Archeologia e Calcolatori
15, 2004, 267-286

SISTEMA DI FILOLOGIA COMPUTAZIONALE PER TESTI DEMOTICI

1. Premessa

Il Dipartimento di Storia Antica dell'Università di Pisa, l'Istituto di Linguistica Computazionale del CNR (ILC-CNR, Pisa) e il Dipartimento di Elettronica, Informatica e Sistemistica dell'Università della Calabria (DEIS, Rende) hanno avviato un progetto in collaborazione che ha lo scopo di creare un archivio digitale di testi demotici. Si tratta di circa 1500 testi scritti su ostraka scoperti nel tempio di Medinet Madi (regione del Fayyum in Egitto) nel 1938 da Achille Vogliano. Questi documenti sono oggi conservati nel museo egizio del Cairo.

La ricerca si fonda sulla convinzione che il crescente sviluppo della tecnologia digitale e la conseguente disponibilità di archivi di immagini, sia in strutture pubbliche che private, possano influenzare positivamente anche la ricerca e la didattica nel settore antichistico. Dal momento che il documento digitale può essere sottoposto a trattamento elettronico mediante sistemi computerizzati, viene facilitata la creazione di grandi banche dati con il vantaggio di favorire la conservazione e la fruizione dei documenti stessi[1].

Dal punto di vista della fruizione, per esempio, si intende mettere in rilievo che uno studioso di un testo o di un reperto che sia disponibile su un archivio digitale collegato alla rete Internet può in molti casi avviare una attività di studio semplicemente operando sulla copia digitale e verificando eventualmente l'originale come conferma e controllo finale del proprio lavoro critico editoriale.

Se da questo punto di vista la tecnologia informatica e telematica ha messo a disposizione strumenti certamente molto potenti, non possiamo tuttavia credere che alcuni limiti siano superabili almeno in tempi brevi. Ci si riferisce, in particolare, alla possibilità di effettuare classificazioni automatiche dei caratteri che compongono un testo manoscritto antico, anche se i sistemi di intelligenza artificiale (per esempio le reti neurali) hanno già dimostrato grandi potenzialità (Bozzi 2000). In questo settore degli studi e delle sperimentazioni, ottimi risultati si sono raggiunti là dove i caratteri di un set alfabetico anche manoscritti sono ben separati gli uni dagli altri e dove in essi non si presentino deterioramenti. In queste situazioni operano molto bene anche prodotti che sono oggi in commercio, ma è sufficiente che l'immagine digitale, sia pure

[1] Un esempio dell'uso di un archivio di immagini digitali per l'analisi filologica di manoscritti medievali è descritto analiticamente in Bozzi 1997.

convertita con una risoluzione ottimale, presenti caratteri su un testo a stampa antico con elementi di danneggiamento al supporto cartaceo o evanescenza dell'inchiostro che il risultato della classificazione peggiora e il riconoscimento automatico del testo necessita di un consistente intervento umano.

Queste osservazioni preliminari sono state propedeutiche all'avvio del programma di digitalizzazione di un corpus di ostraka demotici e di una loro gestione mediante uno specifico sistema computerizzato.

Le motivazioni che hanno spinto a selezionare questo archivio di ostraka come base di sperimentazione di un sistema di filologia computazionale sono principalmente le seguenti:

1) questi testi mostrano delle caratteristiche paleografiche che li rendono adatti al trattamento computerizzato, sia pure di tipo sperimentale, descritto in questo lavoro. In particolare troviamo in essi una tendenza verso la semplificazione del disegno grafico delle parole che consiste nell'uso piuttosto regolare di segni monoconsonantici al posto di grafemi più elaborati e nell'uso di un piccolo gruppo di classificatori (determinativi);
2) tutti i testi sono stati rinvenuti nello stesso sito e possono essere datati allo stesso periodo (II-III secolo d.C.), formando, perciò, un gruppo omogeneo con irrilevanti differenze di lingua e di tipo di scrittura;
3) gli ostraka sono disponibili in formato digitale con ottima risoluzione e a colori;
4) questi testi sono stati ripetutamente analizzati dagli studiosi dell'Università di Pisa da quando Edda Bresciani ne intraprese lo studio e la pubblicazione negli anni '70. Un numero significativo di tali documenti è stato già edito (Bresciani *et al.* 1983; Gallo 1997) e la pubblicazione di altri è tuttora in corso.

