizzo di kernel. Le possibilità di iperparametri includono C che controlla il termine di regolarizzazione, gamma che controlla l'influenza dei singoli esempi di addestramento e kernel che specifica il tipo di kernel da utilizzare (lineare o radiale).

- Decision Tree: I decision tree sono stati scelti perché sono semplici da interpretare e possono catturare relazioni non lineari tra le feature. Le possibilità di iperparametri includono criterion, che specifica la funzione per misurare la qualità della divisione (gini o entropia).

Il conseguente utilizzo della Grid Search Cross-Validation a 5 fold ha permesso di esaminare le possibili combinazioni di iperparametri per ogni modello e trovare quelle che massimizzano l'accuratezza durante la fase di addestramento. In questo modo, è stato possibile ottenere modelli ottimizzati per ciascun algoritmo, garantendo prestazioni migliori rispetto a configurazioni di iperparametri arbitrariamente selezionate. La successiva combinazione (paragrafo 4) di modelli eterogenei ha consentito di sfruttare le peculiarità di ciascun modello per migliorare le prestazioni complessive del sistema di classificazione.

- Per il modello K-NN (K-Nearest Neighbors), la migliore configurazione di iperparametri è stata trovata con un valore di Nneighbors pari a 1. Ciò significa che il modello considera solo il campione più vicino quando effettua una previsione, rendendolo più sensibile ai singoli punti nel dataset di addestramento. L'accuratezza ottenuta con questa configurazione è del 96.49

- Per il modello Softmax Regression la migliore configurazione di iperparametri è risultata con un valore di C pari a 1 e l'utilizzo della penalizzazione L1. L'accuratezza ottenuta con questa configurazione è del 95.97

- Per il modello SVM (Support Vector Machine), la migliore configurazione di iperparametri è stata trovata con un valore di C pari a 100.0, gamma pari a 0.001 e l'utilizzo del kernel 'rbf' (Radial Basis Function). L'accuratezza ottenuta con questa configurazione è del 97.49

- Per il modello Decision Tree, la migliore configurazione di iperparametri è stata trovata con il criterio 'entropy'. L'accuratezza ottenuta con questa configurazione è del 96.97

4 Ensemble

E' stato infine effettuato un ensemble utilizzando uno Stacking Classifier con una Regressione Logistica come modello finale. Questa scelta è motivata dal fatto che il task consiste in un problema di classificazione multiclasse in cui l'obiettivo è assegnare una delle quattro tipologie

di foresta a ciascuna istanza. La Regressione Logistica è quindi un modello ideale per la classificazione multiclasse e può gestire più classi in modo efficiente, è inoltre semplice e diminuisce i tempi computazionali. Questo approccio di ensemble permette di ottenere un modello complessivo robusto e flessibile. L'utilizzo di uno Stacking Classifier ha permesso di combinare i punti di forza dei diversi modelli di base per ottenere prestazioni migliori rispetto all'utilizzo di un singolo modello.

4.1 Risultati Cross Validation Ensemble

I risultati ottenuti dalla cross validation dello Stacking Ensemble mostrano che il modello ha ottenuto buone prestazioni nella classificazione di tipi di foresta. Le metriche di valutazione utilizzate sono F1-score e Accuracy, entrambi valutati tramite una cross-validation a 5 fold. Il valore del F1-score ottenuto dallo Stacking Ensemble è pari a 0.969, mentre l'Accuracy è pari a 0.970. Questi valori indicano che il modello ha una buona capacità di generalizzazione, ha inoltre ottenuto risultati molto simili su entrambe le metriche di valutazione.

5 Valutazioni Metriche Finali e Conclusioni

	KNN	SOFTMAX REGRESSION	SVM	DECISION TREE	STACKING ENSEMBLE
ACCURACY	81,85%	83,38%	83,08%	79,08%	84,62%
PRECISION	82,38%	83,95%	83,55%	79,35%	84,84%
RECALL	81,85%	83,38%	83,08%	79,08%	84,62%
F1-SCORE	81,88%	83,41%	83,19%	79,08%	84,64%

Figure 2: Tabella Performance Finali

Nella tabella riportata sono presenti i risultati delle metriche di valutazione in fase di testing per ciascun modello testato, incluso lo stacking ensemble.

