

Mappatura di categorie forestali mediante l'applicazione del metodo k-NN

Caterina Gagliano^{1*}, Francesca Incerti¹, Fabio Maselli², Marta Chiesi², Lorenzo Bottai², Flora De Natale¹, Laura Ciaghi¹

(1) Istituto Sperimentale per l'Assestamento Forestale e l'Alpicoltura, P.za Nicolini 6, 38050 Villazzano, Trento, Tel +39 0461 381122, Fax +39 0461 381131, caterina.gagliano@entecra.it

(2) CNR – Istituto di Biometeorologia, Via Madonna del Piano 10, 50019 Sesto Fiorentino, Firenze

Riassunto

Lo scopo del presente lavoro è quello di valutare l'efficacia di un approccio non parametrico nella classificazione delle coperture forestali di un'area del Trentino orientale. Lo studio è eseguito mediante l'uso di un campione di dati rilevati a terra utilizzato per la classificazione di un'immagine Landsat 5.

L'applicazione di un classificatore k-NN con diverse configurazioni dei parametri dell'algoritmo è valutata mediante le accuratze raggiunte. Poiché la configurazione classica del k-NN tende a distorcere le stime delle aree coperte dalle diverse classi, si è valutata la possibilità di apportare una soluzione a tale problema. La procedura consiste nel calcolo e nell'attribuzione di opportune probabilità a priori a ciascuna classe di copertura, in modo da guidare l'attribuzione dei pixel alle varie classi.

Abstract

The present paper aims at evaluating the performance of a non parametric approach applied to the classification of forest types in an area of Northern Italy (Trentino). The investigation has been carried out by classifying a Landsat 5 TM image on the basis of a reference data set derived from ground survey.

The work evaluates the accuracy obtained with different configurations of a k-NN classifier. As the conventional configuration of the method leads to a distortion in the estimate of the different class surfaces, a correction methodology is also considered. This consists of computing and applying appropriate prior probabilities to each forest type in order to influence the pixel attribution to the various classes.

Introduzione

Gli strumenti principali per la conoscenza degli aspetti qualitativi e quantitativi del patrimonio forestale sono ancora oggi le indagini di campo, ma le nuove tecnologie disponibili hanno reso possibile l'applicazione di metodi di analisi alternativi.

In particolare il telerilevamento da satellite si è dimostrato uno degli strumenti di monitoraggio più efficaci ed economicamente competitivi per tali studi grazie alla sinotticità, all'alta risoluzione spaziale ed alla frequenza dei dati acquisibili; inoltre costituisce un utile strumento di supporto ai metodi tradizionali di raccolta dei dati in campo.

L'impiego operativo di dati satellitari per la stima di attributi forestali deve tuttavia confrontarsi con una serie di problemi, legati alle complesse relazioni analitiche tra risposte spettrali e attributi come densità, volume, etc. Molti autori (Tomppo, 1990; Maselli et al., 2005) evidenziano la possibilità di superare questo limite attraverso l'uso di classificatori più flessibili come quelli non parametrici.

Uno dei più comuni stimatori non parametrici applicato in studi forestali è l'algoritmo k-NN.

Gli algoritmi 'NN' sono stati ampiamente utilizzati come strumenti di riconoscimento dei *pattern* fin dagli anni '50 (Fix & Hodges, 1951) e attualmente trovano applicazione anche negli studi geostatistici, oltre che nella stima di variabili forestali discrete (uso del suolo) e continue (volume, area basimetrica, etc.) (Muinonen e Tokola, 1990).

Le stime non parametriche non richiedono la conoscenza a priori della funzione che descrive il fenomeno oggetto di analisi (Tomppo, 2004; Maselli et al., 2003) ma si basano sull'ipotesi che la variabilità totale sia ben rappresentata dal campione a terra.

Il successo di questa tecnica è dovuto al fatto che permette di derivare strati informativi che descrivono in modo spazialmente continuo variabili quantitative e qualitative a partire da dati puntuali.

