Riconoscimento automatico di forme in archeologia: il caso della necropoli di Sala Consilina

C. Davino*, R. Di Martino*, F. Mola**, D. Vistocco*

*Dip.to di Matematica e Statistica, Università degli Studi di Napoli - Federico II
**Dipartimento di Economia, Università di Cagliari

1. Introduzione

Scopo della presente comunicazione è proporre un approccio integrato per il riconoscimento di forme in archeologia basato sul simultaneo utilizzo di uno strumento tipico dell'Intelligenza Artificiale, le *reti neuronali* (Bishop, 1995) e di alcune metodologie di Analisi Multidimensionale dei Dati, quali *l'Analisi delle Corrispondenze Multiple*, la *Cluster Analysis* e la *Segmentazione Binaria a due stadi* (Mola e Siciliano, 1992).

In particolare, la procedura di analisi è finalizzata allo studio di una necropoli (II età del ferro) di Sala Consilina che si inserisce in un lavoro di topografia storica del Vallo di Diano (Salerno). Tale studio mira a definire le caratteristiche delle comunità indigene che occupano questo comprensorio della Lucania interna. I dati analizzati si riferiscono a 173 tombe delle quali sono state rilevate una serie di caratteristiche relative alla loro struttura e alla composizione dei corredi funerari.

La procedura di analisi della necropoli di Sala Consilina proposta è articolata in tre fasi, ciascuna delle quali risponde ad uno specifico obiettivo.

La prima fase, che segue ovviamente un'opportuna ricodifica e sistemazione dei dati rilevati, si propone di estrarre da tali dati, tramite un approccio esplorativo, un contenuto informativo superiore a quello puramente descrittivo al fine di individuare la struttura di relazioni presenti all'interno della necropoli. Il ricorso all'Analisi delle Corrispondenze Multiple che, come è noto, si propone di studiare la struttura del fenomeno in esame mediante l'individuazione di sottospazi di migliore approssimazione su cui proiettare individui e variabili, ha permesso una prima analisi strutturale della necropoli.

La seconda fase privilegia in maniera più specifica gli individui attraverso la ricerca di una tipologia degli stessi in gruppi con caratteristiche simili. Tale tipologia è individuata mediante il ricorso alla *Cluster Analysis* realizzata sui risultati dell'analisi fattoriale precedente, cioè sfruttando le relazioni tra modalità individuate precedentemente.

La terza fase della strategia di analisi riguarda l'individuazione di una regola per classificare le tombe. Lo stato di conservazione della necropoli era tale da non permettere all'archeologo che ha curato lo scavo (Maiello, 1995) un immediato riconoscimento della tipologia di tombe rinvenute. Le informazioni fornite dall'esperto in materia hanno consentito di sintetizzare in un' unica variabile due caratteristiche dei defunti: il genere e l'età. Il 65% delle tombe della necropoli sono state, quindi, classificate come appartenenti a *maschi/adulti*, a *femmine/adulte* o a *bambini*. L'obiettivo della terza fase della strategia è il riconoscimento automatico della tipologia di appartenenza del restante gruppo di tombe (pari al 35% della necropoli) relativamente alle quali la classificazione a priori non aveva fornito alcuna informazione.

A tale proposito sono state utilizzate le reti neuronali, modelli computazionali capaci di apprendere "dall'esperienza" utilizzando un insieme di esempi significativi per acquisire esperienza generalizzabile anche a situazioni nuove. Come è noto infatti, le reti neuronali supervisionate sono particolarmente abili nel risolvere problemi di classificazione ossia nel riconoscere la classe di appartenenza di ciascuna unità di osservazione a partire da una serie di caratteristiche osservate.

Il vantaggio offerto dalle reti neuronali in termini di buona performance non può prescindere dall'aspetto legato all'interpretazione dei risultati ottenuti e del processo che ha generato tali risultati. Nel nostro contesto, non è importante solo il riconoscimento delle tombe ma anche il meccanismo che ha condotto a tale classificazione e la descrizione delle caratteristiche delle tipologie di tombe individuate. Successivamente un confronto dei risultati forniti dalle reti neuronali con quelli forniti da un metodo di discriminazione non parametrica, quale la Segmentazione Binaria, consente di analizzare i legami tra statistica e reti neuronali evidenziando le possibili sinergie che derivano da un utilizzo congiunto delle due metodologie (Davino et al., 1997b) nel riconoscimento di forme in archeologia.

2. La descrizione della necropoli di Sala Consilina

Il materiale e le informazioni raccolte per ciascuna tomba sono state catalogate in due diverse schede, una relativa ai dati anagrafici di ogni tomba, l'altra alla descrizione dei corredi.
L'analisi descrittiva dei dati rilevati ha condotto all'eliminazione di alcune variabili scarsamente rappresentate e alla ricodifica di altre precedentemente definite.

La tabella seguente schematizza le 9 variabili utilizzate nella strategia di analisi.