- L'accuracy varia tra il 79.08 (Decision Tree) e 84.62 (Stacking Ensemble). Questo indica che l'ensemble ha ottenuto le prestazioni migliori in termini di accuratezza nella classificazione delle istanze del dataset. Si sottolinea inoltre un valore elevato nei risultati dei modelli Softmax Regression e SVM.
- La precision varia dal 79.35 (Decision Tree) a 84.84 (Stacking Ensemble). Anche in questo caso, l'ensemble ha mostrato le prestazioni migliori, evidenziando la sua capacità di individuare correttamente le istanze positive.
- La recall varia dal 79.08 (Decision Tree) a 84.62 (Stacking Ensemble).
- L'F1-Score varia da 79.08 (Decision Tree) a 84.64 (Stacking Ensemble).

Complessivamente, i risultati indicano che l'Ensemble Stacking ha ottenuto le prestazioni migliori rispetto ai singoli modelli, fornendo quindi una soluzione più efficace.

Al fine di visualizzare al meglio i risultati dell'ensemble si è infine generata una confusion matrix che consente di osservare e di verificare in pochi istanti le corrette previsioni del modello e il numero complessivo di classi riconosciute correttamente.

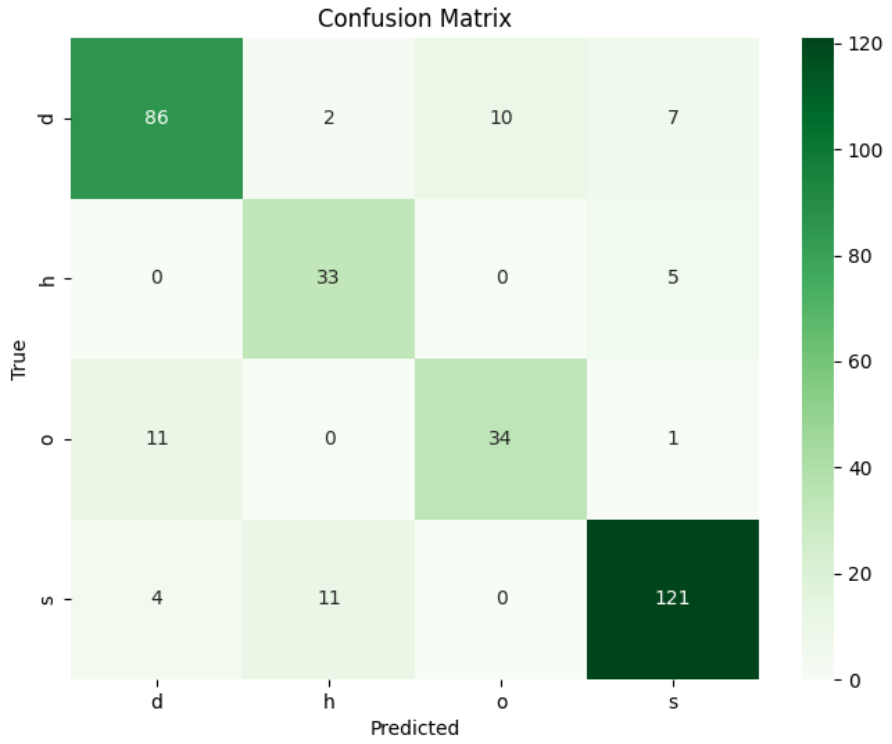

Figure 3: Confusion Matrix

Un'applicazione interessante per questo algoritmo potrebbe essere la previsione del tasso di assorbimento di CO2 da parte delle foreste, un aspetto fondamentale per comprendere il ruolo delle foreste nella sfida contro l'effetto serra. I risultati della classificazione potrebbero essere successivamente combinati con dati di monitoraggio delle concentrazioni di CO2 nell'atmosfera per stimare l'assorbimento di CO2 da parte delle diverse coperture forestali. Questa informazione sarebbe estremamente preziosa per monitorare e valutare l'impatto delle foreste sull'assorbimento di CO2 e contribuire agli sforzi di mitigazione dei cambiamenti climatici.