Il metodo presenta delle limitazioni legate principalmente alla numerosità del campione a terra, che deve essere adeguata a caratterizzare le diverse situazioni presenti. In particolare, le prestazioni del classificatore diventano critiche in situazioni ecologicamente complesse, dove le firme spettrali delle superfici sono determinate da molti fattori che interagiscono fra loro (orografia, densità della copertura arborea, etc.) (Maselli et al., 2003).

Altri fattori che contribuiscono ad influenzare fortemente l'accuratezza delle stime sono lo scostamento delle coordinate dei punti sull'immagine rispetto alla posizione reale al suolo (Halme & Tomppo, 2001; Katila & Tomppo, 2001), l'errore di rettifica dell'immagine e la risoluzione delle immagini utilizzate, il periodo intercorso tra il rilievo a terra e l'acquisizione dell'immagine.

Tali fattori rendono le prestazioni del k-nn dipendenti dalla identificazione di una configurazione ottimale da cercare sperimentalmente per ogni caso considerato (Maselli et al., 2003).

Lo scopo del presente lavoro è quello di valutare l'efficacia di tale approccio nella classificazione delle coperture forestali di un'area del Trentino orientale mediante l'integrazione di immagini telerilevate Landsat 5 con un set di ortofoto e dati rilevati a terra.

Area di studio

L'area di studio corrisponde al Comprensorio dell'Alta Valsugana, ha un'estensione di circa 40.000 ha ed è caratterizzata da un territorio particolarmente complesso dal punto di vista orografico, con altitudini comprese tra 600 e 2400 m.

L'area selezionata è inoltre caratterizzata da un paesaggio molto articolato in cui prevale una copertura forestale interrotta da prati e pascoli. La vegetazione forestale è dominata dalle conifere, in particolare dall'abete rosso, ma sono presenti anche l'abete bianco, il larice, ecc., mentre le latifoglie sono per lo più relegate a cedui presenti nella fascia collinare-submontana e nei fondivalle e sono costituite principalmente da popolamenti di faggio.

Materiali e metodi

Data set e pre-elaborazione delle immagini

La sperimentazione è stata condotta utilizzando una scena Landsat TM del 19 giugno 2000. Per assicurare la corrispondenza tra l'immagine e i punti inventariali in campo la stessa immagine è stata ortorettificata con il software PCI usando un DEM con una risoluzione di 10 m (trasformazione polinomiale di 1° grado e ricampionamento con metodo *Nearest Neighbour*).

Il training set è costituito da un campione di 579 punti distribuiti in modo piuttosto uniforme nell'intera area di studio. Circa il 10 % dei punti è stato rilevato mediante una campagna a terra; i restanti punti sono stati classificati tramite la fotointerpretazione di ortofoto digitali in bianco e nero del 1996 a risoluzione di 1 metro (Agea ex-AIMA).

Nella fase di training sono state individuate 6 classi di copertura forestale.

Classi di copertura forestale	n. di punti	%
Boschi a prevalenza di abete rosso e/o abete bianco	268	46.3
Boschi di altre latifoglie	61	10.5
Boschi di altre conifere	86	14.9
Boschi a prevalenza di castagno	12	2.1
Boschi a prevalenza di faggio	65	11.2
Boschi a prevalenza di larice	87	15.0