Lo scopo del progetto consiste nella realizzazione di una biblioteca digitale con un consistente numero di testi demotici (che in previsione non sarà limitato solo a quelli su ostraka) e nella possibilità di effettuare operazioni di ricerca di vario tipo che possono essere così sintetizzate:

– ottenere la visualizzazione sull'immagine digitale di tutte le zone nelle quali un determinato segno, che sia stato selezionato su una keyboard virtuale o su una tavoletta digitale, si trova nell'archivio digitale degli ostraka demotici (DIADO);
– ottenere la visualizzazione sull'immagine digitale di tutte le zone nelle quali più segni, contigui o meno, che siano stati selezionati su una keyboard virtuale o su una tavoletta digitale, si trovano in DIADO;
– ottenere la stampa di tutte le zone-immagine nelle quali i segni selezionati su una keybord virtuale o su una tavoletta digitale si trovano in DIADO.

Tali finalità sono giustificate dalla impossibilità di ottenere con rapidità e precisione i medesimi risultati in assenza di uno specifico strumento infor-

matico, considerando soprattutto che l'archivio potrebbe raggiungere dimensioni molto consistenti. Ulteriori ricadute si possono avere sia nel versante didattico, sia in quello specialistico: il progetto, qualora produca i risultati attesi, metterebbe a disposizione degli studiosi elementi importanti per l'analisi dei fenomeni paleografici con innegabili conseguenze per quanto concerne l'interpretazione e la datazione dei documenti stessi.

2. Fasi di progettazione

Il progetto di fattibilità prevede uno sviluppo in fasi che possono essere così sintetizzate:

a) digitalizzazione delle immagini degli ostraka originali o di una loro riproduzione fotografica a colori;
b) digitalizzazione dei fac-simile in bianco e nero;
c) progettazione di un sistema computerizzato grazie al quale l'utilizzatore possa evidenziare ciascun segno dell'alfabeto demotico presente sul testo degli ostraka digitalizzati per la sperimentazione;
d) progettazione di una rete neurale artificiale che: 1) apprenda le caratteristiche grafiche di tutti i segni dell'alfabeto demotico sulla base di almeno una trentina di campioni che l'utente ha scelto per ciascuno di essi; 2) classifichi i segni che gli vengano sottoposti in fase di riconoscimento;
e) progettazione di un'interfaccia grafica che faciliti le operazioni di ricerca delle zone-immagine nelle quali uno o più segni demotici si trovino all'interno dell'archivio delle immagini digitali.

Senza entrare per il momento nel dettaglio della suddivisione che i segni demotici posso avere (segni univoci, bilitteri, trilitteri, determinativi, ecc.), vediamo di affrontare alcune caratteristiche del sistema come indicato ai punti c) e d).

2.1 *La preparazione del training set*

Una rete neurale[2] che diventi abile a classificare i segni del demotico deve "imparare" ed estrarre le caratteristiche grafiche da un campione significativo di segni. Riteniamo che 30 segni sia un livello accettabile, anche se, ovviamente, un numero superiore di campioni per ogni singolo segno potrebbe rendere il sistema più specializzato e sicuro anche di fronte alla presenza di variazioni "stilistiche" considerevoli. Dal momento che in questa fase sperimentale alcuni

[2] Il sistema da noi proposto prevede l'utilizzo di una rete neurale per il riconoscimento e la classificazione automatica dei caratteri presenti nelle iscrizioni. La rete neurale utilizzata è un tipo particolare di rete che deriva dallo schema cosiddetto μLVQ2, che a sua volta, come dice il nome stesso, è una rete autoorganizzante LVQ con una leggera modifica capace di migliorarne le prestazioni generali. Tali reti prevedono uno o più modelli (*codebooks*) rappresentativi per ogni classe da riconoscere.

