FUZZY CLUSTERING: LA LOGICA, I METODI

Di Alessandro LA ROCCA

Istat – Direzione centrale per le indagini su condizioni e qualità della vita

INDICE

ABSTRACT

I metodi di classificazione sfocata (fuzzy), sembra non abbiano avuto lo stesso successo degli analoghi che generano veri e propri raggruppamenti (metodi *crisp*). I metodi fuzzy pur non dando luogo a veri e propri raggruppamenti delle unità, nel senso proprio del termine, consentono tuttavia di trattare con efficacia l'imprecisione che spesso è presente nei dati, attribuendo ogni unità a ciascun gruppo con diversi gradi di appartenenza.

L'articolo parte dai due tipi di logica sottostante i due diversi metodi e accenna alla mai sopita polemica sorta tra studiosi di logica fuzzy e studiosi di probabilità; segue una descrizione dei principali algoritmi attraverso cui si giunge a una classificazione sfocata dei dati. Una breve applicazione su alcuni indicatori territoriali conclude il lavoro.

INTRODUZIONE

L'utilizzo delle tecniche di clustering ha registrato negli ultimi anni un crescente interesse sia da parte degli studiosi che si occupano degli aspetti metodologici di queste tecniche, sia di coloro che a vario titolo (Sociologi, Economisti, Statistici applicati) adottano queste tecniche nei loro ambiti di ricerca. A determinarne il successo hanno contribuito l'enorme sviluppo di *package* statistici, che implementano numerosi algoritmi di analisi multivariata dei dati, e la continua evoluzione dell'hardware a cui è demandato il compito di "ospitarli".

L'analisi dei dati spesso coincide con le più note tecniche di clustering: quanto può emergere da una normale ispezione dei dati condotta attraverso normali statistiche descrittive, si preferisce molto spesso evincerlo da tecniche di analisi multidimensionali. Il fenomeno non è chiaramente esente da mode culturali, alimentate il più delle volte da interessi commerciali.

Un esempio di quanto queste tecniche abbiano pervaso molti ambiti applicativi è dato dalla diffusione del data mining, inteso sia come scoperta di conoscenza (*knowledge discovery in databases)*, sia semplicemente come estrazione di dati (*data fishing*, *data dredging)*, in questo ambito le tecniche di clustering sono diventate prevalenti. Il data mining, strumento ormai indispensabile per molte funzioni aziendali (si pensi al marketing), allarga ulteriormente la platea dei suoi utilizzatori, anche fra coloro con competenze non statistiche.

Anche la diffusione dei metodi all'interno della stessa classe non è stata omogenea. Alcuni metodi di cluster analysis, come quelli di classificazione sfocata, non hanno riscosso fra gli utilizzatori lo stesso successo dei metodi di clustering standard così detti *crisp*. Questi non sempre sono adatti a trattare l'imprecisione contenuta nei dati, a differenza dei metodi fuzzy, che benché non offrano classificazioni nel senso proprio del termine, permettono di fornire una lettura semplificata dei dati. La logica duale, dell'inclusione o non inclusione, è preferita alla logica fuzzy, ovvero sfumata. In questo lavoro, una volta definite la logica duale e la logica sfocata, ci si soffermerà sui diversi metodi di fuzzy clustering, nel tentativo di descriverne gli algoritmi e le diverse tipologie di cluster generate. Si concluderà con una breve applicazione di metodi di fuzzy clustering.

CAPITOLO 1

LOGICA DUALE E LOGICA FUZZY

1.1 La logica duale

Gli strumenti logici a disposizione degli studiosi sono il frutto della logica bivalente, ossia aristotelica. Da oltre duemila anni i problemi sono affrontati e risolti con la logica del "Si o No", "Vero o Falso", mentre i problemi a cui la vita reale ci espone hanno poco di bivalente. I principi che sottostanno alla logica bivalente sono: il principio di non contraddizione e il principio del terzo escluso (*tertium non datur)*, il concetto di insieme ci aiuterà ad enunciarli. Il primo afferma che un generico elemento x non può appartenere contemporaneamente ad un insieme A e al suo complemento Ă; il secondo principio afferma che l'operazione di unione tra un insieme e il suo complemento produce l'insieme universo. Quindi un oggetto può o non può appartenere a questo insieme, senza alcuna via di mezzo (insiemi *crisp*).

Ad esempio, nella logica tradizionale (Figura 1.1) una persona può essere considerata adulta, ovvero fa parte dell'insieme "adulti" totalmente, quando supera il diciottesimo anno d'età, altrimenti rientra interamente nell'insieme "giovani".

Figura 1.1

1.2 La logica fuzzy

La pubblicazione dell'articolo dal titolo "Fuzzy sets" di Lofti A. Zadeh del 1965, sancisce la nascita della logica fuzzy (logica sfumata). Sebbene esistessero dei precedenti storici, in quanto tale logica era stata già intuita da Cartesio, Bertrand Russel, Albert Einstein, Werner Heisemberg, all'autore va attribuito il merito di aver riorganizzato e formalizzato il tutto in un articolo originale e destinato a lasciare il segno. Le reazioni a questo lavoro da parte del mondo accademico furono diverse. Alcuni studiosi si dissero entusiasti delle nuove idee, altri accolsero il lavoro con un misto di scetticismo e ostilità. La tradizione cartesiana per ciò che è quantitativo e preciso, e il disprezzo per ciò che è qualitativo e impreciso era troppo radicata per essere superata senza resistenza. Anche se dalle sue origini alcuni filosofi ne avevano messo in luce i punti deboli (si pensi ai famosi paradossi di Zenone).

Lord Kelvin nel 1883 scrisse: "Nelle scienze fisiche un primo essenziale passo nella direzione di apprendere una qualche materia è quello di trovare principi di calcolo numerico e metodi praticabili per misurare alcune qualità ad essa connesse. Spesso affermo che quando puoi misurare quello di cui stai parlando ed esprimerlo in numeri, allora conosci qualcosa di esso; ma se non puoi misurarlo, se non puoi esprimerlo in numeri, la tua conoscenza è di un tipo insoddisfacente: potrebbe essere l'inizio della conoscenza, ma nei tuoi pensieri sei appena approdato allo stato di scienza, di qualunque questione si tratti" (Zadeh, 1990, pp. 95-96).

Il tono della polemica restò all'indomani dell'articolo di Zadeh sempre molto alto, soprattutto nei primi anni di vita della logica fuzzy, quando le applicazioni pratiche erano ancora lontane.

Nel 1974 E.H. Mamdani applicò nell'ambito del controllo la logica fuzzy per un motore a vapore (Mamdani, 1981). Da allora gli studi sulle possibili applicazioni della *fuzziness* sono cresciuti, specialmente in Giappone dove le idee pionieristiche di Zadeh hanno trovato molti sostenitori. Tra le più recenti applicazioni fuzzy, ricordiamo la realizzazione verso la metà degli anni 80 di un sistema di controllo per la metropolitana di Sendai, (Giappone) dove accelerazione e frenatura sono ottenute in modo molto più morbido (a detta dei progettisti) dell'intervento umano. Ancora nel 1985 apparve il primo *digital fuzzy* chip nei laboratori *AT&T Bell*. Gli sviluppi mantengono una caratterizzazione prevalentemente ingegneristica, risentendo del luogo d'origine e si affrancano completamente da speculazioni di ordine filosofico, nonostante l'idea di base risalga al confronto con il linguaggio naturale.

Nella logica fuzzy (Figura 1.2) il passaggio dall'insieme giovani all'insieme adulti è graduale, sfumato sul confine dei due insiemi.

Nella Figura 1.3 è rappresentata la curva degli opposti tra i due insiemi fuzzy.