Tabella 1: Le informazioni sulla necropoli di Sala Consilina

Variabile	N.ro modalità	Modalità
Conservazione	3	1) frammentaria 2) integra 3) non definita
Lunghezza	3	1) < 120cm 2) 121-160cm 3) > 160cm
Profondita`	3	1) < 80cm 2) 81-160cm 3) > 160cm
Genere/età	4	1) maschio 2) femmina 3) bambino 4) non definito
N.ro oggetti	3	1) < 9ogg. 2) 9-16ogg. 3) > 17ogg.
Note	3	1) devastata 2) legata a tomba 4) completa
Materia	2	1) argilla 2) metalli
Forma	5	1) armi 2) ornamenti 3) ogg.vari 4) ogg.non identificati 5) ceramica
Produzione	2	1) Sala Consilina 2) altra produzione

La variabile *materiale* è stata ricodificata in una variabile dicotomica (*argilla*, *metallo*) che assegna a ciascuna tomba la modalità argilla o metallo a seconda se il numero di oggetti presenti nella stessa siano realizzati in prevalenza in argilla o in metallo.

I diversi *tipi di oggetti* sono stati, quindi, sintetizzati nelle seguenti cinque categorie: *armi* (coltelli - giavellotti o lance - altre armi), *ornamenti* (anelli - bracciali - collane - fibule - altri ornamenti), *oggetti vari* (spiedi, fusaiole, rocchetti, tintinnabulum, oggetti vari), *oggetti non identificati*, *ceramica* (ceramica tipo greco, ceramica indigena forme aperte e chiuse, ceramica d'uso). Ciascuna tomba è stata poi assegnata ad una di tali categorie sulla base dell'identico criterio di prevalenza utilizzato per la variabile "materiale".

Le informazioni relative al genere e all'età dell'individuo sono state sintetizzate, come è stato detto, in un'unica variabile definita *genere/età* con 4 possibili modalità: i *maschi adulti*, le *femmine adulte*, i *bambini* e gli *individui non classificabili* dal punto di vista del genere. Tutte le fasi di strategia dell'analisi hanno confermato l'importanza di questa variabile, importanza rivelata intuitivamente dall'archeologo stesso.

3. Individuazione di una struttura: l'Analisi delle Corrispondenze Multiple.

Obiettivo dell'Analisi delle Corrispondenze Multiple è spiegare la struttura delle associazioni tra i diversi caratteri rilevati sulle tombe in funzione di variabili latenti o fattori sottostanti ai dati. Infatti, il problema principale che si presenta quando si dispone di molte caratteristiche relative ad un certo insieme di individui è quello di sintetizzare la maggior parte dell'informazione in esse contenuta in maniera chiara e quanto più rapida possibile. Tale obiettivo viene raggiunto tramite un approccio geometrico rappresentando *j* punti (dove *j* rappresenta la somma delle modalità - nel nostro caso 29 - di tutte le caratteristiche relative alle tombe analizzate) nello spazio *n-dimensionale* delle tombe o alternativamente delle 173 tombe nello *spazio delle modalità*. Per interpretare la nube che si viene a creare si cerca di proiettarla in uno spazio di dimensioni ridotte come la retta o il piano, cercando di deformare il meno possibile le distanze tra punti in modo da minimizzare la perdita di informazione.

L'Analisi delle Corrispondenze Multiple ha individuato i fattori latenti la cui percentuale di variabilità spiegata, definita dai corrispondenti autovalori più significativi, è riportata nella tabella 2. Poiché sono state considerate ben 19 modalità attive, la variabilità totale è ripartita su un numero di fattori eccessivo per cui la quota di variabilità spiegata dai primi fattori risulta molto pessimistica. Un indice che esprime la percentuale di variabilità spiegata da ciascun autovalore a_i tenendo conto di tale aspetto, è dato dal rapporto:

$$Q_i^* = \frac{a_i^2}{\sum_i^t a_i^2}$$ (*i=1,2,.....,m;* con m pari al numero delle modalità).

Secondo tale formula si possono ottenere percentuali di variabilità più realistiche. Nel nostro caso si sono presi in considerazione i primi due fattori che spiegano circa il 40% dell'informazione strutturale contenuta nei dati di partenza mentre quelli successivi individuano solo aspetti marginali.

Tabella 2: Gli autovalori e la % di variabilità spiegata

Num	autovalore	% var	% var. corretta	% var. corretta cumulata
1	0.2806	13.29	25.04	25.04
2	0.2182	10.34	15.14	40.17
3	0.1827	8.65	10.61	50.79
4	0.1595	7.55	8.09	58.88
5	0.1466	6.94	6.83	65.71
6	0.135	6.40	5.79	71.51
7	0.133	6.30	5.62	77.13

La prima considerazione che emerge dalla lettura della posizione delle modalità[1] sul primo piano fattoriale (Fig. 1) è che il primo asse vede la contrapposizione netta di quelle caratteristiche tipiche degli adulti (maschi e femmine) a quelle tipiche dei bambini. Il secondo asse vede invece la contrapposizione delle caratteristiche degli adulti maschi con quelle delle femmine adulte; tale contrapposizione mette in evidenza la relativa facilità di individuare differenze tra tombe di maschi adulti dalle altre tombe e la difficoltà di discriminare in modo netto le tombe dei bambini. La modalità *genere non definito*, è posizionata in prossimità dell'origine a conferma della presenza di tombe di diversa tipologia in questo gruppo.

