APPLICAZIONE DEI ROUGH SETS AI SISTEMI INFORMATIVI GEOGRAFICI. REALIZZAZIONE DI UN NUOVO MODULO IN GRASS

Antonio BOGGIA, Gianluca MASSEI

Dipartimento di Scienze Economico-Estimative e degli Alimenti dell'Università degli Studi di Perugia. Borgo XX Giugno, 74 Perugia 06121. Tel.: 075 5857136; fax: 075 5857146; e-mail: boggia@unipg.it, g_massa@libero.it

Riassunto

L'obiettivo del presente lavoro è quello di implementare un nuovo modulo in GRASS che, attraverso l'impiego della *rough sets theory,* estragga informazione geografica. L'elaborazione di un set di dati secondo la teoria dei *rough sets* porta a definire regole decisionali che descrivono la realtà esaminata. Lo strumento utilizzato per implementare gli algoritmi è GRASS 6.3 che è stato arricchito di un modulo specifico scritto in C. Il nuovo modulo costruito è stato anche inserito nel toolbox di GRASS in QGIS 0.10.0. La verifica del modello teorico e del modulo applicativo è stata fatta su un'applicazione concreta riferita alla problematica degli incendi boschivi nel territorio umbro.

Abstract

The aim of this paper is to present the implementation of a new module in GRASS, to extract geographic information using the rough sets theory. According to the rough sets theory, processing a data set leads to define decision rules, able to describe the object studied. To implement the algorithms GRASS 6.3 has been used, adding a specific module written in C language. In addition, the new module has been included in the GRASS toolbox, in QGIS 0.10.0. To test the theoretical model and the applied module a case study on the forests fire in Umbria is presented.

1. Introduzione

La crescente disponibilità di informazione geografica richiede l'uso di strumenti in grado di trattare efficacemente i dati e di estrarne le informazioni necessarie per la programmazione del territorio, sia in fase di analisi che di progetto. In questo ambito applicativo l'impiego di tecniche di *data mining* e *machine learning* sembra costituire un'interessante ambito di ricerca applicativa, soprattutto se integrata nei più diffusi software GIS oggi disponibili. L'obiettivo del presente lavoro è quello di implementare un nuovo modulo in GRASS che, attraverso l'impiego della *rough sets theory,* estragga informazioni da database geografici.

2. Teoria dei rough sets e applicazione ai database geografici

La teoria dei *rough sets* (insiemi approssimati), introdotta da Pawlak (1982), si fonda sull'assunzione che ad ogni elemento dell'universo sono associate delle informazioni (dati, conoscenza), espresse utilizzando opportuni attributi. Elementi caratterizzati dalla stessa descrizione sono indiscernibili con riferimento alle informazioni disponibili. La relazione di indiscernibilità (*Ip*) così generata costituisce il fondamento matematico della teoria dei *rough sets*:

$$Ip = \{ (x,y) \in U \times U : \quad f(x,q) = f(y,q) \ \forall q \in P \} \qquad [1]$$

Un insieme di elementi indiscernibili è denominato "insieme elementare" e costituisce un "granulo" della conoscenza dell'universo considerato. L'unione degli insiemi elementari può essere preciso (ordinario) oppure *rough* (impreciso, vago). Ogni *rough set*, quindi, è costituito da quei casi che non possono essere classificati esattamente. L' insieme impreciso è definito da due insiemi ordinari, chiamati approssimazione inferiore (denominata $\underline{\ }(X)$) e superiore (denominata $\overline{\ }(X)$):

$$\underline{\ }(X) = \{ x \in U : Ip(x) \subseteq X \} \qquad [2]$$

$$\overline{\ }(X) = \{ x \in U : Ip(x) \cap X \neq \emptyset \} \qquad [3]$$

Gli elementi di $\underline{\ }(X)$ sono tutti e solo gli $x \in U$ appartenenti a tutte le classi generate dalla relazione di indiscernibilità Ip e contenuti in X. Gli elementi di $\overline{\ }(X)$ sono tutti e solo gli $x \in U$ appartenenti a tutte le classi generate dalla relazione di indiscernibilità Ip che hanno almeno un rappresentante appartenente ad X.
La frontiera di X, denominata con $Bn_P(X)$, è:

$$Bn_P(X) = \overline{\ }(X) - \underline{\ }(X) \qquad [4]$$

Pertanto, se un oggetto x appartiene a $\underline{\ }(X)$ esso è certamente anche un elemento di X, mentre se x appartiene a $\overline{\ }(X)$ esso può appartenere all'insieme X. $Bn_P(X)$ costituisce, quindi, la "regione del dubbio" di X: nulla può dirsi con certezza circa l'appartenenza dei suoi elementi all'insieme X.
Due indici, l'accuratezza dell'approssimazione $\alpha_P(X)$, e la qualità dell'approssimazione $\gamma_P(X)$, sono definiti come segue:

$$\alpha_P(X) = \frac{\#\underline{\ }(X)\#}{\#\overline{\ }(X)\#} \qquad [5]$$

$$\gamma_P(X) = \frac{|\underline{P}(X)|}{|X|} \qquad [6]$$

Lo scopo è quello di derivare un insieme di regole decisionali basate su determinati attributi, anche in presenza di situazioni di "approssimazione". Le regole possono essere utilizzate ai fini di classificazione, oppure per determinare relazioni di causa-effetto. Per questa sua capacità la *rough sets theory* è utilizzata come strumento di supporto alle decisioni. L'induzione delle regole derivate grazie all'uso della *rough sets theory* è ben rappresentata nella figura 1, proposta dal Obersteiner e Wilk nel 1999.

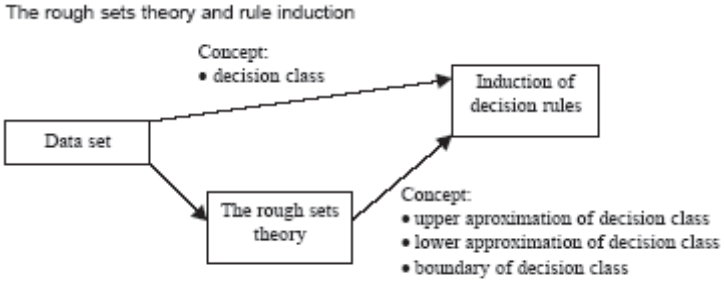

Figura 1 – Rough sets theory e induzione di regole (fonte: M. Obersteiner, S. Wilk, 1999)

La teoria dei *rough sets* dunque, si propone di analizzare le relazioni di causa-effetto tra dati caratterizzati da incertezza e vaghezza. Essa presenta alcuni punti di contatto con altre teorie che trattano l'incertezza e l'imprecisione: teoria della probabilità, teoria dei *fuzzy sets*, analisi discriminante, ecc. Essa offre notevoli potenzialità specialmente per quanto riguarda la descrizione e valutazione della dipendenza fra variabili e l'analisi della significatività di attributi rilevanti caratterizzati da informazioni di tipo qualitativo o quali-quantitativo. E' stata applicata con successo a numerosi problemi reali di classificazione in differenti campi. Gli interessanti risultati ottenuti hanno spinto recentemente studiosi di differenti discipline ad interessarsi allo studio di tale teoria ed alla sua implementazione. Per una raccolta di saggi su applicazioni dei *rough sets* a problemi reali si veda Slowinski (1992). Una breve ma esaustiva rassegna delle più importanti applicazioni si può trovare in Pawlak (1997).
La *rough sets theory* è basata sulle informazioni, e da esse avvia un percorso che porta fino alle regole decisionali. Sta qui la forza dell'integrazione di un modulo *rough set* all'interno di un sistema GIS. E' evidente, infatti, il vantaggio dell'estrazione delle informazioni da un database geografico, per la abbondanza e la collocazione spaziale dei dati, particolarmente utile quando si affrontano problemi decisionali legati alla gestione dell'ambiente e del territorio. In fase di output, inoltre, l'integrazione del modulo in sistema GIS consente la rappresentazione puntuale sul territorio dei risultati ottenuti.