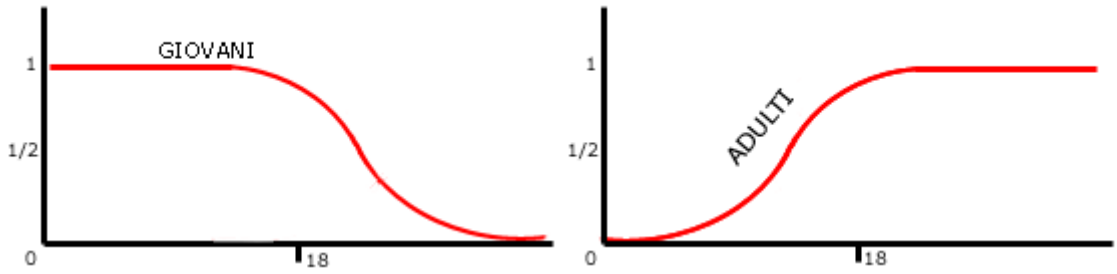

Figura 1.2

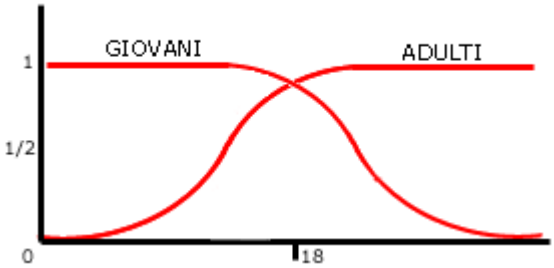

Figura 1.3

1.3 Insiemi fuzzy ed incertezza

Passando dalla logica tradizionale, bivalente, a quella che utilizza gli insiemi fuzzy, sfumata, possiamo definire quest'ultima come un insieme di oggetti dal confine non definito in termini di appartenenza e non appartenenza, in altre parole, la caratteristica distintiva di questo approccio è la ridefinizione del concetto di appartenenza ad un insieme. La definizione fu data la prima volta da Zadeh nel 1965: "Un insieme fuzzy è una classe di oggetti con un continuum di gradi di appartenenza". Tale insieme è caratterizzato da una funzione di appartenenza (caratteristica) che assegna ad ogni oggetto un grado di appartenenza compreso tra 0 e 1. Le nozioni di inclusione, unione, intersezione, complemento, relazione, convessità, ecc. sono estese a tali insiemi, e diverse proprietà di queste nozioni sono stabilite nel contesto degli insiemi fuzzy (Zadeh, 1965, p. 338). Nella logica fuzzy, in altre parole, il concetto di appartenenza è ridefinito in maniera quantitativa, associando ad ogni elemento il grado di appartenenza a quella classe. Formalizzando:

$\mu A : X \rightarrow [0, 1]$

dove μA(x) è interpretato come grado di appartenenza dell'elemento x nell'insieme fuzzy A per ogni x$\in$X.

L'insieme A sarà l'insieme delle coppie:

A={(u, μA(u))|u $\in$ X}.

Si definisce numero fuzzy, un insieme fuzzy normale e convesso tale che esiste al meno un x_0 per cui:

$$\begin{cases} \mu A(x_0) = 1 \\ \mu A = continuo \end{cases}$$

Un insieme fuzzy A si dice normale sul dominio X se la sua funzione di appartenenza assume valore 1 per almeno un elemento x$\in$X; si definisce convesso, se:

$$\forall x, y \in x, \forall \lambda \in [0,1] \Rightarrow \mu a(\lambda x + (1-\lambda)y) \geq \min(\mu a(x), \mu a(y)).$$

Gli insiemi classici, quindi, possono essere visti come un sottoinsieme di quelli fuzzy dal momento che ammettono solo due valori di appartenenza (0 e 1).

1.4 Funzione di appartenenza di un insieme fuzzy e probabilità

I rapporti tra logica fuzzy e teoria della probabilità sono molto controversi e hanno dato luogo a polemiche aspre e spesso non costruttive tra i seguaci di ambedue gli orientamenti. Da una parte, i probabilisti, forti di una tradizione secolare e di una posizione consolidata, hanno tentato di difendere il monopolio storicamente detenuto in materia di casualità ed incertezza, asserendo che la logica sfumata è null'altro che una probabilità sotto mentite spoglie, sostenuti in tale convinzione dalla circostanza, da ritenersi puramente accidentale, che le misure di probabilità, al pari dei gradi d'appartenenza agli insiemi fuzzy, sono espresse da valori numerici inclusi nell'intervallo reale [0,1]. Dall'altra mostrando che anche la teoria probabilistica nelle sue diverse impostazioni (classica, bayesiana soggettivista), è una teoria del caso ancora ancorata ad una logica dicotomica e bivalente. I motivi principali che alimentano la confusione tra i due approcci sono: 1) entrambi tentano di fornire un modello per trattare l'incertezza; 2) la coincidenza dell'intervallo dei valori che possono assumere la funzione di appartenenza in un caso e la probabilità di un evento dall'altro.

La differenza sostanziale tra la logica fuzzy e la teoria della probabilità, è che la prima tratta eventi deterministici, mentre la teoria della probabilità attiene alla verosimiglianza di

eventi stocastici. La teoria della probabilità attiene all'incertezza relativa al risultato di un esperimento casuale, mentre la funzione di appartenenza non è una funzione aleatoria: l'evento si è già verificato ma non si conosce in che misura. Esemplificando, dire che una determinata persona, ad esempio, ha un grado di appartenenza alle persone dotate di intelligenza pari a 0,70 è diverso dall'affermare che questa ha una probabilità di superare un esame in corso pari a 0,70 (tuttavia una volta svolto l'esame, o lo si è superato oppure no e le probabilità collassano a 0 o a 1). Infine, l'incertezza della logica probabilistica muta con l'osservazione mentre quella della logica fuzzy resta immutata, ossia se osserviamo che il tasso di intelligenza è un certo numero, questo non implica che la persona sia del tutto intelligente.

La vaghezza del mondo che ci circonda, la sua polivalenza nonché la vicinanza al modo di ragionare del pensiero umano fanno di questo approccio un ottimo strumento per diversi ambiti disciplinari. Questo tipo di logica fatto di "sfumature" è risultata particolarmente efficace nel campo delle tecnologie dell'informazione, perché in grado di potenziare le capacità dei sistemi esperti tradizionali introducendo flessibilità e robustezza. Ma quando risulta conveniente preferire questo tipo di logica? In generale possiamo dire che dove non c'è incertezza, imprecisione, o dove non si cercano soluzioni deterministiche e elastiche, non ne è consigliata l'adozione.

CAPITOLO 2

CLUSTERING

2.1 Classificazione tipologia e tassonomia

Prima di approfondire i metodi che consentono di classificare insiemi di oggetti, è meglio ridurre l'eccesso di termini utilizzati quali sinonimi per designare il risultato di un clustering. In questo contesto si è deciso di seguire Marradi nella definizione che egli ha dato di classificazione, tipologia e tassonomia per l'enciclopedia delle scienze sociali (Marradi, 1993, pp. 22-30).

Data una classe di oggetti o eventi, è possibile effettuare su di essi tre diverse operazioni che generano tre distinte classificazioni:

1. Riduzione dell'estensione di un concetto in due o più entità con un minor livello di generalità.
2. Operazioni con cui oggetti ed eventi sono raggruppati in sottoinsiemi a seconda delle similarità percepite su uno o più stati di proprietà (Zoologia, Botanica).
3. Operazioni attraverso le quali un oggetto è assegnato a una classe gia costituita attraverso operazioni di tipo 1 e 2.

Poiché le operazioni di cui al terzo punto implicano quelle di cui ai punti precedenti ci occupiamo di queste altre. Le prime sono "intensionali", perché il loro fine è quello di definire l'intensione delle classi, cioè definire queste ultime come concetti e denominarle in modo appropriato. Può essere necessario in questa operazione esaminare le distribuzioni di una o più proprietà esibite dagli oggetti. Si costruiscono gruppi di oggetti o eventi con apposite procedure, preoccupandosi in un secondo momento di trovare un concetto con relativo termine per ogni combinazione di oggetti o eventi per le proprietà che definiscono un gruppo. Le classificazioni intensionali sono a volte denominate categorizzazioni o divisioni.

Le seconde presuppongono matrici dei dati, ovvero, vettori di oggetti, vettori di proprietà. Si costruiscono gruppi di oggetti o eventi con apposite procedure, preoccupandosi in un secondo momento di trovare un concetto con relativo termine per ogni combinazione di oggetti o eventi per le proprietà che definiscono un gruppo. Il risultato di questo tipo di operazioni prende il nome di classificazione estensionale.