L'osservazione del primo piano fattoriale, inoltre, sembra mostrare una serie di relazioni che consentono di descrivere le tombe distinguendole in base al genere: gli oggetti delle tombe degli adulti sono stati prodotti per la maggior parte a Sala Consilina, mentre quelli delle tombe dei bambini e delle tombe non definite hanno avuto una produzione piuttosto varia, le dimensioni delle tombe sono nettamente distinte in quanto procedendo dal basso verso l'alto si osservano tombe via via sempre meno lunghe e profonde. La caratterizzazione completa delle tombe, infine, viene realizzata attraverso le variabili relative alla forma degli oggetti e al materiale che li costituisce, infatti le tombe dei maschi sono caratterizzate dalla prevalenza di armi e di oggetti di ceramica realizzati in gran parte con l'argilla, mentre le tombe delle femmine sono caratterizzate da ornamenti in metallo; le tombe dei bambini rivestono ancora un ruolo un pò ibrido in quanto non si evidenzia la prevalenza di un tipo particolare di oggetti ed inoltre esse risultano molto vicine alla modalità *legata a tomba* che lascia presupporre l'eventuale vicinanza con la tomba di un genitore e alla modalità *devastata*.

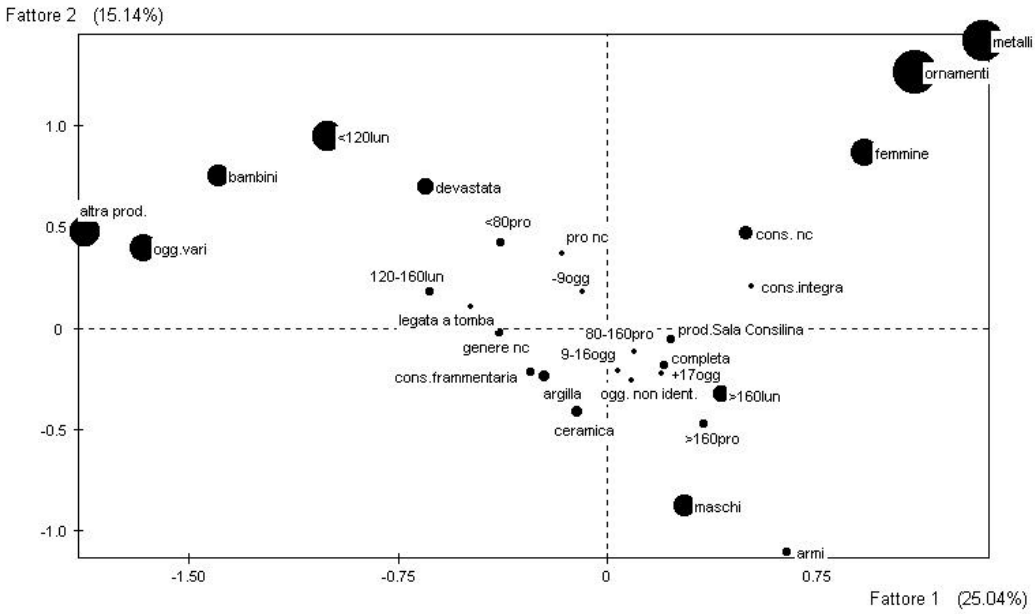


Figura 1: Il primo piano fattoriale

[1] La dimensione dei punti che rappresentano le modalità attive è proporzionale al contributo assoluto delle stesse; tale contributo consente di valutare l'incidenza di ciascuna modalità sull'orientamento dell'asse.

4. Individuazione di una tipologia: la Cluster Analysis

Lo scenario definito dall'osservazione del primo piano fattoriale individuato dall'analisi multidimensionale può essere completato con la ricerca di tipologie omogenee all'interno dell'insieme osservato, ossia di gruppi quanto più possibili omogenei al loro interno e diversi l'uno dall'altro. Tali gruppi vengono ottenuti attraverso le tecniche di analisi dei gruppi (*cluster analysis* o classificazione automatica) che, a differenza dei metodi fattoriali che analizzano le relazioni tra le variabili, operano una riduzione nello spazio degli individui. Nella presente indagine la cluster analysis è stata effettuata sui risultati dell'A.C.M. precedentemente presentati. In particolare, si è fatto riferimento ai primi due fattori che erano stati ritenuti soddisfacenti a sintetizzare la maggior parte dell'informazione contenuta nei dati originari ed è stato utilizzato, come criterio per raggruppare gli individui, il metodo di "Ward" che si propone di minimizzare la variabilità all'interno di ciascuna classe, cioè di giungere alla massima omogeneità interna alle singole classi e a diversificare il più possibile tra loro queste ultime.

In questa fase dell'analisi sono state analizzate in maniera più specifica le singole tombe, cercando di passare da queste a gruppi di tombe caratterizzate da similarità di comportamento al fine di individuare una tipologia di tombe.

L'albero di classificazione ottenuto è riportato nella figura seguente:



Figura 2: L'albero di classificazione

L'analisi del dendrogramma e lo studio degli indici relativi al livello di aggregazione, suggerivano di "tagliare" l'albero in due punti diversi, fornendo due diverse partizioni, rispettivamente in due e tre classi; il carattere esplorativo dello studio ci ha indotto quindi ad analizzare entrambe le partizioni.

I gruppi di ciascuna partizione individuata possono essere definiti confrontando, per ciascuna modalità di risposta, le frequenze percentuali osservate nel gruppo stesso con le frequenze percentuali totali: quanto più i valori sono differenti tanto più quella modalità caratterizzerà quel gruppo.