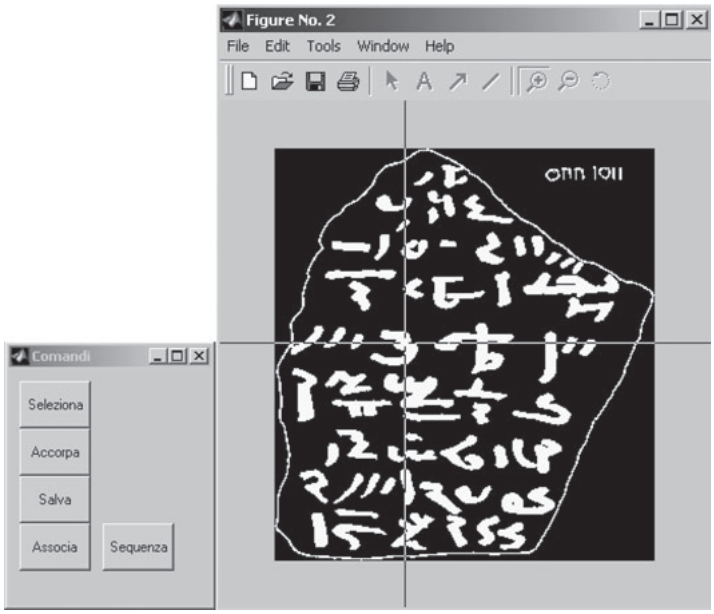

Fig. 1 – Sistema di selezione dei segni grafici da sottoporre alla rete neurale.

segni sono presenti sui documenti in numero troppo esiguo, ne sono stati presi in considerazione alcuni e sono stati graficamente modificati mediante un programma. La selezione dei segni grafici da sottoporre alla rete neurale avviene con un sistema del tipo di quello proposto nella Fig. 1.

L'utente colloca il selettore (che si presenta come il punto in cui una linea orizzontale ed una verticale si incrociano) su un segno grafico e agisce sul tasto sinistro del mouse. Appare una finestra nella quale si vede stilizzato ed ingrandito il segno demotico selezionato (Fig. 2).

Per inserire tale segno nella lista di quelli che la rete deve apprendere (il *training set*) è sufficiente selezionare il bottone "Memorizza": compare una tastiera virtuale che serve per assegnare il valore che deve essere associato al segno selezionato (Fig. 3).

Si noti che questa tastiera è costituita da segni demotici non normalizzati, ma estratti dai fac-simile e considerati significativi da un punto di vista stilistico. Questo tipo di attività serve per supervisionare la rete neurale la quale apprende solo sulla base di conoscenze che le vengono fornite in una fase che, appunto, viene definita "di apprendimento". Nel caso in cui il segno demotico sia costituito di più grafemi diversi e separati fra loro, il sistema consente di selezionarli tutti e di accorparli: il risultato dell'accorpamento si può verificare nella finestra di controllo (Fig. 4).

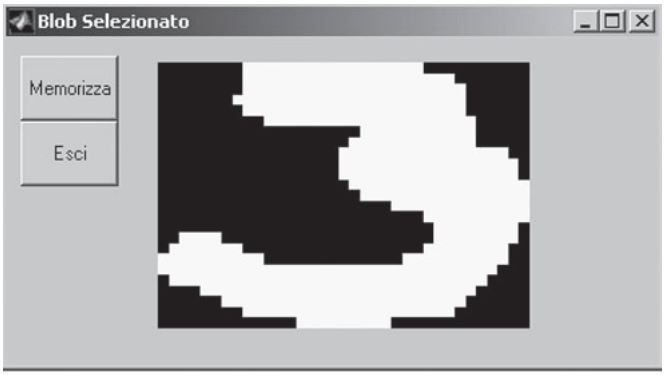

Fig. 2 – Stilizzazione e ingrandimento del segno demotico selezionato.

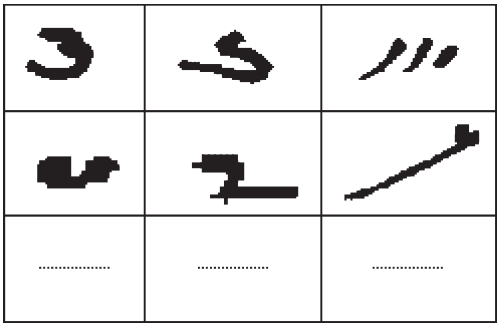

Fig. 3 – Tastiera virtuale per assegnare il valore da associare al segno selezionato.

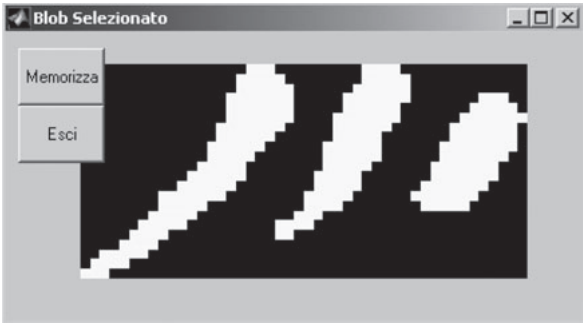

Fig. 4 – Finestra di controllo della procedura.