Tabella 1. Training set

Metodo di stima

Il metodo k-NN si basa sull'ipotesi che nell'area in esame esistano soprassuoli forestali con caratteristiche simili e che la loro risposta radiometrica sia dipendente dagli attributi considerati. Con tale metodo per ogni pixel p_i della classe l nell'immagine viene misurata la distanza spettrale da ciascuno dei pixel per i quali sia stato rilevato l'attributo forestale di interesse. Per il calcolo l'algoritmo utilizza la distanza euclidea calcolata come sommatoria quadratica non pesata delle differenze spettrali nelle varie bande tra i pixel in esame e quelli di *training*. In caso di variabili discrete, come la classe di uso del suolo, l'algoritmo attribuisce ai pixel incogniti il valore modale tra i k pixel di *training* più vicini spettralmente:

$$l_{(p_i)} = \text{classe più frequente nel set con } i = 1, \dots, k$$

Calibrazione dell'algoritmo

Prima di applicare l'algoritmo, è necessario trovare la sua migliore configurazione tramite la modifica di alcuni parametri di stima. Per minimizzare l'influenza dell'errore di posizionamento dei plot nelle stime a livello di pixel e dell'errore dovuto alla registrazione dell'immagine è stato applicato un filtro medio a finestra variabile (3x3 e 5x5). In questo modo la risposta spettrale delle unità campione è stata calcolata come media dei valori spettrali degli nxn pixel circostanti. L'applicazione dell'algoritmo su immagini filtrate è stata confrontata con quella effettuata sull'immagine originale semplicemente corretta geometricamente.

Sebbene la selezione di un'area geografica di riferimento non sia un'operazione indispensabile e non faccia parte della procedura classica (Katila e Tomppo, 2001), nel presente studio si è scelto di valutare quali fossero le distanze orizzontali e verticali entro le quali selezionare i pixel di training. Poiché la selezione delle bande può influire sull'accuratezza della stima (Franco-Lopez, 2001), la configurazione dell'algoritmo è stata verificata, oltre che sulle singole bande, anche su combinazioni multispettrali.

Le *performance* della calibrazione sono state valutate in termini di accuratezza totale, media ed areale (Maselli et al., 1990) mediante la procedura di cross validation *leave one out*, che consiste nell'omettere uno per volta i punti di *training*, nel calibrare l'algoritmo sui punti rimanenti e nell'applicarlo sul punto omissso (Gong, 1986).

In tutte le prove è stata utilizzata una maschera bosco/non bosco dell'area di studio.

Configurazione alternativa dell'algoritmo

Una volta trovata la configurazione ottimale in grado di minimizzare l'entità dell'errore di classificazione, l'algoritmo è stato modificato al fine di considerare diverse probabilità a priori per ciascuna classe.

L'obiettivo è quello di migliorare la stima areale del classificatore, cioè la corrispondenza fra le superfici realmente coperte dalle varie classi e quelle ad esse attribuite.

Le probabilità a priori (*PP*) sono state calcolate iterativamente come rapporti tra il numero di punti realmente appartenenti ad ogni classe l ed il numero di punti attribuiti ad ogni classe l dal classificatore elevato ad un fattore α rapportati alla sommatoria di tali rapporti per tutte le classi.

$$l_{(p_i)} = \text{classe più comune nel set pesato con PP}$$

$$PP = \frac{\left[\frac{C_l}{R_l} \right]^\alpha}{\sum_l \left[\frac{C_l}{R_l} \right]^\alpha}$$

C_l = somma dei punti realmente appartenenti alla classe l

R_l = somma dei punti attribuiti alla classe l dal classificatore

All'iterazione successiva alla prima, nella quale si utilizza una probabilità a priori uguale per ogni classe, il numero di punti realmente appartenenti a ciascuna classe viene messo in rapporto con il numero di punti attribuiti dal classificatore per ottenere un nuovo peso. Il processo viene ripetuto fino a quando le probabilità a priori tendono a stabilizzarsi per tutte le classi.

Risultati

Inizialmente l'algoritmo è stato utilizzato su dati monospettrali. In accordo con quanto riportato in letteratura (Lillesand & Kiefer, 2004), la banda spettrale in grado di fornire il miglior risultato in termini di accuratezza è risultata la banda 4 (infrarosso vicino).

Questi risultati sono in contrasto con quanto dimostrato da altri autori (McRoberts et al., 2002), ma tali differenze sono giustificabili dal fatto che la banda in grado di fornire il miglior risultato cambia al variare della densità dei soprassuoli, in funzione del campione di training utilizzato e della stagione vegetativa considerata (Spanner et al., 1984).