Una misura di quanto una modalità caratterizzi un gruppo può essere rappresentata da un valore-test che assume valori prossimi allo zero quando la modalità non è significativa e valori assoluti crescenti all'aumentare della significatività. Come regola empirica, è possibile porre il valore 2 come valore soglia e ritenere una modalità tanto più significativa per un gruppo quanto più il valore-test è, in valore assoluto, superiore a 2. Nella colonna CLA/MOD è riportata la percentuale di individui che, tra tutti quelli che hanno indicato una data modalità di risposta, appartengono alla classe di riferimento. Nella colonna MOD/CLA è riportata la percentuale di individui del campione che ha indicato una determinata modalità di risposta. Nella colonna GLOBAL è indicata la percentuale di individui del campione che ha indicato una determinata modalità di risposta.

La tabella 3 mostra la descrizione di ciascuna classe della prima partizione analizzata attraverso le variabili che la caratterizzano, che al tempo stesso consente la definizione della seguente possibile tipologia:

- *I maschi*: tombe caratterizzate prevalentemente da ceramiche realizzate in argilla, provenienti da altre zone del comprensorio lucano. Significativa anche la presenza di oggetti vari non meglio identificati a conferma di uno stato di conservazione della tomba piuttosto frammentario. Le caratteristiche relative a tale classe si presentano sia nelle tombe definite come appartenenti a maschi adulti che a bambini.
- *Le femmine*: tombe con caratteristiche tipicamente femminili, ossia con prevalenza di ornamenti in metallo di produzione Sala Consilina.

Tabella 3: Descrizione della partizione in due classi

Classe 1/2 81.50%						
v.test	Prob	CLA/MOD	MOD/CLA	GLOBAL	Modalità	variabile
8.89	0.000	93.92	98.58	85.55	argilla	Materia
7.66	0.000	99.07	75.18	61.85	Ceramica	Forma
3.30	0.000	89.29	70.92	64.74	Cons.frammentaria	conservazione
2.93	0.002	94.44	36.17	31.21	Maschi	genere/età
2.23	0.013	100.00	14.18	11.56	bambini	genere/età
2.23	0.013	100.00	14.18	11.56	ogg.vari	forma
2.14	0.016	100.00	13.48	10.98	altra produz.	produzione
Classe 2/2 18.50%						
v.test	prob	CLA/MOD	MOD/CLA	GLOBAL	Modalità	variabile
10.88	0.000	86.11	96.88	20.81	Ornamenti	Forma
8.89	0.000	92.00	71.88	14.45	Metalli	Materia
6.63	0.000	58.97	71.88	22.54	Femmine	Genere/età
2.14	0.016	20.78	100.00	89.02	Prod. Sala Consilina	Produzione

La tabella 4 mostra la descrizione della partizione in tre classi che consente di far emergere ulteriori caratteristiche delle tombe e di discriminare le tombe che, nella partizione a due classi, confluivano nella prima grossa classe. La nuova tipologia presenta quindi le seguenti caratteristiche:

- *I maschi adulti*: tombe molto lunghe, con presenza di ceramiche di produzione Sala Consilina e armi. Le caratteristiche di tale gruppo di tombe sono ben definite grazie ad un buono stato di conservazione delle stesse e alla completezza dei corredi funerai in esse presenti.
- *I non classificabili:* tombe di dimensioni medio-piccole, con presenza di oggetti in argilla di provenienza varia. La devastazione degli oggetti trovati e lo stato di conservazione della tomba conferma il carattere indefinito delle tombe classificate come *genere nc*. Tali caratteristiche di frammentarietà sono tipiche anche delle tombe dei bambini che spesso si presentano legate ad altra tomba.
- *Le femmine*: tombe con caratteristiche tipicamente femminili, ossia con prevalenza di ornamenti in metallo di produzione Sala Consilina.

Tabella 4: Descrizione della partizione in tre classi

Classe 1/3 41.04%						
v.test	Prob	CLA/MOD	MOD/CLA	GLOBAL	Modalità	Variabile
8.33	0.000	60.17	100.00	68.21	>160	Lunghezza
6.52	0.000	77.78	59.15	31.21	maschi	Genere/età
5.46	0.000	57.01	85.92	61.85	ceramica	Forma
5.46	0.000	52.31	95.77	75.14	completa	Note
4.10	0.000	46.10	100.00	89.02	Sala Consilina	Produzione
3.67	0.000	46.62	97.18	85.55	Argilla	Materia
3.22	0.001	100.00	11.27	4.62	Armi	forma
Classe 2/3 40.46%						
v.test	Prob	CLA/MOD	MOD/CLA	GLOBAL	Modalità	Variabile
6.77	0.000	89.19	47.14	21.39	<120	Lunghezza
5.87	0.000	100.00	28.57	11.56	ogg. Vari	Forma
5.87	0.000	100.00	28.57	11.56	bambino	Genere/età
5.69	0.000	100.00	27.14	10.98	altra prod.	Produzione
4.85	0.000	47.30	100.00	85.55	Argilla	Materia
4.80	0.000	78.79	37.14	19.08	Devastata	Note
4.23	0.000	88.89	22.86	10.40	120-160	Lunghezza
3.64	0.000	60.00	51.43	34.68	genere nc	Genere/età
3.03	0.001	62.50	35.71	23.12	<80	Profondità
2.69	0.004	48.21	77.14	64.74	frammentaria	Conservazione
2.28	0.011	80.00	11.43	5.78	legata a tomba	Note
Classe 3/3 18.50%						
v.test	Prob	CLA/MOD	MOD/CLA	GLOBAL	Modalità	Variabile
10.88	0.000	86.11	96.88	20.81	Ornamenti	Forma
8.89	0.000	92.00	71.88	14.45	Metalli	Materia
6.63	0.000	58.97	71.88	22.54	Femmine	Genere/età
2.14	0.016	20.78	100.00	89.02	Sala Consilina	prooduzione