2.2 *Pre-elaborazione delle immagini per l'estrazione delle caratteristiche grafiche (feature)*

Quella dell'estrazione delle feature è una delle fasi più importanti per la realizzazione di un buon classificatore. Attraverso le feature si ottiene una rappresentazione dei caratteri da riconoscere in una forma diversa da quella in due dimensioni che sostanzialmente corrisponde all'immagine digitale nella quale ciascuno di essi è contenuto. Prima della fase dell'estrazione delle feature, per cercare di eliminare una parte delle informazioni inutili, si passa attraverso una fase di *pre-processing* necessaria per una regolarizzazione delle immagini che rappresentano i caratteri da riconoscere. Questa fase prevede l'applicazione di tecniche per la rimozione del rumore e di tecniche per l'evidenziazione dei bordi che, nel caso del riconoscimento di caratteri, sono la parte dell'immagine che racchiude tutta l'informazione necessaria.

Per la fase di rimozione del rumore, di solito, si opera con delle tecniche di media effettuate su intorni opportuni dei pixel dell'immagine. Questo tipo di soluzione porta a degli effetti collaterali, consistenti in una sfocatura dell'immagine risultante. La tecnica utilizzata in questo lavoro per la rimozione del rumore è quella del filtraggio mediano che provoca una sfocatura dell'immagine risultante decisamente inferiore rispetto a quella prodotta dalle tecniche di media locale. I risultati ottenuti si vedono nella figura che mostra nell'ordine: l'immagine originale (a sinistra), quella prodotta con l'applicazione delle medie locali (in alto a destra) e quella ottenuta con il filtro mediano (Fig. 5).

L'evidenziazione dei bordi si effettua tramite convoluzione con una maschera di dimensione 3×3 che realizza un filtraggio del tipo "passa alto" dell'immagine. La maschera utilizzata in questo lavoro è la seguente:

$$h(i,j)=\begin{pmatrix}-1 & -1 & -1\\ -1 & 9 & -1\\ -1 & -1 & -1\end{pmatrix}$$

Essa ci permette di ottenere un buon effetto di evidenziazione dei bordi come si può osservare nella Tav. III, a.

Dall'immagine ottenuta dalla fase di *pre-processing* si effettua l'estrazione delle feature secondo le modalità di seguito specificate.

2.2.1 Estrazione delle feature: le proiezioni

Una proiezione orizzontale $y(x_i)$ rappresenta il numero di pixel, di valore pari ad uno, con ascissa x_i (Fig. 6). Questo tipo di feature può essere reso indipendente dalla scala utilizzando un numero fisso di divisioni per ogni asse, sommando cioè le proiezioni per i punti con coordinate adiacenti, ed

Fig. 5 – Tecniche per la rimozione del rumore.

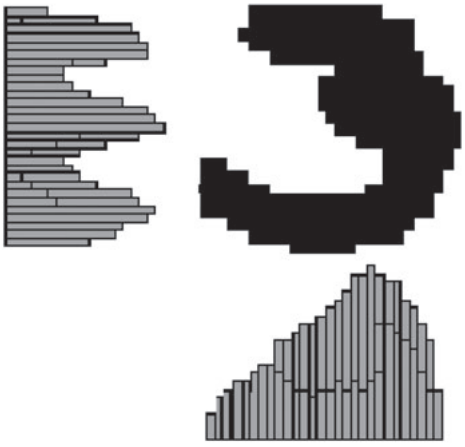

Fig. 6 – Istogramma delle proiezioni sugli assi *x* ed *y* di un carattere.

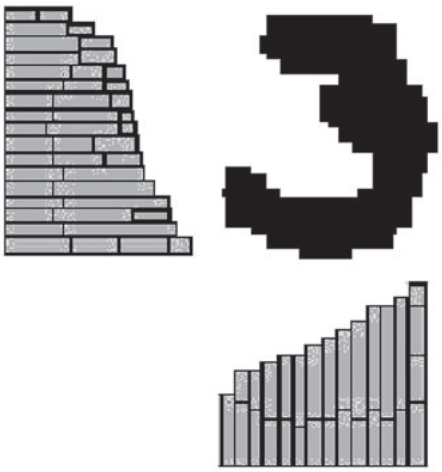

Fig. 7 – Istogramma cumulativo delle proiezioni sugli assi *x* ed *y* di un carattere.

inoltre dividendo per il numero totale di pixel, di valore uno, presenti nell'intero carattere. Il risultato ottenuto è un istogramma che rappresenta la distribuzione dei pixel nelle diverse zone del carattere e ci permette di ottenere delle indicazioni utili alla sua identificazione. Questo tipo di feature, comunque, è molto sensibile alle rotazioni ed alle variazioni dello stile di scrittura. Le proiezioni verticali $x(y_i)$, che rappresentano il numero di pixel con ordinata y_i, sono abbastanza immuni alle variazioni nell'orientamento del carattere, mentre quelle orizzontali lo sono molto meno.