L'aspetto del concetto intensionale che viene articolato per formare le classi, viene chiamato *fundamentum divisionis*. Parleremo di tassonomia in presenza di più *fundamenta divisionis*, se le divisioni sull'estensione del concetto sono fatte in successione su concetti di generalità decrescente; di tipologia, se le divisioni sull'estensione del concetto sono fatte simultaneamente su più concetti. Le tipologie sono molto più frequenti nella classificazione estensionale, le tassonomie nella classificazione intensionale.

Citiamo un esempio di classificazione intensionale riportato dall'autore: per un sistema politico un aspetto della sua intensione (una delle proprietà dei sistemi politici) è la legittimazione dei governanti, suo *fundamentum divisionis:* le classi possono essere sistema politico teocratico, autocratico, aristocratico, plutocratico, democratico, ecc.

La classificazione estensionale è il risultato di operazioni che raggruppano gli oggetti o eventi di un insieme in due o più sottoinsiemi in modo da massimizzare la somiglianza (negli stati su una serie di proprietà considerate) fra membri dello stesso sottoinsieme e la dissomiglianza fra membri di sottoinsiemi diversi. Come accade spesso trattandosi di più proprietà parleremo di tipi e non di classi. Nella classificazione estensionale la registrazione degli stati degli oggetti su più proprietà, matrice dei dati, è elemento essenziale di tutto il processo. Altre denominazioni correnti sono: classificazione, tassonomia numerica, e cluster analysis. Quest'ultima, a detta dell'autore, dovrebbe designare solo una parte delle tecniche che generano una classificazione estensionale. Le classi devono possedere la proprietà dell'esaustività, e della mutua esclusività, requisiti che molto spesso alcune classificazioni disattendono.

La sistematizzazione operata da Marradi ha il pregio di fornire i fondamenti logici e semantici della classificazione, molto spesso disattesi dall'uso indiscriminato delle tecniche di classificazione automatica. La classificazione denominata estensionale è quella a cui faremo riferimento nel nostro lavoro. I concetti di proprietà, oggetti, eventi, usati correntemente dall'autore nella definizione possono essere tradotti rispettivamente come variabili e unità statistiche. Infatti, in questo tipo di classificazione si introduce il concetto di matrice dei dati, utilizzata per registrare stati su più proprietà di più oggetti o eventi. Tra le denominazioni usuali di questa classificazione si riporta anche quella di cluster analysis.

La classificazione estensionale a rigore produrrebbe tipologie, e non classi: infatti, molto spesso nella scuola statistica francese di analisi dei dati, si usa il concetto (termine) di analisi tipologica.

2.2 Obiettivi e scopi della classificazione

I metodi di classificazione, (clustering) sono finalizzati a classificare le unità statistiche attraverso procedure che in genere presuppongono che su ogni unità siano state rilevate più variabili. I metodi erano già noti alla fine del XIX secolo, ma l'interesse da parte degli statistici si è avuto attorno agli anni 60.

Ad oggi, grazie soprattutto alle potenzialità del software e dell'hardware a disposizione si contano una enorme varietà di algoritmi, sempre più efficienti e con diversi gradi di difficoltà computazionale. I campi di applicazione sono numerosi, dalle scienze fisiche (fisica, medicina, biologia), a quelle sociali (economia, sociologia, psicologia).

Un clustering può essere definito come una partizione di un insieme di unità elementari in modo che la suddivisione risultante goda di alcune proprietà considerate desiderabili; oppure, come raggruppamento di unità molto simili tra loro in gruppi che abbiano la caratteristica di essere il più possibile distinti tra loro. Si tratta di un insieme di metodi di apprendimento non supervisionato.

2.2.1 Metodi di clustering

I fattori che caratterizzano una tecnica di clustering sono:

1. le misure di similarità;
2. l'algoritmo di individuazione dei cluster.

Attraverso il criterio riportato nel punto 2 (tipo di algoritmo) possiamo distinguere i metodi gerarchici da quelli non gerarchici.

Quelli gerarchici producono raggruppamenti successivi ordinabili in base a valori crescenti della distanza, tali che ogni gruppo è incluso in gruppo di ampiezza superiore fino al cluster che contiene tutte le unità considerate.

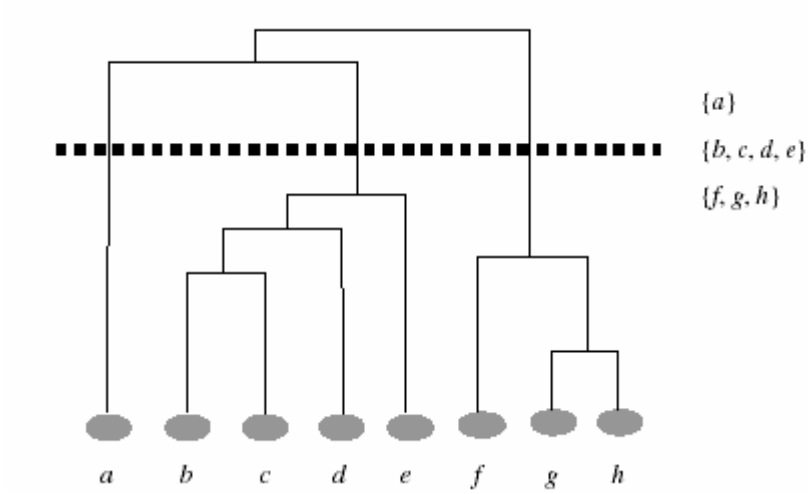

Figura 2.1

La rappresentazione riportata nella Figura. 2.1 denominata dendrogramma, illustra il processo di aggregazione gerarchico dove, a diversi livelli di distanza si aggregano le unità differenti.

All'interno degli algoritmi gerarchici possiamo distinguere quelli aggregativi da quelli scissori: i primi partono da *n* elementi distinti e determinano un numero decrescente di gruppi di ampiezza crescente, fino ad associare in un unico cluster tutte le unità di partenza. I secondi partono da una unica partizione che contiene tutte le unità di partenza, e ripartiscono le stesse in gruppi sempre più piccoli, finché il numero di cluster viene a coincidere con il numero di unità. Di questi non ci occuperemo vista la loro scarsa diffusione nella statistica applicata.

I metodi non gerarchici, invece, producono una unica partizione delle *n* unità iniziali in g gruppi prefissati, partizione considerata ottimale rispetto ad una specifica funzione obiettivo. Sono incluse anche le classificazioni prodotte dai metodi di programmazione matematica e quelle che, partendo da una partizione iniziale provvisoria, tentano di migliorare la suddivisione delle unità attraverso una serie di riallocazioni iterative che terminano quando viene soddisfatto un determinato criterio di ottimalità.

I metodi di suddivisione iterativa eseguono una suddivisione effettiva delle unità attraverso l'individuazione di nuclei (centroidi, semi ecc.) intorno ai quali raggruppare le unità e procedere iterativamente. I metodi di programmazione matematica si basano su spostamenti virtuali delle unità fatti secondo la soluzione di un minimo o massimo vincolato.

Per concludere questa breve introduzione ai metodi di clustering elenchiamo i più diffusi algoritmi geararchici e non gerarchici (Tabella 2.1), senza entrare nel merito delle caratteristiche che li contraddistinguono e li rendono applicabili a determinate strutture di dati. Il dettaglio sul

funzionamento di alcuni di questi algoritmi sarà comunque affrontato successivamente, ma in un ottica fuzzy clustering.

Tabella 2.1 – Principali algoritmi di clustering classico.

Cluster gerarchico	Cluster non gerarchico
Metodo del legame singolo	*Metodo delle K medie*
Metodo del legame completo	
Metodo del legame medio	
Metodo del centroide	
Metodo di Ward	

2.2.2 Sfocature, ricoprimenti e partizioni

Nel paragrafo precedente, i metodi di clustering sono stati suddivisi in base al tipo di algoritmo utilizzato; questo criterio, molto frequente in letteratura, non esclude però che si possano distinguere rispetto al risultato prodotto. Se si ipotizza che per ogni unità venga fornita una funzione di appartenenza, che indichi in che misura una determinata unità appartiene a diversi gruppi, è possibile ottenere la seguente suddivisione dei metodi di clustering (Tabella 2.2).

Tabella 2.2 – Classificazione dei metodi di clustering rispetto ai risultati prodotti.

