Valutazione delle funzioni di appartenenza in una classificazione Fuzzy K-Means per l'analisi dell'uso del suolo da immagini IKONOS

Eva Savina Malinverni

Università Politecnica Marche - Facoltà di Ingegneria – DARDUS, Via Brecce Bianche – 60131 Ancona
e-mail: e.s.malinverni@univpm.it

Abstract:
La sperimentazione si basa sull'applicazione e il confronto di due classificatori *unsupervised*: *K-means* e *Fuzzy K-means*. Le immagini multispettrali appartengono al sensore *IKONOS* e riprendono un'area attorno alla provincia di Ancona comprendente il Parco del Conero con una grande variabilità di aspetti morfologici e un'ampia rappresentatività di coperture del suolo. Selezionando le bande ottimali e definendo un diverso grado di legenda (4, 9, 15 classi) sono state ottenute diverse interpretazioni tematiche impiegando i due diversi algoritmi implementati allo scopo in un ambiente GIS *open source* (ILWIS 3.6). Pur avendo avuto in altre occasioni l'opportunità di applicare la classificazione automatica supervisionata a tali dati, si è voluto in questo caso sperimentare le caratteristiche di una classificazione *unsupervised,* mettendo in evidenza come un approccio *fuzzy* abbia una migliore capacità interpretativa. La tecnica *fuzzy* si è infatti rivelata più opportuna in fase di assegnazione delle classi. Rinunciando ad attribuire in modo univoco ogni pixel ad una determinata classe, come in una classificazione di tipo "*hard*" e, utilizzando il grado di appartenenza di ogni singolo pixel ad ognuna delle classi in legenda, è stato possibile valutare al meglio le soluzioni miste e darne la definitiva assegnazione. Mentre l'accuratezza globale di entrambi i classificatori soddisfa i limiti ammessi dalla letteratura, l'indice di confusione delle mappe *fuzzy* da una valutazione puntuale sulla stabilità della classificazione, consentendo insieme ad altri indicatori di valutare le iterazioni della classificazione stessa.

This research compares two unsupervised classifiers, K-means and Fuzzy K-means, applied to IKONOS imagery. The select territory (part of the Province of Ancona) has the presence of different urban fabric density, commercial areas, agricultural and forest environment. Selecting the optimal bands and defining a different degree of nomenclature (4, 9, 15 classes), different thematic land use maps were derived. To test the unsupervised classification characteristics and the capacity of a fuzzy approach to discover and assign the correct classes I implemented ex-novo in open source GIS environment (ILWIS 3.6) the personalized procedures. The fuzzy approach is resulted more opportune in presence of mixed covers, because the values of the membership functions allow to determine the class as well as the "hard" classification. The overall accuracy of both methods is in the limits admitted by the literature, on the contrary the confusion index of the fuzzy maps gives information on the stability of the classification and other parameters help to stop the procedure.

Introduzione

Il suolo, inteso come territorio, è lo spazio sul quale si svolgono quasi tutte le attività umane; intendendo per attività non solo quelle produttive, ma qualsiasi azione compiuta dall'uomo. In questo contesto risulta evidente che informazioni riguardanti il suolo sono necessarie per una corretta pianificazione del suo utilizzo e conservazione. Queste informazioni, utilizzate in molteplici ambiti, sia della pubblica amministrazione che della ricerca scientifica, sono organizzate in carte tematiche. Le carte uso del suolo, rappresentando una situazione legata strettamente alla dimensione temporale, raggiungono il grado di obsolescenza molto velocemente e quindi la loro