5. Individuazione di una regola di classificazione

La terza fase della strategia di analisi si propone di creare una regola di classificazione, che consenta di riconoscere dalle caratteristiche delle tombe rinvenute il genere/età del defunto, ossia se si tratti di tombe appartenute ad un maschio/adulto, ad una femmina/adulta o ad un bambino. La regola di classificazione viene creata utilizzando le informazioni relative alle 113 tombe (65% delle tombe rinvenute), di cui è nota la classificazione a priori della variabile genere/età. I risultati ottenuti sono poi utilizzati a scopo previsivo per classificare il restante gruppo.

5.1 Le Reti Neuronali

Le reti neuronali (Bishop, 1995) sono state sviluppate per studiare ed approfondire l'attività delle reti neuronali biologiche al fine di riuscire a creare un "uomo artificiale". Con il tempo le reti neuronali hanno continuato a destare un forte interesse presso gli studiosi di diverse discipline, non come cloni biologici dei neuroni naturali, ma come modelli matematici finalizzati ad imitare principalmente le enormi potenzialità del cervello umano quali, ad esempio, il parallelismo nell'elaborazione dell'informazione e la flessibilità e capacità di apprendere dall'esperienza.

Percorrendo il quadro storico della nascita e sviluppo della tecnologia neuronale si nota un forte grado di interdisciplinarietà: ai primi testi di biologi e psicologi sono seguiti numerosi contributi di ingegneri, matematici, fisici, statistici. In realtà l'approccio verso tale tecnologia è stato inizialmente di tipo "ingegneristico", molto attento quindi agli aspetti informatici e computazionali. In particolare, il contesto, in cui le reti neuronali hanno avuto origine ed in cui trovano principale campo di applicazione, è quello caratterizzato da applicazioni di tipo *data intensive*, ossia con molti più dati che calcoli, in cui i dati stessi sono spesso incompleti o affetti da rumore e non registrati su archivi, ma continuamente aggiornati, e per i quali i risultati ottimali sono ottenuti utilizzando macchine con dispositivo di calcolo parallelo.

Recentemente le reti neuronali stanno trovando applicazione in molti altri campi per la risoluzione di problemi tipicamente affrontati con metodologie statistiche. In tali contesti, ai problemi più specificatamente legati alla costruzione, taratura ed utilizzo dei modelli neuronali (scelta dell'architettura, inizializzazione dei pesi, definizione della forma funzionale del modello) si affianca la forte sensibilità delle reti neuronali ad alcuni problemi tipici delle applicazioni statistiche tradizionali, quali il trattamento delle variabili, la numerosità del campione e la rappresentatività dello stesso. Sia gli uni che gli altri influiscono in maniera determinante sulla capacità di generalizzazione di una rete ovvero la capacità di classificare dati mai visti in precedenza.

Nell'analisi della necropoli di Sala Consilina, quindi, le reti neuronali sono state utilizzate non come semplice strumento computazionale, ma come momento di una più generale strategia di analisi basata su un utilizzo statistico di tali strumenti dell'Intelligenza Artificiale.

1. In primo luogo, le informazioni disponibili sono state sottoposte ad una fase di *pre-processing*. Troppo spesso, infatti, le reti neuronali vengono utilizzate in maniera inefficiente, ossia i dati vengono brutalmente inseriti nella rete senza considerare alcuna trasformazione a priori degli stessi. Le 9 variabili prese in considerazione (8 esplicative ed 1 dipendente), tutte qualitative, danno luogo ad una configurazione iniziale della rete neuronale con 24 e 3 neuroni, rispettivamente di input e di output. In realtà una modalità per ciascuna variabile è stata eliminata in quanto essa risulta univocamente determinata dalle altre per cui il numero di neuroni considerato è pari a 16 per i neuroni di input e 2 per quelli di output. Una volta definita la rappresentazione delle variabili all'interno della rete, tali variabili sono state sottoposte ad una operazione di normalizzazione secondo la seguente formula:

$$ValNorm = ValMin + \left(\frac{Val - Min}{Max - Min} (ValMax - ValMin) \right)$$

   dove *ValMin* e *ValMax* sono il limite inferiore ed il limite superiore del range di normalizzazione, *Min* e *Max* rispettivamente il minimo e il massimo della variabile da normalizzare e *Val* il valore originario della variabile.

   In particolare, utilizzando la funzione logistica come funzione di trasferimento, i valori *ValMin* e *ValMax* sono posti pari a –5 e +5 per i neuroni di input e 0.1 e 0.9 per quelli di output. Nel caso invece di una funzione tangente, è opportuno normalizzare tutti i valori nel range [-1 +1].
2. Il controllo della complessità della rete è fondamentale per garantire una buona capacità di generalizzazione della rete stessa, ossia di trattare dati che, pur appartenendo allo stesso spazio campionario dei dati su cui la rete è stata addestrata, presentano rispetto a questi caratteristiche differenti.