3. Il modulo *r.roughset*

Lo strumento utilizzato per implementare gli algoritmi di analisi basati sui *rough sets* è GRASS 6.3 che è stato arricchito di un modulo specifico scritto in linguaggio C (Bellini, Guidi, 1994) e basato, oltre che sulle librerie standard di GRASS, sulle *rough set library (RSL ver. 2.0)* pubblicate da M.Gawrys´ e J.Sienkiewicz (1993).
Il nuovo modulo di GRASS si chiama *r.roughset* (Fig. 2), ed accetta in input tutti i file raster che rappresentano gli attributi descrittivi del fenomeno da studiare e un file raster con l'identificazione delle aree decisionali.
L'estrazione delle regole decisionali può avvenire secondo otto differenti algoritmi ed il risultato viene registrato in due diversi files. Nel primo vengono riportati in forma descrittiva le regole decisionali estratte, gli indici di qualità dei *rough sets* e ulteriori informazioni utili per fare valutazioni quali-quantitative basate proprio sulla teoria degli insiemi approssimati.

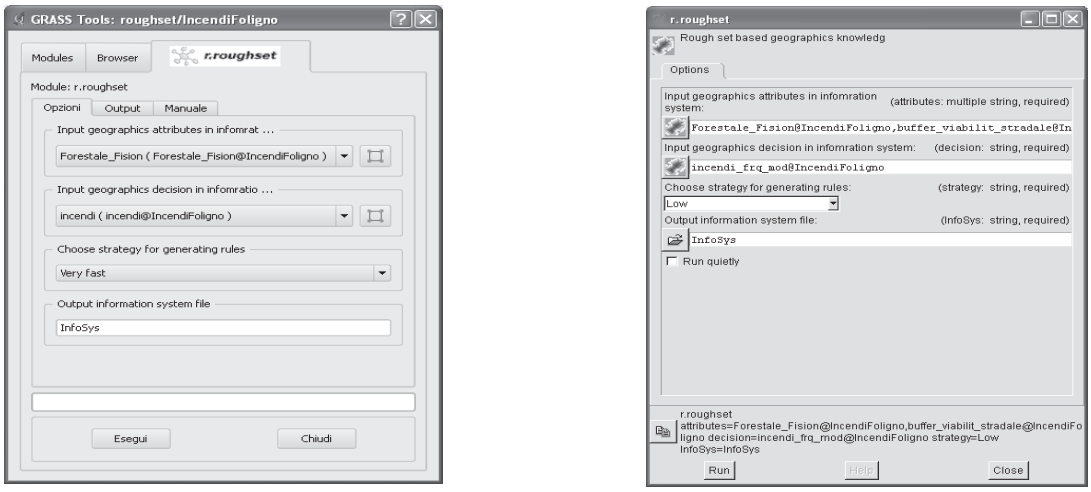

Figura 2 - Interfaccia del modulo r.roughset in GRASS 6.3 e in QGIS 0.10.0

Il secondo file di output è in realtà uno script della *shell bash* in stile *UNIX like* che applica le regole estratte ai file raster presenti nella location e che genera la relativa mappa di classificazione basate sui *rough sets*. Lo script può essere lanciato immediatamente dall'utente da una *shell* oppure può essere opportunamente modificato con l'aggiunta di ulteriori comandi di GRASS per approfondire l'analisi.
L'approccio descritto potrebbe ricordare quello della classificazione delle immagini nell'ambito dell'*image processing*. Anche se ciò può rappresentare un interessante campo di esplorazione dei *rough sets*, i principi di base sono sostanzialmente differenti. Il modulo *r.roughset* non implementa un algoritmo di classificazione delle immagini ma, piuttosto, costituisce un'interfaccia per le *rough sets library* e consente l'accesso a tutte le funzioni di analisi da queste implementate. La successiva fase di classificazione viene ottenuta da un'elaborazione dei file testuali di output e dall'applicazione ricorsiva di alcuni comandi standard di GRASS (*r.mapcalc, r.patch, g.remove, r.colors*).