utilità può esaurirsi nel giro di pochi anni. L'avvento della tecnologia satellitare con immagini che hanno raggiunto risoluzioni spaziali paragonabili a quelle delle fotografie aeree, ha fornito l'opportunità di produrre cartografia tematica non solo in maniera fotointerpretativa ma anche e soprattutto in maniera automatica. Il metodo della interpretazione analogica, ancora oggigiorno ampiamente utilizzato, può dare buoni risultati in termini di precisione (sia a livello di classe identificata che a livello di definizione areale), se è condotto da personale esperto che segue procedure codificate, ma rimane comunque un processo prettamente soggettivo, molto lento e quindi assai dispendioso. La classificazione numerica diventa quindi preferibile quando si lavora con aree ampie o servono informazioni in tempi brevi e a costi relativamente contenuti.
L'area di indagine di questa sperimentazione è il territorio che si sviluppa attorno alla provincia di Ancona con un'estensione di circa 150 kmq (Figura 1)[1], caratterizzato da aspetti sia costieri che montani, paesaggi urbani e rurali, ambienti naturali ed insediamenti antropici. Il dataset di immagini multispettrali a 11 bit è stato acquisito nel Luglio 2006 dal satellite IKONOS[2]; successivamente è stato georeferenziato sulla base del Datum UTM-WGS84.
Pur avendo avuto in altre occasioni l'opportunità di applicare la classificazione automatica supervisionata a tali dati (Zingaretti et al., 2008) si è voluto in questo caso sperimentare le caratteristiche di una classificazione *unsupervised*. In particolare sono stati implementati ex-novo nel software GIS *open source* ILWIS 3.6 due algoritmi: uno basato sull'approccio *K-Means* e l'altro che sviluppa la tecnica *Fuzzy K-means*; al fine di ottenere interpretazioni sempre più oggettive e nel contempo avere strumenti idonei per la valutazione della correttezza dei risultati.

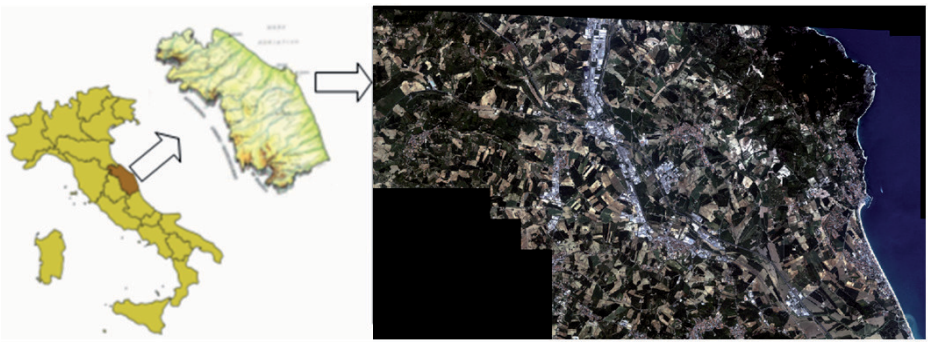

Figura 1 Area di studio e scena test acquisita dal sensore IKONOS sulla Regione Marche

Metodologia di classificazione: algoritmi *K-Means* e *Fuzzy K-Means*
Nella classificazione non supervisionata (*unsupervised*) l'assegnazione dei pixel ai *cluster* è condotta senza necessità di dati esterni ma solo in funzione dei valori radiometrici dello spazio multispettrale, una volta definito inizialmente il numero dei *cluster* da creare. La relazione di quest'ultimi con le classi informative di uso del suolo avviene solo successivamente attraverso sopralluoghi sul campo o mediante il confronto con mappe fotografiche.
I metodi di clusterizzazione presenti in letteratura si basano su differenti misure di distanza. Gli algoritmi implementati in questa ricerca usano la media come misura di distanza. L'idea che è alla base dell'algoritmo *K-Means* è quella di fissare a priori il numero di *cluster k* e di migliorare via via la clusterizzazione, assegnando al *cluster* k_i quei pixel che hanno valore più prossimo a quello della sua media m_k, ricalcolando, di volta in volta, la media stessa che corrisponde al centro del *cluster*. Inizialmente l'algoritmo sceglie in maniera arbitraria *k* centri rappresentativi di ogni *cluster* e le iterazioni terminano quando i centri non subiscono più consistenti variazioni.

[1] Per una migliore visualizzazione nelle figure successive si riporta solo una porzione dell'area indagata
[2] I dati remote sensing del sensore IKONOS sono stati messi a disposizione dalla Regione Marche (progetto Anconapaco) con un accordo di ricerca con l'Università Politecnica delle Marche.