La selezione dell'architettura è stata effettuata addestrando più reti con diverse architetture e scegliendo quella che presenta la migliore performance. Tale valutazione viene in genere realizzata osservando l'andamento dell'errore quadratico medio (EQM) o della percentuale di corretta classificazione (PCC). In realtà tali misure non sono appropriate in quanto non tengono conto del numero di parametri (pesi) che la rete deve stimare. A tale proposito si preferisce combinare l'informazione fornita dall'EQM e dalla PCC con quella fornita dall'*Information Criterion* di Akaike (1970) che invece tiene conto dei gradi di libertà del modello (Moody, 1994).

Sulla base delle precedenti considerazioni e tenuto conto dei risultati in tabella 5 la rete scelta è quella che presenta 8 neuroni nello strato intermedio, funzione di trasferimento logistica per lo strato nascosto e lineare per quello di output. La regola di previsione ottenuta, sfruttando i risultati della rete selezionata a vantaggio del gruppo di 60 tombe non classificate, conduce a definire 39 tombe come maschi, 15 come femmine e 9 come bambini.

Tabella 5: I risultati dell'approccio neuronale

Architettura input/hidden/output	Funz. Trasferimento hidden/output	Algoritmo di apprendimento	# parametri	EQM	PCC	AIC
16-10-6-2	Logistica lineare	Backpropagation adattivo	235	0.10	90.26%	1.83
16-10-6-2	Logistica lineare	Backpropagation Lr=0.1	235	0.27	85.84%	2.86
16-8-2	Logistica lineare	Backpropagation Lr=0.5	218	0.10	90.26%	1.53

3. Per garantire una buona capacità di generalizzazione della rete selezionata si propone, a questo punto, una valutazione e validazione del rischio di predizione attraverso l'utilizzo dei metodi di ricampionamento (Davino, Gherghi, Vistocco, 1997a). In particolare, i concetti di stabilità interna ed esterna (Greenacre, 1984), relativi all'influenza di osservazioni anomale all'interno del campione ed alla rappresentatività dello stesso rispetto alla popolazione di provenienza, possono essere ricondotti agli analoghi concetti di resistenza al rumore (o a dati anomali) e di generalizzazione propri delle reti neuronali.
   Al fine di raggiungere l'obbiettivo di questa terza fase della strategia proposta, l'attenzione è concentrata sull'aspetto relativo alla stabilità esterna. A tale proposito le informazioni che hanno consentito di creare la precedente regola di classificazione, capace di fornire una percentuale di corretta classificazione pari al 90.26%, sono state sottoposte ad una procedura di ricampionamento nota come bootstrap (Efron, 1979). Tale tecnica effettua delle perturbazioni sulla matrice dei dati originari per generare dalla stessa un numero elevato di campioni, ossia 200 matrici sono state generate estraendo casualmente con ripetizione le osservazioni presenti nei dati iniziali.

   In figura 3 è riportato l'istogramma della percentuale di corretta classificazione ottenuta dall'addestramento delle 100 matrici bootstrap. Il valore ottenuto dalla rete selezionata in tabella 5 (90.26%) risulta ben all'interno dell'intervallo centrato sulla media e di ampiezza pari a due volte lo scarto quadratico medio e comprendente circa il 95% della distribuzione dei valori, prova di una buona capacità di generalizzazione della rete selezionata.

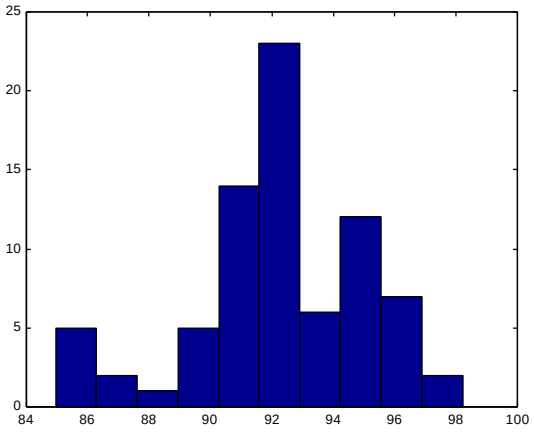


Figura 3: PCC di 100 matrici bootstrap

5.2 La Segmentazione Binaria a due stadi

Le procedure di Segmentazione Binaria sono sorte a partire dagli anni '60 in seguito allo sviluppo dell'informatica e di potenti calcolatori che consentono il trattamento di enormi quantità di dati. Tali tecniche, come si evince dal nome, si propongono di classificare un collettivo di individui in gruppi (*segmentazione*), omogenei al loro interno e quanto più possibile differenziati, attraverso una serie di suddivisioni dicotomiche (*binaria*).

Un punto di svolta nella storia degli alberi di decisione binaria è rappresentato dal CART (*Classification And Regression Trees*) proposto da Breiman, Friedman, Olshen e Stone nel 1984. L'introduzione della metodologia CART ha dato la sensazione, in un primo momento, che il capitolo segmentazione binaria fosse ormai chiuso; i numerosi articoli pubblicati successivamente su riviste internazionali nonché le numerose tecniche proposte in letteratura, hanno invece testimoniato un rinnovato interesse verso l'argomento.