4. Applicazione di *r.roughset* all'analisi degli incendi boschivi

Gli incendi boschivi rappresentano un elemento di forte criticità per gli ecosistemi forestali mediterranei. In questi ultimi anni la sensibilità del pubblico, gli strumenti di programmazione territoriale ed i controlli sul territorio hanno rivolto maggiore attenzione al problema con l'obiettivo di una progressiva riduzione della superficie forestale incendiata. Purtroppo, nonostante ciò, il fenomeno si ripete annualmente e gli interventi di emergenza non sempre riescono a contenere efficacemente i danni.
Le cause degli eventi sono sempre piuttosto difficili da individuare con certezza. Se si escludono i casi di tipo doloso che si riesce ad accertare, vi è una grande difficoltà ad individuare le motivazioni che hanno portato all'innesco dell'incendio. Gli stessi modelli matematici che sono stati sviluppati e sempre più spesso vengono impiegati nella pratica dell'antincendio non sempre riescono a fornire risposte attendibili nella previsione dell'evento.
La teoria dei *rough sets*, attraverso il modulo *r.roughset* implementato in GRASS 6.3, è stata utilizzata per analizzare database geografici di un comune umbro (Foligno) per il quale è disponibile un catasto decennale degli incendi boschivi con l'individuazione dei punti di innesco.
E' stato costruito un database geografico contenente l'uso del suolo, l'inventario forestale regionale, la viabilità stradale e ferroviaria, gli elettrodotti di alta tensione ed i centri abitati, la pendenza e l'esposizione dei versanti. Il sistema informativo geografico ottenuto, che costituisce l'insieme degli attributi che hanno determinato il verificarsi degli incendi, è stato integrato con un tema geografico che materializza il catasto delle aree percorse da incendio nel decennio 1998 – 2007. Tutti i punti di innesco degli incendi boschivi sono stati riclassificati in quattro differenti classi in base alla frequenza di eventi ripetutesi nel corso del decennio. Il sistema così ottenuto rappresenta la base per il calcolo del "Sistema Informativo" definito secondo la teoria dei *rough sets*.
Attraverso l'applicazione del modulo *r.roughset* è possibile estrarre le regole decisionali che secondo i presupposti della teoria e sulla base delle variabili territoriali fornite come input influenzano l'innesco di un incendio boschivo.
Le informazioni più significative ottenute dall'analisi sono riportate in Figura 3 ed i numeri identificano i temi geografici utilizzati nell'analisi secondo l'ordine di inserimento.
La parte più importante dell'output è rappresentato dalle regole decisionali che vanno lette da sinistra a destra secondo la sintassi: *if ... then* .....A fini esemplificativi, la prima regola decisionale ottenuta e riportata in Figura 3 va letta " se il bosco è classificato con categoria 6 allora la classe di frequenza di incendi è 2".
La traduzione cartografica in ambiente GIS è riportata in Figura 4 ed è generata dallo script che costituisce il secondo output del modulo *r.roughset*.

Mean number of attributes in rule = 1.7
Condition attributes are
{ 0,1,2,3,4,5 }
Decision attributes are
{ 6 }
DependCoef = 0.751843
CORE = { 0,2,4,5 }
RedOptim = { 0,2,4,5 }
Few reducts (1):
{ 0,2,4,5 }
5 **strategy of generating rules**
Time of generating rules = 1234s
Rules (6)
Forestale_Fision@IncendiFoligno=6 => incendi_frq_mod@IncendiFoligno=2 (65 objects)
Forestale_Fision@IncendiFoligno=10 => incendi_frq_mod@IncendiFoligno=4 (43 objects)
Forestale_Fision@IncendiFoligno=9 => incendi_frq_mod@IncendiFoligno=4 (70 objects)
Forestale_Fision@IncendiFoligno=5 buff_elettrodotti@IncendiFoligno=4 => incendi_frq_mod@IncendiFoligno=3 (62 objects)
buff_elettrodotti@IncendiFoligno=3 centri_buffer@IncendiFoligno=5 => incendi_frq_mod@IncendiFoligno=3 (20 objects)
UsoSuolo@IncendiFoligno=2 buff_strade@IncendiFoligno=3 => incendi_frq_mod@IncendiFoligno=3 (61 objects)
Mean number of attributes in rule = 1.5