Più in generale è possibile definire la proiezione di un'immagine su una qualsiasi retta di pendenza θ, sommando i suoi elementi lungo una famiglia di rette perpendicolari a θ. Anziché utilizzare i semplici istogrammi, ottenuti dalle proiezioni, è più significativo utilizzare un istogramma cumulativo, che rappresenta la somma degli istogrammi delle singole divisioni secondo la seguente formula, valida per il caso delle proiezioni orizzontali:

$$Y(x_k) = \sum_{i=1}^{k} y(x_i)$$

dove $y(x_i)$ rappresenta la proiezione precedentemente descritta. A differenza delle suddette proiezioni, questa funzione cumulativa gode della proprietà di una maggiore immunità a piccole differenze nell'allineamento dei picchi dominanti negli istogrammi originali quando si effettuano confronti fra istogrammi relativi a caratteri diversi (Fig. 7).

La scelta degli assi da utilizzare per le proiezioni, in questo lavoro, è ricaduta su quelli principali e sulla famiglia di rette perpendicolari alle diagonali, principale e secondaria, del rettangolo che racchiude esattamente il carattere da riconoscere. La strada per ottenere tali proiezioni passa, come già anticipato, attraverso le fasi di pre-elaborazione dell'immagine. Una volta ripulita l'immagine dal rumore, essa contiene esattamente un carattere contornato da una zona, più o meno estesa, di bordo in cui non è presente alcuna informazione. Il primo passo della pre-elaborazione consiste proprio nell'eliminare dall'immagine queste zone inutili, ottenendone un'altra che racchiude esattamente il carattere da riconoscere.

Successivamente, data la scelta di utilizzare nove divisioni per le proiezioni e considerando che il rettangolo che racchiude il carattere ha dimensioni qualsiasi, si opera una maggiorazione delle dimensioni dell'immagine fino a raggiungere il più vicino multiplo del numero di divisioni utilizzate in modo da considerare, per ogni divisione, una identica regione dell'immagine. Per ottenere un numero fissato di divisioni, si potrebbe anche riportare ogni immagine ad una dimensione precostituita, ma tale scelta può portare a delle distorsioni sul carattere presente nell'immagine. Il metodo adottato in questo lavoro consente di limitare tale inconveniente in quanto le dimensioni originali dell'immagine subiscono una modifica minima.

Per ottenere una maggior discriminazione fra le diverse cifre, oltre alle proiezioni precedentemente descritte che ci indicano come i pixel che formano ogni cifra sono distribuiti spazialmente rispetto al rettangolo che la racchiude, si è pensato di aggiungere delle proiezioni particolari che tengono conto della diversa morfologia dei caratteri a scapito dei possibili orientamenti che queste possano presentare. Queste proiezioni, prese in forma cumulativa come le precedenti, sono relative a due assi ortogonali la cui origine coincide con il centroide del carattere in esame. Il centroide, chiamato anche centro di gravità, è il punto di bilanciamento della funzione immagine *f(x,y)* tale che la massa, rappresentata dal numero di pixel con valore uno, che si trova a destra ed a sinistra della sua ascissa X_k e quella che si trova sopra e sotto la sua ordinata Y_j sia uguale. I rapporti:

$$\begin{cases} X_k = \dfrac{M(1,0)}{M(0,0)} \\ \\ Y_j = \dfrac{M(0,1)}{M(0,0)} \end{cases}$$

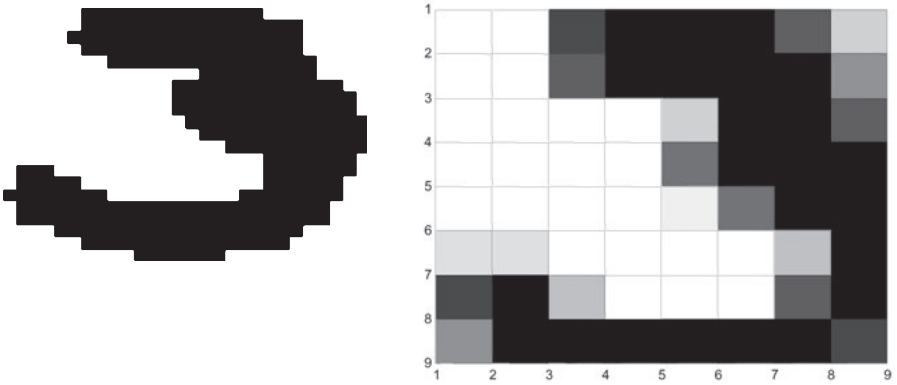

Fig. 8 – Un esempio di zoning con una griglia 9×9.