E' stato proprio il riconoscimento della natura vaga ed imprecisa di molti fenomeni, soprattutto, nel campo ambientale, che ha portato al superamento della logica di tipo binario a cui è legata una classificazione di tipo "*hard*". La teoria dei *fuzzy sets* (Zadeh, 1965) indica che ogni elemento x_i appartiene ad un dato insieme A secondo un certo grado di appartenenza compreso nell'intervallo [0, 1]; gli estremi indicano rispettivamente il minore ed il maggiore grado di appartenenza, mentre tutti i valori intermedi indicano il *continuum* delle appartenenze. Questo concetto di appartenenza sfumata torna utile nei problemi di classificazione di dati telerilevati soprattutto a bassa risoluzione dove è più facile trovare i cosiddetti *mixel*: pixel la cui appartenenza integralmente ad una classe non è ben definita.
L'algoritmo *K-Means* diventa così *Fuzzy K-Means* (Dunn, 1973, Bezdek, 1981). Esso utilizza un procedimento iterativo che assegnando i pixel ai *cluster* sulla base sempre della ben nota misura della distanza Euclidea ne determina anche il loro valore di appartenenza (*membership value*), che fornisce il grado di affinità con il centro del *cluster* (Bezdek et al., 1984). Si dimostra come i *cluster fuzzy k-means* sono "reciprocamente esclusivi", nel senso che la somma delle loro appartenenze è uguale a 1 (Figura 2).

Figura 2 Membership function del Cluster 4, i valori chiari indicano l'alto grado di appartenenza a quel cluster, per gli altri pixel abbiamo valori sfumati fino al nero che indica nessuna appartenenza

Una volta che i valori di appartenenza sono stati calcolati, un'osservazione può essere assegnata definitivamente ad una classe "*hard*", formata da tutte le osservazioni che hanno la loro più alta appartenenza a quella classe. Questa operazione prende il nome di "defuzzificazione" (*defuzzification*). Osservazioni che non possono essere attribuite ad una classe "*hard*" possono essere "*intragrades*", ossia assumono valori intermedi tra due o più classi; oppure "*extragrades*", ossia che appartengono un po' a tutte le classi. Vedremo in particolar modo nell'applicazione qui presentata come le funzioni di appartenenza possono guidare la successiva assegnazione della classe uso del suolo al *cluster* anonimo con maggior facilità.
L'applicazione dei due algoritmi *K-Mean*s e *Fuzzy K-Mea*ns è stata eseguita sulla composizione di immagini Red, Blue e NiR, considerata combinazione ottimale dall'analisi della matrice di correlazione e dall'indice statistico *Optimum Index Factor* (*OIF*), che in *ILWIS* fornisce la combinazione di immagini con il più alto valore di informazione (riassumibile nella più alta somma della deviazione standard) evitando ripetizioni di bande correlate. Un'ulteriore sperimentazione è stata eseguita applicando i due algoritmi a delle pseudobande, frutto della combinazione delle bande originali, calcolate per mezzo di una trasformazione per Componenti Principali (*PCA*). In questo caso sono state utilizzate le prime tre componenti che spiegano oltre il 99% della varianza totale delle quattro bande originali.
In figura 3 sono messe a confronto le classificazioni *unsupervised* ottenute mediante l'algoritmo *Fuzzy K-Means,* applicato sia alla combinazione delle immagini originali che alle pseudobande. Risultati analoghi si sono ottenuti applicando l'algoritmo *K-Means*.