Figura 3 - Output testuale generato dal modulo r.roughset

La scelta di rimandare ad una successiva fase la generazione della mappa "*rough*" è dovuta al fatto che in base agli algoritmi di analisi scelti e alla complessità del fenomeno le regole generate possono essere moltissime con richieste di tempi di calcolo elevate. Inoltre, l'utente avanzato può facilmente modificare lo *script* per adattarlo a specifici scopi di analisi. Si pensi, al riguardo, alle analisi statistiche che possono essere generate autonomamente o alla verifica di corrispondenza tra le aree decisionali e i risultati ottenuti dalla classificazione.
L'estensione dall'analisi numerica a quella spaziale, resa possibile dall'implementazione del modulo all'interno di GRASS, apre una serie di possibilità aggiuntive importanti, consentendo una più immediata e efficace comprensione dei risultati, a vantaggio della funzione di supporto alle decisioni.

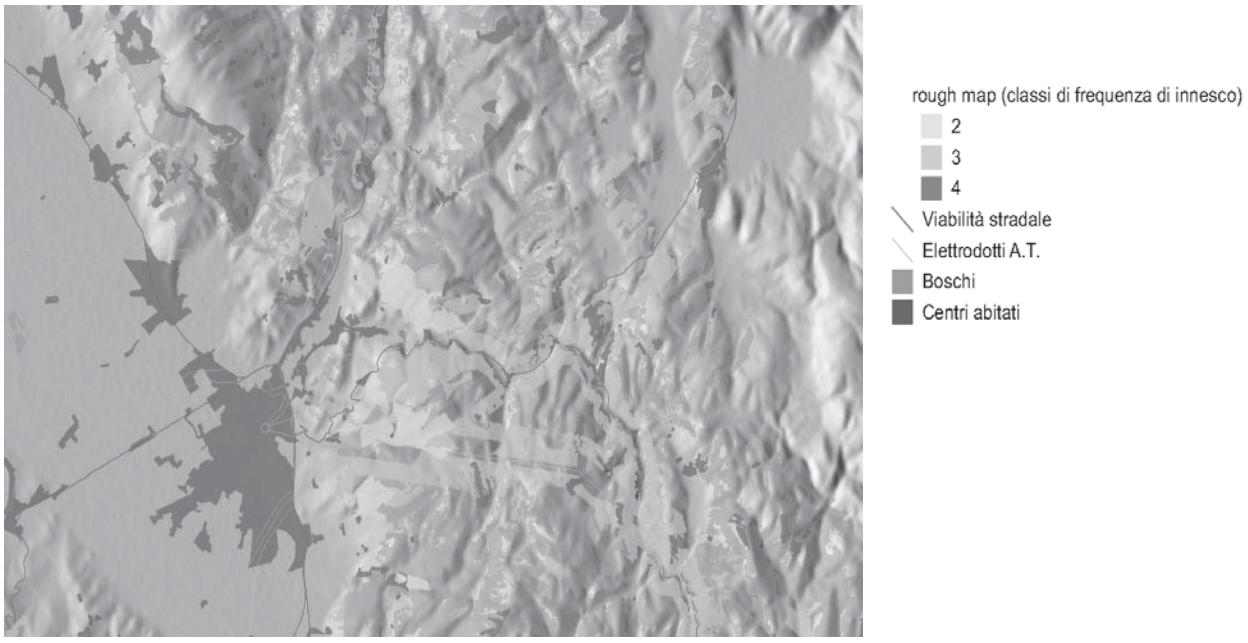

Figura 4 - Output grafico generato sulla base delle regole decisionali ottenute dalla rough set analysis

5. Conclusioni e sviluppi futuri

Il modulo *r.roughset* costituisce il primo passo di un programma di ricerca che ha l'obiettivo finale di integrare la teoria dei Dominance-based Rough Set Approach (DRSA) (Greco, Matarazzo, Slowinski, 2004) in software GIS quali, appunto, GRASS e QGIS. Il passaggio dalla teoria dei *rough sets* tradizionale a quella basata sulla dominanza comporta l'ampliamento del confine metodologico da quello del *data mining* a quello dell'analisi multicriterio, con la conseguente introduzione di un criterio di preferenza e, quindi, di ordinamento, altrimenti non gestibile con i *rough sets* tradizionali.