	Grado di appartenenza / **Vincoli**	**Metodi classici**	**Metodi sfocati**
		{0,1}	[0,1]
Partizioni	$\sum_{I=1}^{M} U_{i,j} = 1$	*Clustering classico*	*Clustering sfocato*
Ricoprimenti	$\sum_{I=1}^{M} U_{i,j} \geq 1$	*Clustering sovrapposto*	*Clustering sfocato sovrapposto*

Si definisce $U_{i,j}$ una funzione a j valori (j è il numero di gruppi della partizione o del ricoprimento) che associa ad ogni unità i, j numeri ognuno dei quali indica il grado di appartenenza dell'unità i-esima al j-gruppo; sia inoltre M il numero noto di cluster. La matrice dei gradi di appartenenza $\mathbf{U}$ di dimensione $N*M$ avrà elemento generico $\left[\mathbf{u}_{i,j}\right]$, che rappresenta il valore numerico della funzione di appartenenza dell'elemento i-esimo al cluster j.

I metodi di clustering, in altre parole, non sempre sono ben definiti come accade nel caso del clustering classico, ma si possono avere classi dai confini indefiniti che si sovrappongono, oppure casi in cui gli oggetti per loro natura non possono essere classificati univocamente e

presentano valori di appartenenza sfocati. Nei metodi di classificazione classica rientrano tutti quelli che forniscono delle partizioni che producono cluster disgiunti, la cui unione corrisponde alla totalità delle unità. Ma anche i metodi di classificazione sovrapposta, che danno suddivisioni delle unità in cluster non disgiunti: una stessa unità può appartenere a più cluster.

I metodi di classificazione sfocata assegnano le unità ai cluster in modo che una unità appartiene per una parte ad un cluster e per la parte restante agli altri cluster. I metodi di classificazione sovrapposta sfocata sono la combinazione dei due metodi precedenti: questi producono dei ricoprimenti sfocati, cioè dei cluster sfocati sovrapposti. In base alla definizione di Marradi, i metodi che generano sfocature o ricoprimenti non consentono quindi di parlare di classificazione giacché non è rispettata la proprietà della mutua esclusività.

.

CAPITOLO 3

FUZZY CLUSTERING

3.1 Introduzione

Nel nostro lavoro vengono descritti solo metodi fuzzy che usano dati non sfocati, tralasciando altri metodi che danno una iniziale sfocatura ai dati attraverso una funzione caratteristica che assegna ad ogni unità una misura della quantità di carattere posseduta, rispetto a quella posseduta da altre unità. I dati, quindi, nel nostro caso sono quelli realmente osservati e il problema è assegnare le unità ad un certo numero di cluster, in modo tale che ognuna di loro appartenga ad un cluster con un diverso grado di appartenenza espresso da una funzione di appartenenza che assume valori nell'intervallo [0,1]. Tra questi metodi, come già detto si distinguono quelli che producono clustering sfocati, da quelli che producono ricoprimenti (*clump*) a seconda dei vincoli imposti alla funzione di appartenenza.

I metodi di fuzzy clustering sono stati poco sviluppati, rispetto a quelli classici detti *crisp* o di *hard* clustering, gli algoritmi a disposizione, infatti, risultano abbastanza ridotti. Questi metodi fanno ricorso alla teoria degli insiemi fuzzy, e permettono di associare una unità ai gruppi con un certo grado di appartenenza. L'interesse per questi metodi nasce dalla consapevolezza che esiste una certo grado di imprecisione nei dati e quindi che un metodo siffatto è in grado di rappresentarli più di quanto possa fare un metodo *crisp*.

Nella Figura 3.1 è rappresentata una situazione ideale: i punti sono perfettamente separati in due cluster, e una situazione più vicina alla realtà: i punti si distribuiscono in modo tale che risulta difficile attribuire un punto ad un cluster oppure ad un altro.

I metodi di fuzzy clustering sono, inoltre, più ricchi di informazioni in quanto forniscono il grado di coerenza di una unità con ciascun cluster, consentendo di stabilire una gerarchia di gruppi (la gerarchia è data dal diverso grado di appartenenza dell'unità ai gruppi) a cui può appartenere l'unità, in virtù del fatto che i gruppi sono visti come insiemi fuzzy. Inoltre, non hanno alcuna pretesa di fornire risposte precise su come si aggregano i dati. Anche per questi metodi è possibile una suddivisione tra metodi gerarchici e non gerarchici, i primi sono spesso delle estensioni dei metodi classici al caso fuzzy.

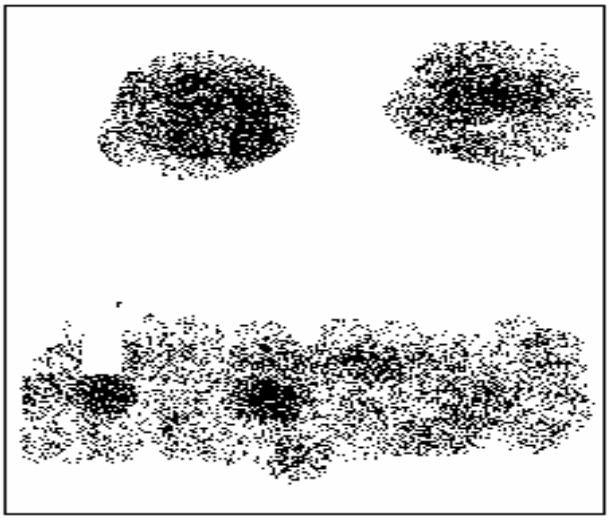

Figura 3.1

3.2 Fuzzy clustering gerarchico

I metodi gerarchici di fuzzy clustering hanno la caratteristica di prevedere due fasi: nella prima si calcola una misura di similarità tra coppie di unità, nella seconda si assegna ciascuna unità ai gruppi formati con un certo grado di appartenenza. La prima fase è analoga alla classificazione *crisp*, la differenza sta nel modo in cui vengono attribuite le funzioni di appartenenza ai cluster. Restano, come nel caso della classificazione classica, le differenze nel modo in cui attribuire le distanze tra una unità e un gruppo, o tra due gruppi, oltre al modo in cui attribuire il grado di appartenenza di una unità ad un cluster sfocato. Rispetto all'attribuzione delle distanze è possibile trovare delle similitudini con i metodi, ad esempio, del legame singolo e del legame completo. I metodi descritti in seguito possono essere modificati per renderli più adatti ai casi reali che volta per volta si presentano.

3.2.1 Metodo della sintesi di più partizioni (Zani, 1989)

Il metodo parte da P partizioni delle unità e termina con una classificazione sfocata di esse. Si suppone che per le N unità siano state rilevati caratteri qualitativi e quantitativi. La similarità tra una coppia di unità è determinata attraverso un indice di similarità su tutti i caratteri, dato dalla frequenza relativa delle partizioni, una per ogni carattere, in cui le due unità si trovano incluse nello stesso gruppo. Agli indici di similarità ottenuti si applica una procedura di classificazione che genera una successione gerarchica di partizioni, che ad ogni livello di

similarità considerato produce una partizione sfocata delle unità, attraverso l'assegnazione ad ogni gruppo di una unità funzione di appartenenza, sotto il vincolo che la somma dei gradi di appartenenza per ciascuna unità sia uguale a 1. Per i caratteri qualitativi la scelta delle partizioni iniziali è abbastanza naturale in quanto le stesse modalità comportano per ogni carattere la partizione migliore. Per i caratteri quantitativi le cui modalità non possono che essere classi è inevitabile un elemento di soggettività dato dal criterio scelto per la suddivisione in classi: quartili, minima varianza ecc.

Dati *M* caratteri misti su *N* unità, e C_K gruppi della partizione indotti dal k-esimo carattere, il grado di appartenenza congiunto tra due unità a_i e a_j è dato da:

$$Z_{ij} = \frac{1}{M}\sum_{k=1}^{M}\delta_K(i,j) \qquad [3.1]$$

dove δ_K vale 1 se a_i e a_j sono nello stesso gruppo nella partizione k-esima, 0 altrimenti.

Questo indice di similarità assume valori nell'intervallo [0,1], il suo valore massimo coincide con la piena appartenenza delle due unità allo stesso gruppo in ciascuna delle *M* partizioni. Costruita la matrice delle similarità si procede alla determinazione della classificazione sfocata, il grado di appartenenza di una unità a ciascun cluster sfocato può essere attribuito attraverso il criterio del minimo delle similarità tra l'unità e ciascuna delle unità incluse nel gruppo (analogamente al metodo del legame completo nella classificazione *crisp*), oppure, del massimo delle similarità tra l'unità e ciascuna delle unità incluse nel gruppo (analogamente al metodo del legame singolo nella classificazione *crisp*).