definiscono le coordinate del centro di gravità di un'immagine, dove *M(m,n)* rappresenta il momento spaziale di ordine (n+m), espresso dalla seguente formula:

$$M(m,n) = \frac{1}{J^n K^m} \sum_{j=0}^{J-1} \sum_{k=0}^{K-1} (x_k)^m (y_j)^n f(j,k)$$

in cui *J* e *K* rappresentano le dimensioni dell'immagine *f(j,k)* e x_k ed y_j sono le coordinate trasformate dal dominio del continuo a quello delle immagini digitali, dove l'origine del sistema è posto nell'angolo in alto a sinistra dell'immagine stessa, secondo le seguenti relazioni:

$$\begin{cases} x_k = k - \frac{1}{2} \\ y_j = J - j + \frac{1}{2} \end{cases}$$

Questo tipo di feature, quindi, fornisce informazioni in merito a come la massa del carattere sia distribuita attorno al suo centroide permettendo un migliore discernimento fra caratteri diversi.

2.2.2 Estrazione delle feature: analisi zonale o "zoning"

L'analisi zonale consiste nella sovrapposizione di una griglia $n \times n$ all'immagine che rappresenta il carattere ed al calcolo, per ognuna delle $n \times n$ zone, del valore medio dell'insieme dei pixel che essa racchiude (Fig. 8). Questo

valore, nel caso di immagini binarie, rappresenta la percentuale di pixel neri rispetto al numero totale di pixel presenti in ogni zona. In questo modo si ottiene una sorta di compressione dell'informazione contenuta nell'intera forma del carattere, da utilizzare direttamente come ingresso alla rete neurale.

Questo tipo di feature è fortemente influenzato dalle dimensioni dell'immagine in quanto, assegnando alla griglia un numero uguale di regioni, le zone appartenenti a caratteri diversi potrebbero essere di dimensioni molto diverse fra loro. Per ovviare a questo problema si deve effettuare una normalizzazione dei valori che rappresentano le percentuali di pixel in zone relative a diverse immagini. Il metodo utilizzato in questo lavoro per ottenere tale normalizzazione consiste nell'estrarre, fra le diverse zone che interessano uno stesso carattere, quella con la concentrazione massima di pixel e quindi normalizzare tutte le altre rispetto ad essa, semplicemente effettuando una divisione fra i valori di tutte le componenti del vettore per il valore che corrisponde alla zona con la concentrazione massima. Come nel caso precedente la dimensione scelta per la griglia è nove. Complessivamente, allora, dalla fase di estrazione delle feature si ottiene un vettore così composto:

– *9×9* informazioni relative allo zoning;
– *27* proiezioni orizzontali, ulteriormente suddivise in:
  – *9* proiezioni sull'asse delle ascisse dell'immagine,
  – *9* nella parte superiore dell'asse passante per il centroide dell'immagine,
  – *9* nella parte inferiore dell'asse passante per il centroide dell'immagine;
– *27* proiezioni verticali, suddivise in maniera analoga a quelle orizzontali;
– *9* proiezioni sul fascio di rette passante per la diagonale principale del rettangolo racchiudente il carattere;
– *9* proiezioni sul fascio di rette passante per la diagonale secondaria dello stesso rettangolo.

2.3 *Classificazione semiautomatica dei segni su un testo demotico*

Terminata la fase di addestramento, il sistema è predisposto per aiutare a classificare i segni contenuti su testi digitali degli ostraka, ma vi sono buone probabilità che le capacità classificatorie della rete neurale siano in grado di operare anche su testi demotici su papiro. La fase di classificazione procede in maniera assai simile a quella usata per la costituzione del *training set*. Si seleziona sull'immagine il segno (eventualmente accorpandone i tratti che lo compongono) e lo si sottopone al riconoscimento: la tastiera virtuale mostra in colore rosso il simbolo al quale quello riconosciuto è stato associato con il massimo grado di somiglianza (*dependebility rate*) (Tav. III, b). È possibile inoltre far sì che la rete neurale mostri, eventualmente con colori differenti, altre associazioni in ordine inverso di probabilità.