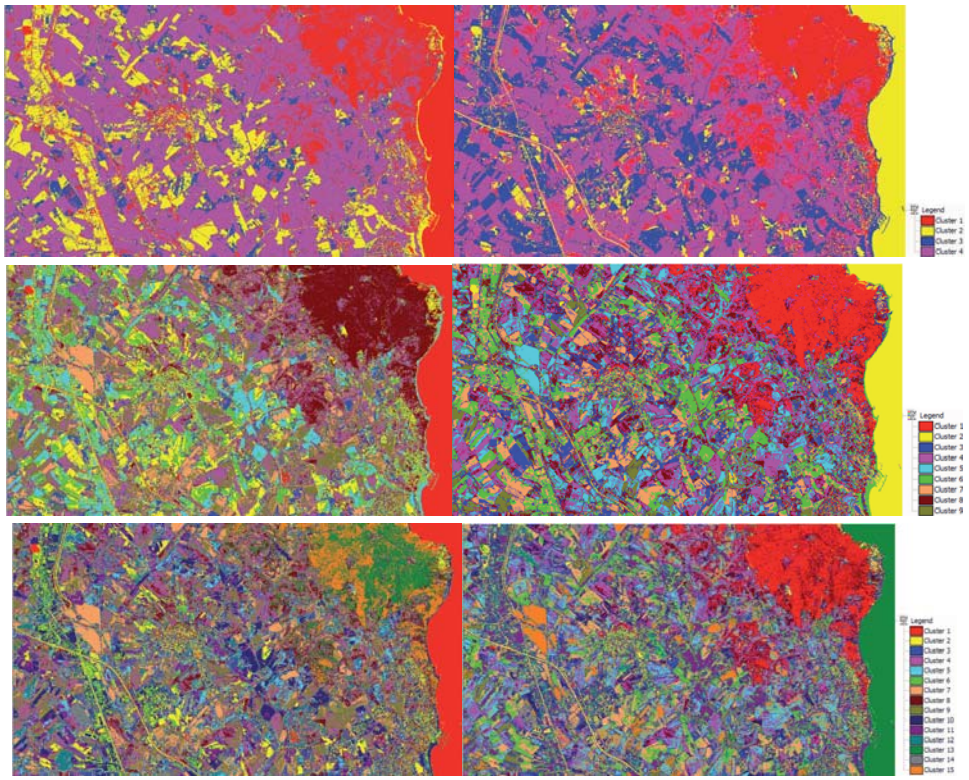

Figura 3 Classificazioni Fuzzy-K-means a 4-9-15 cluster, impiegando le immagini PCA (sx) e le immagini RBNiR (dx)

Uso delle funzioni di appartenenza per l'assegnazione delle classi ai *cluster*: generazione della carta tematica

Per ottenere la mappa tematica uso del suolo un aspetto delicato da affrontare è quello della definizione consistente e adeguata delle classi in legenda. La definizione di una classe diviene un problema di scelta di convenzione, che dipende, chiaramente, dal tipo di applicazione e di enfasi che si pone su un particolare tipo di uso del suolo. La legenda adottata nella sperimentazione si basa sulla struttura *Corine Land Cover*, opportunamente dettagliata ed accorpata per adeguarla alle caratteristiche del territorio in esame. Partendo da un concetto di classe gerarchico inizialmente si è definito un numero limitato di classi al livello più alto (4 classi: "Superfici artificiali", "Vegetato", "Ambienti agricoli e seminaturali" ed "Acque superficiali"), per poi approfondirle ulteriormente a 9 ed infine a 15 classi. Una fase importante della classificazione *unsupervised* è l'interpretazione e assegnazione della classe ai *cluster*, di norma eseguita da un operatore sulla base delle proprie capacità interpretative. Per cercare di automatizzare il processo e renderlo il meno soggettivo possibile si è eseguito un *overlay* tra i dati classificati e i campioni delle verità a terra. La difficoltà maggiore si è riscontrata nell'assegnazione delle classi alla classificazione *K-Means*, per la presenza di *mixel* appartenenti a più classi spettrali, soprattutto all'aumentare del numero di classi. Per la classificazione *Fuzzy K-Means* invece, sfruttando le informazioni fornite dalle *membership function* di ogni *cluster,* è stato possibile assegnare con certezza quelle classi che presentavano pixel con una probabilità di appartenenza in un *range* piuttosto elevato (80%-100%). A titolo di esempio in Figura 4 è rappresentato il processo decisionale di assegnazione della classe "*Boschi*" al *Cluster 1* nella mappa ottenuta con l'algoritmo *Fuzzy K-Means*.