La procedura per ottenere la classificazione si articola in tre fasi:

1. Individuazione dei nuclei, dove per nucleo si intendono le coppie di unità con similarità uguale a 1;
2. individuazione delle coppie di unità con similarità pari a: $\alpha = \frac{M-1}{M}$;
3. si itera il punto 2 per valori decrescenti: $\frac{M-2}{M}$, $\frac{M-3}{M}$…….;

al punto due possono verificarsi tre casi:

1. Le coppie di unità con similarità α non erano ancora incluse in nessun gruppo, in questo caso formeranno un nuovo gruppo con grado di appartenenza α;
2. una delle due unità è già inclusa in un gruppo, in questo caso anche l'altra unità verrà inclusa nello stesso gruppo con grado α;

3. entrambe le unità sono state classificate in gruppi diversi, si assegna ciascuna delle unità all'altro gruppo con grado di appartenenza α, purché la somma dei gradi di appartenenza di ciascuna unità sia uguale a 1. Nel caso contrario si attribuisce il valore massimo che soddisfa tale vincolo, qualora fosse 0 l'unità non può più essere assegnata a nessun gruppo.

L'algoritmo appena descritto genera una successione gerarchica di cluster sfocati in corrispondenza di livelli decrescenti di similarità.

3.2.2 Metodo dei ricoprimenti sfocati

Anche in questo contesto si parte da una matrice delle similarità, l'indice di similarità scelto sarà quello che si adatta al minimo livello di misura delle variabili a disposizione, evitando in questo modo di utilizzare diverse misure di similarità a seconda del tipo di carattere.

In una prima fase viene calcolata la matrice delle similarità, successivamente vengono definiti i cluster e attribuite le misure di appartenenza delle unità ai gruppi. Ricordiamo che la caratteristica di questo tipo di clustering è che la somma delle funzioni di appartenenza delle unità ai cluster può essere maggiore o uguale a 1. Supponiamo che siano state osservate *N* unità statistiche su cui sono stati rilevati *M* caratteri. I casi che si possono verificare sono i seguenti:

1. I caratteri sono quantitativi, per ogni carattere si possono calcolare le distanze relative e farne il complemento a 1:

$$V_{i,j}(K) = 1 - \frac{d_{i,j}(K)}{\max_{i,j} d_{i,j}(K)} \qquad [3.2]$$

dove $d_{i,j}(K)$ è la distanza di Hamming generalizzata tra l'i-esima e la j-esima unità, relativamente al k-esimo carattere. L'indice di similarità complessivo relativo alle unità i e j è dato da:

$$S_{i,j} = \sum_{k=1}^{m} V_{i,j}(K) * p(K) \qquad [3.3]$$

dove $p(K)$ è il peso, non negativo, assegnato al *k*-esimo carattere sotto il vincolo:

$\sum_{1}^{m} p(K) = 1$, la [3.3] è nota come indice di Gower;

2. i caratteri sono ordinali, si calcola la distanza relativa come nel caso (1) tenendo presente però che si tratta di una distanza tra posizioni in una graduatoria;

3. i caratteri sono sconnessi, non essendo possibile calcolare la distanza di Hamming generalizzata tra le unità, si adotta la [3.1] proposta da Zani nella sintesi di più partizioni, e si procede nella individuazione dei cluster e della funzione di appartenenza;

Alla matrice di similarità **S** ottenuta si applica una procedura di classificazione analoga al legame completo e cioè:

a. Si ricercano nella matrice di similarità **S** le coppie di unità con similarità pari a 1;
b. si individuano le coppie di unità con indice di similarità massimo, ma minore di 1, supponiamo α, si formeranno, quindi, tanti gruppi quanti sono le coppie con questa similarità, le unità incluse nei cluster avranno grado di appartenenza α, l'appartenenza degli altri elementi ai cluster è determinata dal minimo delle similarità tra l'unità e ciascuna delle unità incluse nel gruppo;
c. analogamente alla fase precedente, si ricerca nella matrice **S** il livello di similarità massimo non ancora considerato, supponiamo β in corrispondenza della coppia $\mathbf{S_{h,l}}$ che costituiranno un gruppo, a meno che una delle due unità, ad esempio *h*, sia inclusa già in un gruppo e l'altra unità *l* presenti con le unità di questo gruppo similarità non inferiori a β, in questo caso anche *l* viene assegnato al gruppo a cui appartiene *h*.
d. si itera il passo (c) per valori decrescenti dell'indice di similarità, fino a considerare tutti i casi.

Gli aspetti distintivi del metodo dei ricoprimenti sfocati, rispetto ai metodi che producono cluster *crisp* sono: 1) una unità già inclusa in gruppo, può anche appartenere ad un altro gruppo a condizione che esistano le condizioni relative al livello di similarità con le unità di quel gruppo, 2) il grado di appartenenza di una unità ad un gruppo è dato dalla similarità minima tra questa unità e le restanti del gruppo.

3.2.3 Metodo del legame medio sfocato

L'indice di similarità utilizzato in questo metodo per la costruzione della matrice di similarità è quello usato nel metodo dei ricoprimenti sfocati [3.2], si tratta quindi di un metodo adatto per caratteri quantitativi. La classificazione avviene attraverso i seguenti passi:

1. Si ricercano nella matrice di similarità le coppie di unità con similarità massima, questi gruppi rappresentano i nuclei iniziali, si procede con il calcolo dei centroidi dei gruppi. Il grado di appartenenza di ciascuna unità al gruppo è dato in proporzione inversa alla

distanza di tale unità dal centroide del gruppo (più sarà distante, minore sarà il grado di appartenenza), con il vincolo che la somma dei gradi di appartenenza sia uguale a 1;

2. si ricerca il livello di similarità più alto tra quelli non ancora considerati. Al livello di similarità, supponiamo α, se una delle unità, ad esempio i, della coppia (i, j) è gia inclusa in un gruppo allora j sarà inserita nel medesimo gruppo solo se la similarità media tra j e tutte le unità che compongono il gruppo è maggiore di α. In caso contrario la coppia (i, j) formerà un nuovo gruppo;
3. si itera il punto (2) fino a quando tutte le unità formano un unico gruppo.

3.3 Fuzzy clustering non gerarchico

I metodi di classificazione non gerarchici hanno la caratteristica di fornire direttamente un determinato numero di gruppi fissato a priori, attraverso procedure di tipo iterativo che cercano di ottimizzare una funzione obiettivo. I diversi algoritmi si differenziano a seconda della funzione obiettivo adottata e quindi della diversa procedura iterativa scelta per calcolare i gradi di appartenenza delle unità ai gruppi. La funzione obiettivo determina per ogni soluzione una misura dell'errore, in termini di efficienza o costo basandosi sulla distanza tra i dati e gli elementi rappresentativi dei cluster. Ogni formulazione della funzione obiettivo incorpora dei vincoli: la soluzione ottima corrisponde al valore ottimo della funzione. Si tratta, quindi, di un problema di ottimizzazione. Il metodo delle k medie, fuzzy k means, rappresenta sicuramente il metodo più noto.

Esistono anche altri metodi di fuzzy clustering non gerarchico basati su relazioni fuzzy che danno origine a matrici di similarità fuzzy, di cui però non si discuterà in questo lavoro.