Fig. 9 – Caratteri utilizzati nell'esempio.

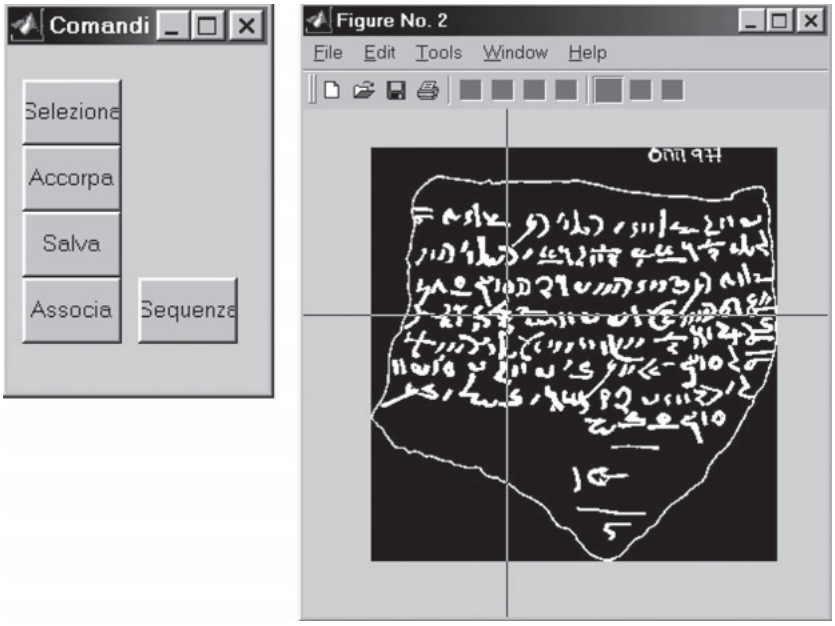

Fig. 10 – Programma per la selezione dei caratteri.

3. Realizzazione di un esempio

A titolo di esempio abbiamo costruito una rete su un set di 240 caratteri suddivisi in ugual numero su 6 classi. I caratteri relativi alle diverse classi utilizzate sono indicati nella Fig. 9. Per l'estrazione dei singoli caratteri dalle immagini che rappresentano le iscrizioni si è realizzato un programma che permette di selezionarli semplicemente cliccando sulla parte dell'immagine che ci interessa ed offrendo la possibilità di memorizzare le immagini catturate in file separati (Fig. 10).

Classe di appartenenza	'1'	'2'	'3'	'4'	'5'	'6'
Codebooks assegnati	3	1	3	2	3	2

	Classe di assegnazione					
Classe di appartenenza	31	1	1	0	0	2
	0	30	0	5	0	0
	0	0	33	2	0	0
	0	0	0	35	0	0
	0	0	0	0	35	0
	0	0	0	0	0	35

Tab. 1

Attraverso il programma realizzato abbiamo selezionato 40 caratteri diversi per ognuna delle classi scelte. Utilizzando 5 caratteri per classe come insieme di addestramento e 35 per la fase di verifica si ottengono i seguenti risultati rispettivamente per i *codebooks* assegnati alle diverse classi e per i caratteri riconosciuti ed erroneamente assegnati dalla rete racchiusi in una matrice di confusione (Tab. 1).

Si noti che la percentuale di riconoscimento ottenuta è pari circa al 94,8%. La scarsità dei dati è il motivo principale che ci porta a considerare il risultato ottenuto in maniera molto poco indicativa delle effettive potenzialità della rete neurale. D'altra parte il rapporto fra il numero di caratteri utilizzati per l'addestramento e quelli per la fase di verifica ci conforta della bontà del modello studiato. Per cercare di rendere più significativo l'esempio, abbiamo estratto 10 caratteri per classe dall'insieme a nostra disposizione, sovrapponendovi del rumore, e con essi addestrare la rete; tutti gli altri caratteri sono stati usati per la fase di verifica. In questo esempio, ai caratteri delle prime tre classi si è aggiunto un rumore Gaussiano con media nulla e varianza pari a 0,1, mentre ai caratteri delle classi rimanenti si è aggiunto un rumore di tipo *salt & pepper* con una densità di rumore pari a 0,2. I risultati ottenuti in questo caso sono riportati nella Tab. 2.