HRA (km)	VRA (m)	accuratezza
2	200	0.656
2	300	0.661
2	400	0.663
2	500	0.667
2	700	0.658
3	200	0.653
3	300	0.653
3	400	0.658
3	500	0.658
3	700	0.653
4	200	0.651
4	300	0.651
4	400	0.642
4	500	0.641
4	700	0.627
5	200	0.665
5	300	0.656
5	400	0.648
5	500	0.632
5	700	0.618

Attraverso prove successive (tab. 2) sono stati individuati il *range* orizzontale (HRA) e quello verticale (VRA) in grado di minimizzare l'errore di classificazione: tali valori corrispondono alla distanza massima entro cui sono individuare i pixel di riferimento e sono risultati pari a 2 km e 500 m rispettivamente.

Tabella 2. Accuratezze ottenute con le diverse combinazioni di HRA e VRA.

Il processo di calibrazione è stato concluso confrontando l'uso di dati multispettrali con quello di dati monospettrali. L'accuratezza più bassa, pari al 60%, è stata ottenuta in modalità monospettrale con la banda 6 e tale risultato può essere ampiamente giustificato dalla sua minore risoluzione geometrica. La migliore configurazione dell'algoritmo, ottenuta combinando le bande TM2, TM3 e TM7, ha permesso di raggiungere accuratezze del 67%.

La classificazione ottenuta alla fine del processo di calibrazione (tab. 3 e fig. 1) risulta essere fortemente sbilanciata verso la categoria più frequente, quella dei boschi di abete rosso e/o abete bianco, che tende ad essere dominante nel processo di attribuzione dei pixel. La classe "lariceti", se pur rappresentata da un cospicuo numero di punti campione, non viene classificata correttamente per la similarità spettrale con la classe delle abetine.

		data set						
		ab	laric	altrconif	fag	cast	altrlat	
classification data	ab	236	36	16	16	0	3	307
	laric	7	29	7	3	1	5	52
	altrconif	9	7	46	3	1	6	72
	fag	11	7	7	34	0	5	64
	cast	0	1	1	0	7	4	13
	altrlat	5	7	9	9	3	38	71
n. punti totali		268	87	86	65	12	61	579
n. punti corretti		236	29	46	34	7	38	
accuratezza %		0.88	0.33	0.53	0.52	0.58	0.62	

Tabella 3. Matrice di confusione ottenuta alla fine del processo di calibrazione

L'opzione di attribuire iterativamente diverse probabilità a priori ad ogni classe considerata, calcolate sulla base del reale comportamento del classificatore, permette di ridurre questo squilibrio associando alle classi meno rappresentate una probabilità maggiore di attribuzione e a quelle più rappresentate una probabilità minore.

L'effetto conseguente (fig. 2) è una leggera diminuzione dell'accuratezza totale e media ed un evidente miglioramento delle stime delle superfici forestali in termini di accuratezza areale. Le classificazioni risultanti infatti rappresentano in modo più preciso la reale distribuzione delle superfici forestali considerate.