3.3.1 Metodo delle K medie sfocato

Questo metodo (Bezdek, 1981) è quello più utilizzato e più diffuso tra quelli di classificazione sfocata, si tratta di una generalizzazione del metodo *crisp* delle k medie, particolarmente adatto a trattare matrici di dati di notevoli dimensioni, grazie alla velocità con la quale giunge ad una classificazione ottimale (convergenza). L'algoritmo procede nel modo seguente: si sceglie il numero di cluster C in cui si vogliono classificare le N unità con P caratteri; si sceglie un raggruppamento iniziale o in modo ragionato, o, come più spesso accade,

casualmente. Da questo si procede iterativamente minimizzando una funzione obiettivo. Il grado di appartenenza delle *N* unità ai *C* gruppi soddisfa i seguenti vincoli:

$$0 \le u_{i,k} \le 1$$

$$\sum_{1}^{c} u_{i,k} = 1$$

per l'i-esima unità ed il k-esimo gruppo. Il primo vincolo stabilisce l'insieme di definizione della funzione di appartenenza, il secondo vincola la somma dei gradi di appartenenza di ogni unità ad 1. La matrice **U** contenente i gradi di appartenenza $u_{i,k}$ avrà dimensioni *N*C*. Indichiamo con O_m la funzione obiettivo utilizzata; questa è funzione del quadrato della distanza $d_{i,k}$ tra l'unità i-esima e il centroide del k-esimo gruppo e dipende dal parametro *m* che può assumere qualsiasi valore reale maggiore o uguale a 1.

$$O_m(U,v) = \sum_{k=1}^{c}\sum_{i=1}^{n}\left(u_{i,k}\right)^m * \left(d_{i,k}\right)^2; \qquad [3.4]$$

dove: $\left(d_{i,k}\right)^2 = \left|x_i - v_k\right|$ e $\left|.....\right|$ è una opportuna norma su R^P ad esempio la norma euclidea;

$v_k \in R^p$ è la componente k-esima del vettore dei centroidi $v = \left(v_1,....v_c\right) \in_{np} R^{cp}$;

$x_i \in R^p$ è la componente i-esima del vettore delle unità $x = \left(x_1,....x_n\right) \in R^{np}$;

$m \in \left[1,\infty\right)$.

Le variabili rispetto alle quali effettuare la minimizzazione sono quindi: i centri dei cluster e i gradi di appartenenza. Il significato della funzione obiettivo è che il centroide di ogni gruppo è la migliore rappresentazione delle unità che lo compongono, in quanto rende minima la somma dei quadrati degli errori $\left|x_i - v_k\right|$. La funzione obiettivo O_m è una misura dell'errore quadratico in cui si incorre quando si rappresentano le *N* unità con in *C* centroidi dei gruppi, essa dipende da come le unità sono disposte nei gruppi. La partizione ottima è quella che minimizza O_m. Il parametro *m* esprime il grado di sfocatura (*fuzziness*), cioè quanto è sfocata la partizione risultante. L'algoritmo può essere sintetizzato nei seguenti passi:

1. Si fissano i valori di *m* e di *c* e si sceglie una partizione iniziale delle unità in *c* gruppi rappresentata con la matrice $U^0 = \left[u_{i,k}\right]$ dove l'esponente indica il numero di iterazioni;

2. si calcolano i centri dei gruppi $V_k^{\,0}$ secondo la formula $V_k^{\,0} = \dfrac{\sum_{i=1}^{n} (u_{i,k})^m * x_i}{\sum_{i=1}^{n} (u_{i,k})^m}$ [3.5];
3. si calcola la matrice U^1 alla prima iterazione;
4. si calcola la differenza tra l'ultima e la penultima iterazione secondo una opportuna distanza, se $\left|U^1 - U^0\right| < \delta$ dove δ è un parametro stabilito a priori ci si ferma e si considera come classificazione finale quella dell'ultima iterazione, altrimenti si itera il passo 2 fino a quando la [3.5] non è soddisfatta.

Nel passo 3 possono verificarsi le seguenti condizioni:

1. Se per qualche gruppo (supponiamo *r*) si ha che $d_{i,r} = 0$ si pone $u_{i,r} = 1$ e $u_{i,k} = 0$ per tutti i $k \neq r$;
2. se la condizione precedente non è soddisfatta allora si applica la seguente formula:

$$u_{i,k} = \frac{1}{\sum_{j=1}^{c} \left[\frac{d_{i,k}}{d_{j,k}} \right]^{\frac{2}{m-1}}}.$$

Solo un cenno alle maggiori questioni su cui si sono concentrati alcuni studiosi relativamente a questo algoritmo.

Studi sulla convergenza della funzione obiettivo [3.4] (Bezdek, Hathaway, 1988), hanno dimostrato che la soluzione raggiunta non è sempre punto di minimo globale, in quanto potrebbe essere punto di minimo locale o di flesso.

Anche per quanto riguarda la scelta della partizione iniziale (casuale o ragionata) si rileva (Bezdek, Hathaway, 1988) che nella maggior parte dei casi la soluzione ottenuta dipende dalla partizione iniziale; inoltre se la partizione iniziale non è "buona" aumentano i costi computazionali per il raggiungimento della convergenza . A questo riguardo, sono state proposte soluzioni per l'inizializzazione efficiente dell'algoritmo, come ad esempio: il metodo a due stadi di Zahid ed il *multi stage random sampling fuzzy c-means* di Cheng ed altri.

Per la scelta del valore da attribuire al parametro *m,* che rappresenta il livello di sfocatura, (nel caso assuma valore 1 si ricade nel clustering *crisp*), non esistono basi teoriche a cui affidarsi; infatti, il valore 2 che convenzionalmente si attribuisce è frutto di studi empirici.

3.3.2 Estensioni del metodo delle K medie sfocato

Selim e Ismail (Selim, Ismail, 1984), hanno introdotto alcune variazioni all'algoritmo delle k medie sfocato, giungendo a metodi che forniscono classificazioni semisfocate. Infatti, l'eccessiva sfocatura delle unità all'aumentare del numero dei gruppi conduce spesso a problemi di interpretazione della classificazione finale. Accenniamo solo brevemente alle variazioni imposte dai tre metodi all'algoritmo di base per giungere alla semisfocatura.

Il primo metodo prevede un vincolo scelto all'inizio sul numero massimo di cluster.

Il seconda impone di non attribuire alcun grado di appartenenza a quella unità la cui distanza dal centro del gruppo sia superiore ad un valore prefissato.

Il terzo assegna un valore soglia di accettazione per i gradi di appartenenza delle unità ai gruppi.

Kamel e Selim (Kamel, Selim, 1991) cercano successivamente di migliorare il terzo metodo, prevedendo un valore soglia appropriato del grado di appartenenza delle unità ai gruppi, al di sotto del quale l'unità "smette" di appartenere a quel gruppo. Il metodo denominato *Thresholded Fuzzy C-Means*, (TFCM) prevede di determinare il valore limite della soglia al fine di evitare che la alcune unità non appartengano ad alcun gruppo, o che alcuni gruppi restino vuoti.

CAPITOLO 4

UNA APPLICAZIONE CON R

4.1 R un ambiente statistico open source

R è un ambiente statistico disponibile gratuitamente e scaricabile dal sito *http://www.r-project.org*. Si tratta come si dice in gergo di un "dialetto" del noto pacchetto commerciale S-plus con cui mantiene una certa somiglianza.

Il fatto che sia disponibile gratuitamente non è chiaramente l'unico aspetto che ne fa uno strumento sempre più utilizzato dagli statistici: infatti, l'enorme varietà di rappresentazioni grafiche a disposizione dell'utente, e la varietà di librerie aggiuntive costituiscono aspetti di non poca importanza.

Il software ha raggiunto ormai svariate *release* (la prima nel 1996) ed i moduli aggiuntivi (scaricabili separatamente dal sito), toccano ormai i più disparati ambiti della statistica applicata: dall'analisi delle serie storiche, alla ricerca operativa, alla analisi multivariata dei dati, solo per citare i più noti.

Altro aspetto di non secondaria importanza è la sua leggerezza in termini di impegno hardware, infatti, l'utente può caricare i "contributi" che gli occorrono senza che questi risiedano permanentemente in memoria ogni qual volta si avvia R.

In questo lavoro la scelta di un software open source come R è stata dettata dall'esigenza di disporre di un programma che contenesse una procedura di clustering sfocato. I programmi più diffusi in ambito statistico (ad esempio Sas) non offrono la disponibilità di procedure che implementino metodi di fuzzy clustering, a meno che non si acquistino i moduli aggiuntivi che sono ormai piuttosto costosi (come ad esempio *enterprise data miner*). Non si esclude che si possano sviluppare autonomamente, anche in Sas ad esempio, algoritmi di fuzzy clustering: il tempo richiesto e le competenze possono però molte volte far propendere per altre soluzioni, più semplici.

4.1.1 Il modulo Cluster

Il modulo Cluster permette di utilizzare i metodi più diffusi di clustering, (gerarchici e non gerarchici) con il semplice utilizzo di funzioni già predisposte.