Il punto di partenza di una procedura di segmentazione binaria è rappresentato da un insieme di **N** individui sui quali sono osservate **M** variabili esplicative ed una variabile dipendente in modo da soddisfare un *obiettivo* di tipo *esplicativo/esplorativo*, cioè spiegare la variabile di risposta sulla base delle variabili indipendenti, e un *obiettivo decisionale*, che serve a sfruttare la regola ottenuta per classificare nuovi casi. Le suddivisioni dicotomiche successive del campione iniziale conducono alla costruzione di un *albero di decisione binaria* la cui radice rappresenta la popolazione iniziale, i nodi intermedi rappresentano sottoinsiemi del campione iniziale, e i nodi terminali le "foglie dell'albero" cioè nodi non divisibili che identificano individui appartenenti ad una stessa classe. Il numero delle classi alle quali un individuo può appartenere dipende dal numero di modalità della variabile di risposta.

Nell'analisi della necropoli di Sala Consilina si è utilizzato l'algoritmo di split a due stadi (F. Mola e R. Siciliano, 1992). Tale tecnica evidenzia l'importanza di un effetto di predizione globale delle variabili attraverso una procedura di segmentazione nella cui fase di split viene considerato un criterio a due stadi: considerare innanzitutto il contributo globale di un singolo predittore alla spiegazione della variabile dipendente e poi, per il predittore selezionato, considerare tutte le possibili variabili dicotomiche che si possono ottenere dalle sue categorie. Tale procedura viene iterata più volte fino alla costruzione dell'albero massimo (albero esplorativo) , successivamente viene applicata una procedura di "sfoltimento" dell'albero (in letteratura tale procedura è nota come *pruning* e viene utilizzata per ottenere alberi di taglia ottimale) basata sulla statistica CATANOVA (CATegorical ANalysis of VAriance) generando l'albero decisionale e le regole di classificazione (Mola e Siciliano, 1994; Siciliano, 1998).

L'albero massimo ottenuto (26 nodi terminali) utilizzando le 8 variabili in tabella 1 come variabili esplicative e la variabile genere/età come variabile dipendente, è stato "sfrondato" utilizzando il test Catanova ad un livello di significatività dell'1%. L'albero ottimale ottenuto (Fig. 4) si presenta quindi con quattro nodi terminali.

La variabile considerata maggiormente discriminante, ossia la prima variabile di split, è risultata la lunghezza della tomba per cui una separazione netta viene fatta tra le tombe con lunghezza >160 e quelle medio-corte (<160). I primi due nodi terminali scaturiscono dalla divisione del gruppo di tombe lunghe sulla base delle modalità della variabile forma degli oggetti. Le tombe con presenza con forte presenza di ornamenti

(confermando i risultati della *Cluster Analysis*) individuano il nodo terminale delle femmine mentre le altre il nodo terminale dei maschi.
Le tombe rimanenti si presentano piuttosto confuse come dimostra la divisione generata dalla variabile materiali. L'argilla caratterizza in maniera determinante le tombe dei bambini, mentre solo un piccolo gruppo di tombe si differenzia dalle precedenti per la presenza di oggetti in metallo.

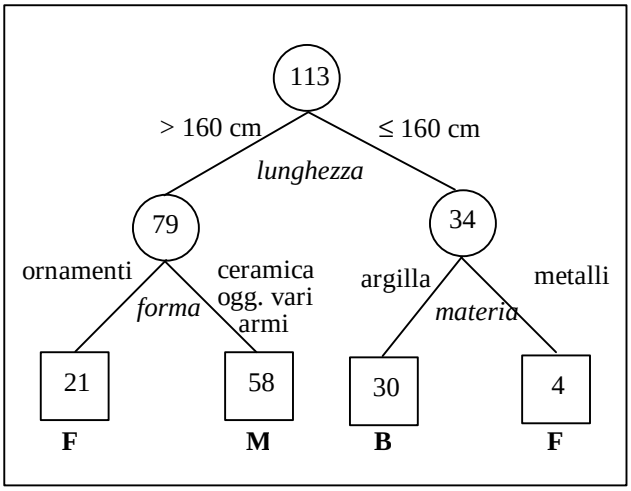


Figura 4: la regola di classificazione

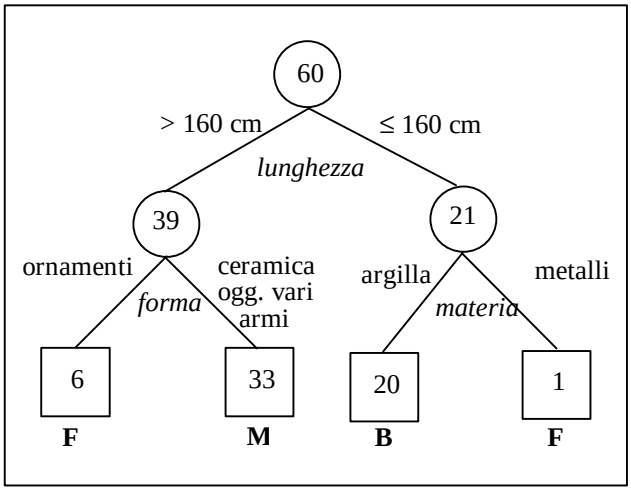


Figura 5: l'albero di previsione

La regola individuata è servita non solo a descrivere le tombe in base alle loro caratteristiche ma anche ad individuare una regola per classificare le tombe in una delle tre modalità della variabile genere/età. In particolare, per avere un'idea del grado di validità della classificazione di base basta osservare che l'83% delle tombe classificate in origine come maschi ha conservato tale connotazione; le tombe di bambino hanno confermato la classificazione per il 100%; le tombe femminili per il 58%.