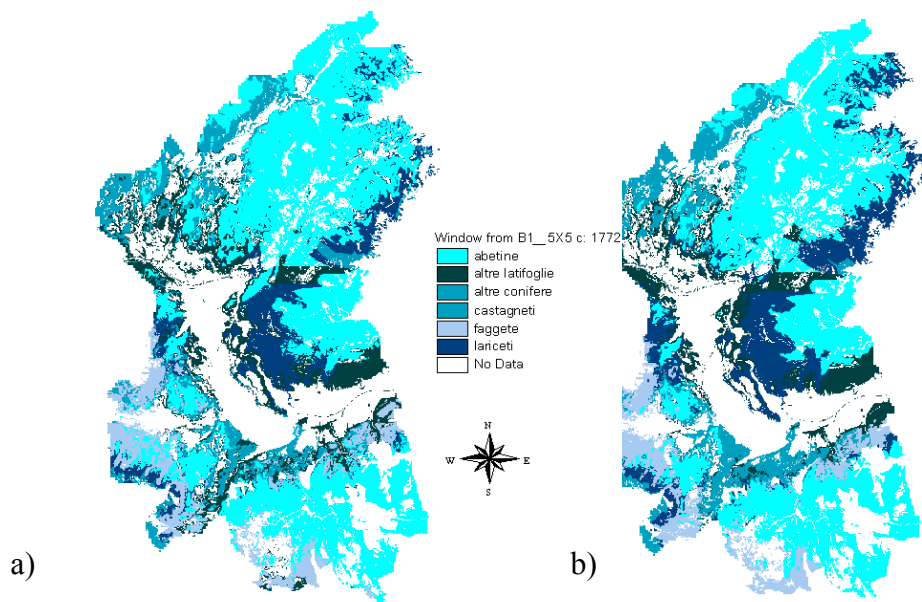


Figura 1. a) classificazione dell'area di studio con probabilità a priori uguali per ogni classe; b) classificazione dell'area di studio con probabilità a priori variabile

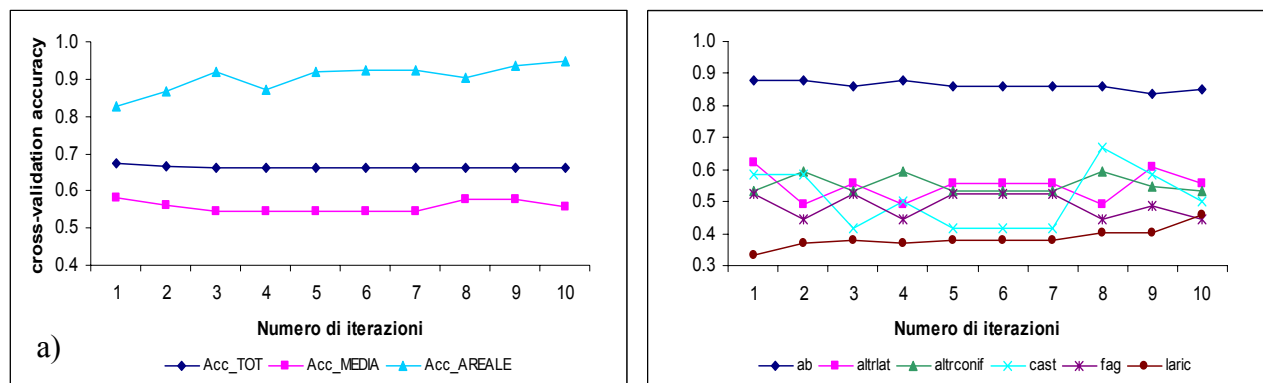


Figura 2. a) Accuratezze totale, media ed areale raggiunte al variare della probabilità a priori; b) accuratezza media per classe raggiunta al variare della probabilità a priori.

Conclusioni

In questo studio è stata testata l'efficacia di un classificatore k-NN applicato a dati derivati da fotointerpretazione e rilevati a terra. I valori di accuratezza totale ottenuti si sono mantenuti pressoché costanti attorno al valore 0.66. Se pur tale risultato sia inferiore rispetto a quanto citato in letteratura, si può supporre che la complessità topografica e la frammentazione che caratterizzano l'area di studio influiscano sull'accuratezza. Inoltre la prova è stata condotta considerando una classificazione piuttosto dettagliata (6 classi forestali). Accuratezze maggiori sono ottenute generalmente per diverse classi di uso del suolo e non per una sottoclassificazione dello stesso uso (nel nostro caso forestale).

L'applicazione di probabilità a priori a ciascuna classe di copertura forestale non ha comportato un cambiamento sostanziale della precisione né totale né media ma tale scelta si è comunque dimostrata utile soprattutto per la classe dei "lariceti", la cui accuratezza è aumentata del 10%.