Oltre ai metodi *crisp,* questa libreria permette di usare un metodo di fuzzy clustering, non gerarchico. L'output di queste procedure può essere arricchito con efficaci rappresentazioni grafiche che visualizzano i più importanti risultati ottenuti.

La funzione che permette di avere una classificazione fuzzy è "fanny" e richiede come parametri la matrice dei dati (casi per variabili) o di dissimilarità, il numero di cluster scelti, il parametro *m* che determina il grado di sfocatura (l'esponente della funzione obiettivo), cui per ragioni empiriche si assegna il valore 2, il numero di iterazioni e la tolleranza. L'algoritmo è molto simile a quello delle k-medie sfocato, già descritto nei suoi aspetti teorici nel paragrafo 3.3.1. Le differenze con il metodo delle k-medie si trovano nella funzione obiettivo adottata: le distanze non sono al quadrato come nel metodo delle k medie, ed il valore del parametro *m* è fissato al valore 2 e non ad un qualsiasi valore maggiore o uguale a 1.

Ne presentiamo una applicazione su alcuni indicatori territoriali relativi alle venti regioni italiane, a scopo puramente esemplificativo.

4.1.2 Procedura fanny: applicazione su alcuni indicatori delle regioni italiane

Gli indicatori territoriali relativi alle regioni italiane (Tabella 4.1), sono desunti dall'indagine sulle forze di lavoro condotta dall'Istat.

Tabella 4.1 – Indicatori territoriali delle regioni italiane – Anno 2003 valori percentuali.

Regioni	T_att.	T_occ.	Occ_agr.	Occ_ind.	Occ_ser.	Don_occ.	T_dis.
Piemonte	51,70	49,21	3,83	37,53	58,65	41,70	4,80
Valle d'Aosta	55,32	53,08	4,70	23,38	71,92	42,30	4,06
Lombardia	53,51	51,59	2,11	40,29	57,60	40,49	3,60
Trentino A .A.	56,09	54,72	8,06	27,21	64,73	40,72	2,44
Bolzano	59,88	58,66	11,79	25,36	62,85	41,39	2,03
Trento	52,37	50,85	3,83	29,31	66,86	39,96	2,90
Veneto	53,18	51,36	4,01	41,32	54,67	39,53	3,41
Friuli V. G.	50,18	48,21	3,18	33,13	63,69	41,59	3,93
Liguria	46,25	43,47	3,46	22,00	74,54	40,65	6,02
Emilia R.	54,03	52,38	5,01	35,96	59,04	43,48	3,06
Toscana	50,10	47,75	3,68	32,23	64,09	40,99	4,70

Regioni	T_att.	T_occ.	Occ_agr.	Occ_ind.	Occ_ser.	Don_occ.	T_dis.
Umbria	47,48	45,03	4,67	32,90	62,44	40,28	5,17
Marche	50,90	48,97	3,85	40,17	55,97	41,97	3,78
Lazio	49,47	45,16	2,62	19,78	77,60	38,10	8,71
Abruzzo	46,15	43,67	5,85	30,83	63,32	37,47	5,37
Molise	44,60	39,12	9,24	29,54	61,23	34,34	12,28
Campania	44,40	35,43	6,35	24,70	68,95	29,38	20,20
Puglia	42,84	36,94	10,17	26,75	63,07	29,71	13,78
Basilicata	43,22	36,27	10,27	33,65	56,08	32,24	16,06
Calabria	44,81	34,32	12,79	19,87	67,34	31,81	23,42
Sicilia	42,52	33,96	8,32	20,93	70,75	29,07	20,13
Sardegna	47,04	39,10	8,04	23,97	67,99	34,39	16,88

L'anno di riferimento è il 2003 e gli indicatori sono: il tasso di attività (T_att.), il tasso di occupazione (T_occ.), il tasso di disoccupazione (T_dis.), gli occupati nell'agricoltura (Occ_agr.), nell'industria (Occ_ind.), nei servizi (Occ_ser.) per ogni 100 occupati e le donne occupate ogni 100 occupati (Don_occ.). Abbiamo scelto di classificare le nostre 20 unità territoriali (le regioni) in due gruppi.

La procedura fanny fornisce assieme ai gradi di appartenenza, altre importanti informazioni tra cui il numero di iterazioni necessarie a far convergere l'algoritmo, l'indice di sfocatura di Dunn, ed il cluster classico (*crisp*) più vicino alle unità.

Tabella 4.2 – Principali risultati della procedura fanny.

Regioni	Cluster 1	Cluster 2	Cluster crisp più vicino
Piemonte	0,73	0,26	1
Valle d'Aosta	0,59	0,40	1
Lombardia	0,68	0,31	1
Trentino A .A.	0,55	0,44	1
Veneto	0,65	0,34	1
Friuli V. G.	0,67	0,32	1
Liguria	0,45	0,54	2
Emilia R.	0,65	0,34	1
Toscana	0,65	0,34	1
Umbria	0,59	0,40	1
Marche	0,71	0,28	1
Lazio	0,47	0,52	2
Abruzzo	0,44	0,55	2
Molise	0,35	0,64	2
Campania	0,28	0,71	2
Puglia	0,34	0,65	2
Basilicata	0,40	0,59	2
Calabria	0,32	0,67	2
Sicilia	0,32	0,67	2
Sardegna	0,31	0,68	2

Di seguito si riportano le rappresentazioni delle ripartizioni territoriali sulle prime due componenti e della struttura dei due gruppi (Figura 4.1 e 4.2).