La fase successiva del lavoro comporterà l'arricchimento delle *Rough Set Library* (RSL) con gli algoritmi di DRSA e, quindi, la realizzazione del relativo modulo di GRASS di interfaccia applicativa.

Parallelamente al lavoro di analisi e sviluppo condotta in linguaggio C, è in fase di debugging l'accesso alle RSL da parte di *Python*, attraverso il modulo *ctypes*. In questo modo sarà possibile sfruttare appieno le potenzialità di un linguaggio moderno e orientato agli oggetti, quale è il *Python*, con la potenza del linguaggio C. Inoltre, moltissimi software GIS, sia Open Source (GRASS, QGIS, SAGA GIS, OpenEV, ecc.) che commerciali, stanno adottando il *Python* come linguaggio interno di *scripting*.

Al termine del percorso di ricerca, si disporrà del modulo *r.roughset* per l'applicazione dei rough sets a database geografici complessi, di un modulo analogo ma basato sulla DRSA (che applica in ambito geografico le tecniche di analisi multicriterio basate sui *rough sets*) e di un modulo in *Python* che consente di applicare tutte le funzioni delle RSL ad ogni ambiente GIS in grado di disporre di un interprete del "linguaggio del pitone".

Bibliografia

AA.VV. (2007), *Il modello di monitoraggio software UmbriaSUIT 1.0* Università degli Studi di Perugia, ARPA Umbria, Perugia.

Bellini A, Guidi A. (1994), *Guida al linguaggio C,* McGraw Hill.

Gawrys M., Sienkiewicz J. (1993), *Rough Set Library user's manual* . Institute of Computer Science, Warsaw University of Technology, Nowowiejska 15/19, 00-665 Warsaw, Poland.

Grass development team (2008), *Programmer's Manual* www.grass.itc.it**.**

Greco, S., Matarazzo, B., Slowinski, R. (1999), *The use of rough sets and fuzzy sets in MCDM***.** Chapter 14 [in]: T.Gal, T.Stewart, T.Hanne (eds.), *Advances in Multiple Criteria Decision Making*. Kluwer, pp. 14.1-14.59.

Greco, S., Matarazzo, B., Slowinski, R. (2004), *Dominance-Based Rough Set Approach to Knowledge Discovery (I) General Perspective*. Chapter 20 [in]: N.Zhong, J.Liu, *Intelligent Technologies for Information Analysis* Springer-Verlag, Berlin.

Larson M.**,** Shapiro M., Tweddale S. (1991), *Performing Map Calculations on GRASS Data: r.mapcalc Program Tutorial - U.S. Army Corps of Engineers* **-** Construction Engineering Research Laboratory - Environmental Division - Spatial Analysis Systems Team.

Markus N., Helena M. (2004), *Open source gis: a GRASS approach*. Kluwer Academic Publishers.

Obersteiner M., Wilk S. (1999), Determinants of Long-Term Economic Development. An Empirical Cross-country Study Involving Rough Sets Theory and Rule Induction, Transition Economics Series No. 11, Institute for Advanced Studies, Vienna.

Pawlak, Z. (1991), *Rough Sets. Theoretical Aspects of Reasoning about Data*. Kluwer Academic Publishers.

Pawlak Z. (1997), "Rough sets approach to knowledge - based decision support", in *European Journal of Operational Research*, n. 99.

Slowinski R. (a cura di) (1992), *Intelligent Decision Support, Handbook of applications and advances of the rough set theory*, Dordrecht, Kluwer.