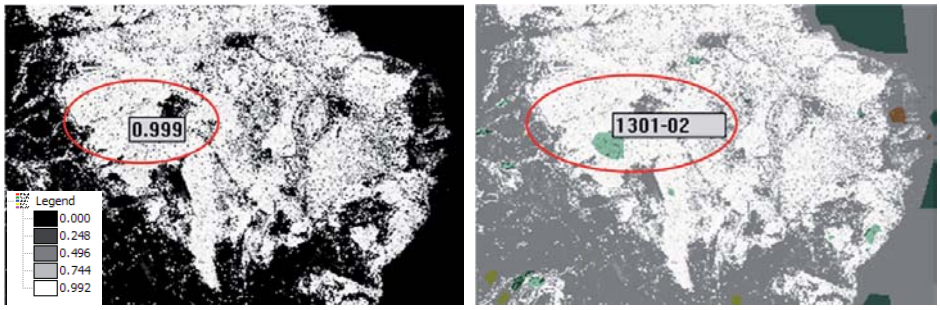

Figura 4 Membership function del Cluster1 (sx) usata in sovrapposizione con la mappa verità a terra assegna, il Cluster 1 alla classe "Boschi" (dx)

Una volta classificati i *cluster* con la maggiore probabilità di appartenenza, si sono affrontati i *cluster* incerti. L'uso delle *membership function* è utile soprattutto in questo caso, infatti valutando i diversi gradi di appartenenza ad una classe fra diversi *cluster*, questa viene assegnata al *cluster* che ha dimostrato un grado di appartenenza più alto. Il beneficio di questo metodo si ha all'aumentare del numero di classi dove si ottiene una migliore discriminazione fra *cluster*, con classi sempre meglio definite. Procedendo in questo modo si è potuto ottenere la mappa uso del suolo della classificazione *Fuzzy K-Means* con maggiore precisione e dettaglio (Figura 5).

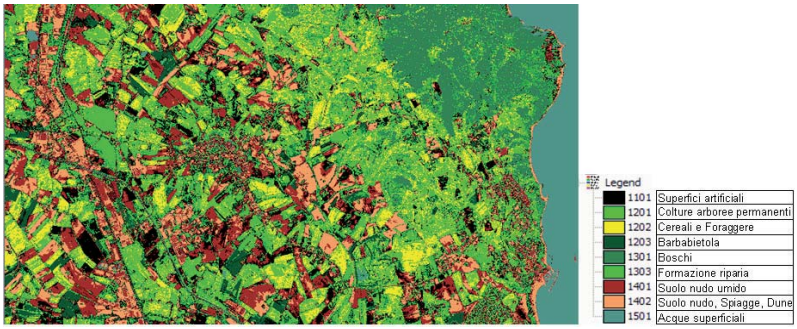

Figura 5 Mappa di uso del suolo della classificazione Fuzzy K-means a 9 classi

Valutazione dei risultati

Per ogni combinazione di set di immagini e di metodo di classificazione applicato, al variare del numero di classi selezionate, è stato operato un confronto con un set di verità a terra al fine di valutare l'accuratezza delle classificazioni. Questa procedura ha visto sia l'utilizzo del metodo classico della matrice di confusione, coi suoi parametri di omissione e commissione e la sintesi nell'*overall accuracy* (accuratezza globale) che l'utilizzo di indici mutuati dalla tecnica *fuzzy*. Tra questi ultimi ampiamente utilizzato è stato il *Confusion Index* (*CI*) che ha permesso di creare delle mappe di confusione per valutare in percentuale (0 perfetta appartenenza, 1 massima confusione) la corretta assegnazione del singolo pixel alla classe. Inoltre il confronto delle mappe del *CI* alle varie iterazioni diventa un modo per poter controllare la procedura iterativa. In figura 6(sx) l'esempio della mappa *CI* a 9 classi. Oltre al *CI* che indica la precisione della classificazione sul singolo pixel ci si può basare sui valori dei coefficienti *F* ed *H*, che rapportati al numero delle classi (cioè scalati), è bene che si attestino su valori prossimi a 1 il primo e a 0 il secondo (Burrough et al., 2000). Essi, in quanto indicativi della qualità della classificazione nel suo complesso, risultano un ulteriore parametro per decidere il numero di iterazioni. Il metodo applicato sia alle bande originali che alle pseudobande ha generalmente fornito dei buoni risultati, come nel caso della classificazione a 9

classi sulle bande originali dove si ottengono valori di F_{scaled} ed H_{scaled} alla quarta ed ultima iterazione rispettivamente pari a 0,70 e 0,05.