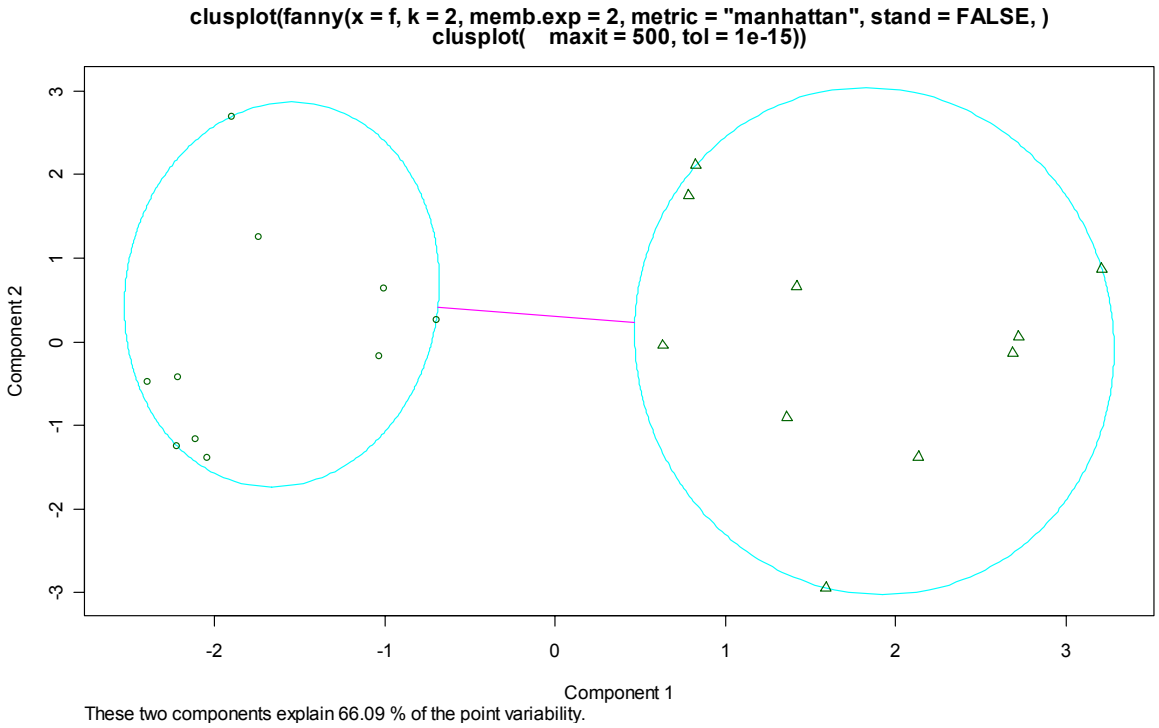

Figura 4.1

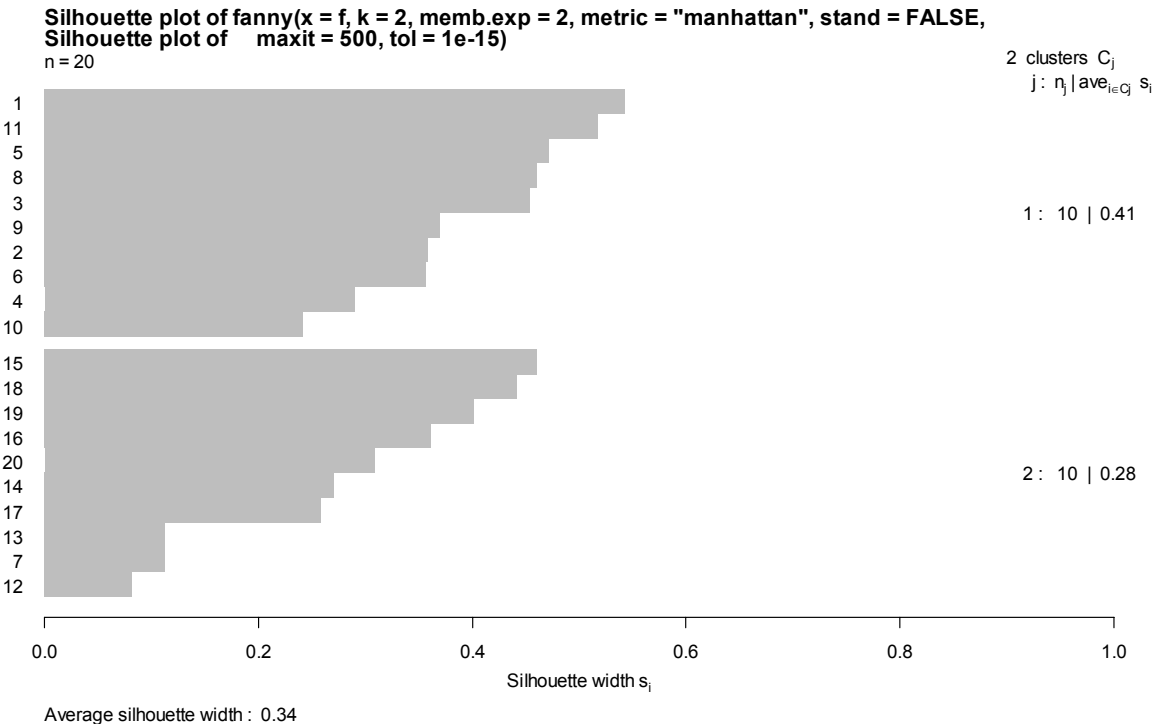

Figura 4.2

Dall'esame dei gradi di appartenenza delle unità, emerge abbastanza chiaramente una distinzione tra regioni del nord e regioni del sud, i gradi di appartenenza ai gruppi sono distanti dal caso di equa appartenenza.

Tuttavia, alcune regioni presentano livelli di sfocatura maggiori: in altre parole hanno caratteristiche economiche tali da non poter essere attribuite in modo preponderante all'uno o all'altro gruppo, ed il grado di appartenenza ripartito quasi al 50 per cento. Queste sono: la Liguria, e in parte anche il Lazio e l'Abruzzo.

I gruppi hanno una struttura debole, il primo pari a 0,41, il secondo pari a 0,28 (Figura 4.2). I valori di riferimento dell'ampiezza sono: (0,71-1,00) struttura forte, (0,51-0,70) struttura plausibile, (0,26-0,50) struttura debole, (<0,26) struttura assente.

Il grado di sfocatura misurato dall'indice di Dunn è pari 0,09, questo indice può assumere valori tra 0 e 1: rispettivamente massimo grado di sfocatura nei gruppi, assenza di sfocatura nei gruppi (clustering *crisp*).

CONCLUSIONI

I metodi di fuzzy clustering fin qui descritti hanno indubbiamente motivo di essere inclusi nel novero delle tecniche multivariate di cui il ricercatore si dovrebbe avvalere nel corso delle proprie indagini, siano esse esplorative o confermative. A differenza dei metodi classici questi hanno il vantaggio di far emergere quelle unità statistiche particolarmente problematiche in una operazione di classificazione, perché non perfettamente includibili in un gruppo, data la loro specificità reale, o in quanto affette da possibili errori di misurazione. I dati desunti da rilevazioni dirette, o da fonte amministrativa possono nascondere problemi di accuratezza, di imprecisione, anche una volta espletate le normali procedure di controllo e correzione dei dati. Oppure, alcune unità pur non essendo totalmente *outlier*, lo sono solo in parte, rispetto ad alcuni caratteri: in questi casi, un metodo sfocato si adatta sicuramente meglio di quanto possa fare un metodo *crisp*.

Queste tecniche non vanno però preferite contrapponendole a quelle così dette classiche, ma "giocate" assieme ad altre, prendendo da ognuna quanto di meglio offre in uno specifico contesto di ricerca.

I metodi sfocati, tuttavia, presentano anch'essi dei limiti non trascurabili, ad esempio: l'elevato grado di sfocatura che aumenta al crescere del numero dei gruppi, crea difficoltà in fase di interpretazione della classificazione finale. Per i metodi non gerarchici il raggiungimento di un ottimo locale e non globale della funzione obiettivo, l'assenza di una base teorica per la scelta del parametro m.

Problema non secondario è la validità dei cluster ottenuti (*cluster validity*); oltre a metodi prettamente empirici (confrontare i risultati di più prove ripetute), esistono in letteratura una varietà di indicatori la cui efficacia è però legata a condizioni quali una buona separazione tra i gruppi, o una determinata forma degli stessi.

Da un punto di vista applicativo, questi metodi risentono della limitata o nulla disponibilità nei pacchetti statistici: all'utente che abbia le competenze necessarie non resta che svilupparli in proprio.

BIBLIOGRAFIA

Bezdek J. C., Hathaway R. J. (1988), "Recent convergence results for the fuzzy cmeans clustering algorithms", in *J. of Classification*, 5, pp.237-247.

Cerbara L., Iacovacci G. (1998), *Tecniche sfocate per la classificazione di dati di popolazione*, Working paper, 2, Irp Cnr, Roma.

Everitt B. S. (1980), *Cluster analysis*, London, Heineman Educational Books.

Kamel M.S., Selim S. Z. (1991), "A thresholded fuzzy c-means algorithm for semifuzzy clustering", in *Pattern Recognition*, 24(9), pp. 825-883.

Kosko B. (1993), *Fuzzy thinking. The new science of fuzzy logic, tr., it., Il fuzzy pensiero. Teoria e applicazione della logica fuzzy,* Baldini & Castoldi, Milano, 1995.

Mamdani E.H., Gaines B.R. (1981), *Fuzzy Reasoning and its Applications*, Academic Press, New York.

Marradi A. (1993), *Classificazioni, Tipologie, Tassonomie*, Enciclopedia delle scienze sociali, 2, Roma, pp. 22-30.

Martin M. (2005), Cluster Analysis Extended Rousseeuw et al., http://www.r-project.org/.

Ricolfi L. (1992), *HELGA - Nuovi principi di analisi dei gruppi*, Franco Angeli, Milano.

Rizzi A. (1985), *Analisi dei dati. Applicazioni dell'informatica alla statistica*, Studi Superiori NIS.

Selim S. Z., Ismail M. A. (1984), "Soft clustering of multidimensional data: a semi-fuzzy approach", in *Pattern Recognition*, 17(5), pp. 559-568.

Zadeh L. A. (1990) "The Birth and Evolution of Fuzzy Logic", *International Journal of General Systems"* 17, pp. 95-96.

Zadeh L. A. (1965), "Fuzzy sets", *Inf. Control*, 8, pp. 338-353.

Zadeh L. A. (1977), "Fuzzy sets and their application to pattern classification and clustering", in Van Ryzin J., *Classification and clustering*, Accademic press, New York.

Zani S. (1988), Un metodo di classificazione "sfocata", in G. Diana, C. Provasi e R. Vedaldi, *Metodi statistici per la tecnologia e l'analisi dei dati multidimensionali,* Università degli Studi di Padova, pp. 281-288.

Zani S. (2000), *Analisi dei dati statistici*, volume 2, Giuffrè Milano, pp. 253-256.