In conclusione, la percentuale di corrispondenza tra la classificazione ottenuta dall'analisi statistica e la classificazione *a priori* è di circa il 76%.

L'albero ottenuto, è utilizzato a scopi previsivi per la classificazione del gruppo di tombe che non ha partecipato alla creazione della regola di classificazione (35% delle tombe rilevate), in quanto le caratteristiche che esse presentano non ne hanno consentito una classificazione a priori. La figura 5 mostra quindi l'albero di previsione ottenuto facendo "scivolare" le 60 tombe nell'albero ottenuto in figura 3 che definisce 33 tombe come maschi, 7 come femmine e 20 come bambini.

6. Conclusioni

Questo lavoro ha analizzato i dati relativi alla necropoli di Sala Consilina e fornito quindi un ausilio ed uno strumento decisionale all'archeologo che ha condotto la rilevazione, permettendo la classificazione (maschi, femmine o bambini) di alcune tombe che per stato di conservazione, localizzazione e contenuto degli oggetti rinvenuti, non erano facilmente classificabili o quantomeno si prestavano a più possibili classificazioni.

D'altro canto, il lavoro ha anche mostrato come l'uso combinato di tecniche di Analisi dei Dati e le reti neuronali possono sinergicamente fornire una strategia di analisi che, in discipline che non possono sempre analizzare elementi certi (tipico dell'archeologia), si presentano come strumenti sussidiari a disposizione dell'esperto (l'archeologo) e possono più o meno suffragare teorie e ipotesi.

E' evidente che un'analisi critica dei risultati deve essere sempre ed obbligatoriamente condotta dall'esperto; è altresì vero, come mostrato nel lavoro, che la semplice applicazione di una singola metodologia (la classica pressione di un tasto del computer) non solo non è sufficiente, ma addirittura fuorviante e quindi estremamente dannosa per la interpretazione del fenomeno analizzato, mentre un uso combinato di tecniche esplorative e decisionali si presenta sicuramente come una strada da seguire.

Sia le reti neuronali che la segmentazione binaria hanno proposto una classificazione (con diverse probabilità di errore) delle tombe che risultavano difficilmente classificabili. Tuttavia una sostituzione meccanica archeologo/macchine è impensabile. Il nostro scopo è, e riteniamo fortemente debba essere, unicamente quello di fornire ausili interpretativi da considerare con giusto senso critico.

7. Bibliografia

Akaike H. (1970). Statistical predictor identification, *Ann. Inst. Statist. Math.*, 22, pp. 203-217.

Bell S., Croson C. (1998). Artificial Neural Networks as a tool for archaeological Data Analysis, *Archaeometry*,40, pp. 139-151.

Bishop C.M. (1995). *Neural Networks for Pattern Recognition*, Oxford University Press.

Breiman L., Friedman, J.H., Olshen, J.H. and Stone, J.H. (1984). *Classification and Regression Trees*, Wadsworth, Belmont

Cheng B., Titterington., D.M. (1994). Neural Networks: a review from statistical perspective, Statistical Science, 9(1), pp. 1-54.

Davino C., Gherghi, M. e Vistocco, M. (1997), Sulla stabilità delle regole di classificazione nelle reti neuronali, in A. Bellacicco, C. Lauro (eds), *Reti Neurali e Statistica*, pp. 137-147, FrancoAngeli, Milano

Davino C., Mola, F., Siciliano, F. and Vistocco D. (1997). A Statistical Approach to Neural Networks, in K. Fernandez-Aguirre, A. Morineau (eds) *Analyses Multidimensionnelle des Données* IV Congrés International NGUS97, CISIA Ceresta, pp. 37-51.

Efron B, (1979). Bootstrap methods:another look at jackknife, *Annals of Statistics*, 7.

Greenacre M.(1984),. *Theory and Applications of Corresnpondence Analysis*, Academic Press, New York

Lebart L., A. Morineau, M. Piron (1995). *Statistique exploratoire multidimensionnelle*, Dunod, Paris

Maiello M. (1995). *Il Vallo di Diano in età arcaica e classica*, Tesi di Dottorato in Archeologia (rapporti Oriente-Occidente), VII ciclo, Istituto Univesitario Orientale.

Mola e Siciliano (1992). A two-stage predictive splitting algorithm in binary segmentation, in DodgeY. and Witthaker J. (eds). *Computational Statistics* (Compstat '92 proceedings) Physica Verlag, Heidelberg, vol. 1.

Mola e Siciliano (1994). Alternative strategies and CATANOVA testing in two-stage binary segmentation, *New Approaches in Classification And Data Analysis* (Diday E. et al. eds), Springer Verlag, Berlin.

Moody J. (1994). Prediction Risk and Architecture Selection for Neural Networks, in V. Cherkassky, J.H. Friedman, H. Wechsler (eds), *From Statistics to Neural Networks: theory and pattern recognition*, Springer Verlag, Berlin ,pp. 146-165.

Ripley B.D. (1996). *Pattern Recognition and Neural Networks*, University Press, Cambridge.

Siciliano R. (1998), Exploratory Versus Decision Trees, in COMPSTAT '98 (Payne R. and Green P. eds), Physica Verlag, Heidelberg.

Smith M. (1993). *Neural Networks for Statistical Modeling*, International Thomson Publishing.