In generale l'algoritmo k-NN si è dimostrato un efficace strumento per la classificazione delle diverse categorie forestali presenti in Trentino e potrebbe trovare una possibile applicazione a più ampia scala, lasciando intravedere un'interessante prospettiva come metodologia a supporto dell'inventario forestale nazionale.

I risultati di questo lavoro potrebbero trovare ulteriori conferme in regioni a diverse caratteristiche topografiche e forestali e dovrebbero essere discussi in funzioni di ulteriori verifiche dell'algoritmo.

Bibliografia

Franco-Lopez H., Ek A.R., Baue M.E. (2001), "Estimation and mapping of forest stand density, volume, and cover type using the k-nearest neighbors method", *Remote Sensing of Environment*, 77: 251-274

Gong G. (1986), "Cross-validation, the jackknife, and the bootstrap: excess error estimation in forward logistic regression", *Journal of the American Statistical Association*, 81 (393), 108-113

Halme M. & Tomppo E., (2001), "Improving the accuracy of multisource forest inventory estimates by reducing plot location error – a multi-criteria approach", *Remote Sensing of Environment*, 78: 321-327

Katila M., & Tomppo E. (2001), "Selecting estimation parameters for the Finnish multisource National Forest Inventory", *Remote Sensing of Environment*, 76, 16- 32

Lillesand T.M., Ralph W. Kiefer, and Jonathan W. Chipman (2004). *Remote sensing and image interpretation*. 5th ed. New York, Wiley, pp 763

Maselli F., Bottai L., Chirici G., Corona P., Marchetti M., Travaglini D. (2003), Stima di attributi forestali in ambiente mediterraneo tramite integrazione di misure a terra e dati telerilevati. *Italia Forestale e Montana* 4: 251-263

Maselli F., Conese C., Zipoli G., Pittau M.A. (1990), "Use of error probabilities to improve area estimates based on maximum likelihood classification", *Remote Sensing of Environment*, 31:155-160

Maselli F., Chirici G., Bottai L., Corona P. Marchetti M. (2005), "Estimation of mediterranean forest attributes by the application of k-NN procedures to multitemporal Landsat ETM+ images", *International Journal of remote sensing*, Vol. 26, No.17 pp. 3781-3796

McRoberts R., Nelson M., Wendt D.G. (2002), "Stratified estimation of forest area using satellite imagery, inventory data, and the k-NN technique", *Remote Sensing of Environment*, 82:457-468

Muinonen E., & Tokola T. (1990), *An application of remote sensing for communal forest inventory*. In: *The usability of remote sensing for forest inventory and planning*. Proceedings from SNS/IUFRO workshop 35- 42. Umea°, Sweden: Remote Sensing Laboratory, Swedish University of Agricultural Sciences (Report 4)

Spanner M.A., Brass J.A., Patterson D.L. (1984), "Feature selection and the information content of thematic mapper simulator data for forest structural assessment", *IEEE Transactions on Geoscience and Remote Sensing*, 22 (6): 482-489

Tomppo E. (1993), *Multi-source National Forest Inventory of Finland*. Proceedings of Ilvessalo Symposium on National Forest Inventories, Finland, 17-21 August, pp. 52-59

Tomppo, E. (1990), *Designing a Satellite Image-Aided National Forest Survey in Finland*. *The Usability of Remote Sensing For Forest Inventory and Planning*, Proceedings from SNS/IUFRO workshop in Umea 26-28 February 1990. Swedish University of Agricultural Sciences, Remote Sensing Laboratory, Report 4. 43- 47. Umea°, Sweden. ISBN 91-576-4208-7

Tomppo E. (2004), "Using coarse scale forest variables as ancillary information and weighting of variables in k-NN estimation: a genetic algorithm approach", *Remote sensing of environment*, 92:1-20