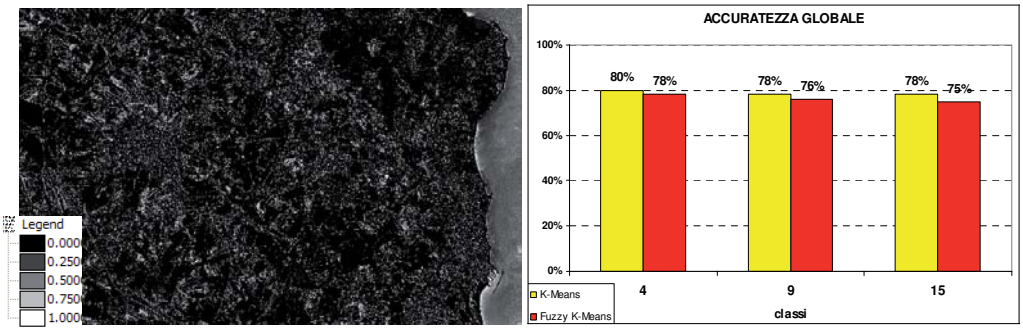

Figura 6 Valutazione della classificazione mappa del CI (sx) e overall accuracy (dx)

Inoltre il confronto con il campione di controllo a terra mediante la classica matrice di confusione ha riportato, per entrambi i metodi di classificazione, dei buoni valori di accuratezza globale, tra loro paragonabili (Figura 6(dx)). L'analisi della matrice di confusione delle diverse classificazioni a differenti classi, evidenzia un numero maggiore di classi classificate con minor incertezza nella *Fuzzy K-Means* grazie all'interpretazione condotta con l'ausilio delle *membership function*.

Conclusioni

L'applicazione di due classificatori *unsupervised*, il tradizionale *K-means* ed il *Fuzzy K-means* ha messo in evidenza la migliore capacità interpretativa di quest'ultimo approccio. La tecnica *fuzzy* rinunciando ad attribuire in modo univoco un pixel ad una determinata classe fornisce una serie di gradi di appartenenza di quel pixel ad ognuna delle classi definite in legenda. Questo basarsi sul concetto di probabilità di appartenere ad una classe piuttosto che ad un'altra permette di caratterizzare quantitativamente l'affidabilità di un risultato, attraverso la "defuzzificazione". Inoltre l'errore, valutato attraverso la messa a punto di indici *fuzzy* oltre alla classica matrice di confusione fornisce risultati più accurati e realistici rispetto al classificatore tradizionale. L'accuratezza globale per entrambi i classificatori soddisfa comunque i limiti generalmente ammessi dai dati forniti dalla letteratura.

Ringraziamenti

Si ringrazia l'Ing. Fabio Coccia che ha applicato, durante la sua tesi di laurea, gli algoritmi sviluppati dall'autore.

Bibliografia

Bezdek J. C. (1981), *Pattern Recognition with Fuzzy Objective Function Algoritms*, Plenum Press, New York.

Bezdek J. C., Ehrlich R., Full W. (1984), *FCM: the fuzzy c-means clustering algorithm*, Computers and Geosciences, n. 10.

Burrough P.A., van Gaans P.F.M, MacMillan R.A. (2000), *High-resolution landform classification using fuzzy k-means*, In Fuzzy Sets and Systems, 113, pp. 37-52.

Dunn J. C. (1973), A *Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters*, Journal of Cybernetics 3, pp. 32-57.

Zadeh L.A. (1965), *Fuzzy Sets. Information and Control*. 8, 338-353.

Zingaretti P., Frontoni E., Bernardini A., Malinverni E. S. (2008), *Automatic Classification of Central Italy Land Cover: Comparative Analysis of Algorithms*, Proceedings GEOBIA2008, ASPRS, ISPRS, CSA-ASC, Calgary, Alberta, Canada.