La percentuale di riconoscimento ottenuta in questo caso sale al 97,5%. Per cercare di esasperare l'esempio precedente, esso è stato ripetuto incrementando il rumore aggiunto ad ogni immagine. In questo caso, un valore della media nullo ed un valore della varianza di 0,2 sono stati applicati alle classi caratterizzate da rumore Gaussiano, mentre un valore di densità di rumore pari a 0,3 per le altre. I risultati sono indicati nella Tab. 3.

Classe di appartenenza	'1'	'2'	'3'	'4'	'5'	'6'
Codebooks assegnati	3	1	2	2	3	2

	Classe di assegnazione					
Classe di appartenenza	39	1	0	0	0	0
	1	37	0	2	0	0
	0	0	39	1	0	0
	0	0	0	39	1	0
	0	0	0	0	40	0
	0	0	0	0	0	40

Tab. 2

Classe di appartenenza	'1'	'2'	'3'	'4'	'5'	'6'
Codebooks assegnati	3	3	6	2	3	2

	Classe di assegnazione					
Classe di appartenenza	33	5	2	0	0	2
	0	36	0	4	0	0
	0	2	36	1	0	1
	0	0	0	38	2	0
	0	0	0	0	40	0
	0	0	0	0	0	40

Tab. 3

Classe di assegnazione	**'3'**	**'6'**	**'1'**	**'3'**
Coefficiente di correlazione	0,57	0,49	0,66	0,51

Tab. 4

Classe di appartenenza	'1'	'2'	'3'	'4'	'5'	'6'
Coefficiente di correlazione	0,93	0,90	0,90	0,89	0,96	0,98

Tab. 5

Fig. 11 – Caratteri di test.

In questo caso la percentuale di riconoscimento è ragionevolmente più bassa rispetto alle situazioni precedenti, attestandosi su valori di poco inferiori al 93%. Un ulteriore esperimento è stato realizzato per cercare di classificare alcuni caratteri appartenenti a classi estranee a quelle utilizzate per la costruzione della rete e verificando poi il livello di compatibilità di questi caratteri con il *codebook* vincente rappresentativo della classe a cui vengono comunque assegnati dalla rete. I caratteri utilizzati sono indicati nella Fig. 11.

La Tab. 4 indica a quale classe questi caratteri sono stati assegnati dalla rete, riportando rispettivamente il coefficiente di correlazione fra di loro ed il codebook vincitore.

Estraendo un vettore per ognuna delle classi adoperate per la costruzione della rete e ripetendo l'esperimento fatto per i caratteri estranei ci si può rendere conto del buon funzionamento del sistema realizzato in quanto, come si può osservare nella Tab. 5, i coefficienti di correlazione con i *codebooks* vincenti hanno dei valori decisamente più alti.

4. Il sistema di filologia computazionale

Il lavoro sopra descritto si inserisce nel quadro di un progetto molto più vasto relativo alla realizzazione di un sistema di filologia computazionale per lo studio e l'edizione di fonti manoscritte e a stampa antiche che è in avanzata fase di sviluppo presso l'Istituto di Linguistica Computazionale del CNR a Pisa, in collaborazione con la società M.E.T.A. di Lucca. Esso è costituito da due diversi ambienti tecnologici: il primo opera come una postazione *stand-alone* e permette operazioni molto avanzate di critica testuale (con-

sultazione simultanea di archivi di fonti digitali collazionate per reperire eventuali varianti al testo, registrazione di apparati, indicizzazioni, ecc.)[3]. Il secondo ambiente è operativo in una rete locale (un Dipartimento, un Centro culturale, una biblioteca, un archivio) e, pur possedendo caratteristiche simili al sistema *stand-alone*, offre strumenti più semplici e minori funzioni di tipo filologico.

Quest'ultimo, tuttavia, si caratterizza per una serie di componenti che lo rendono aderente agli standard internazionali del settore ed ha una struttura a moduli software tale da consentire l'inserimento di nuovi programmi senza dover riprogettare quanto già realizzato. L'architettura hardware si basa su una rete locale nella quale un certo numero di postazioni (client) dialogano con un server: ad uno dei client è connesso uno scanner per l'acquisizione dei dati (da originale o da microfilm) che vengono poi inviati al server che funge da deposito centralizzato. I dati sono memorizzati sul server in forma ridondante in modo da garantire la continuità del lavoro anche nel caso in cui un disco di memoria di massa subisca un danno.

Ogni singolo client accede ad un catalogo generale utilizzando un browser Internet: una volta che l'utente abbia effettuato la scelta, il sistema porta in visione l'immagine digitale del documento selezionato in una finestra collocata nella parte sinistra dello schermo. Il menù relativo alle immagini offre dei bottoni che attivano procedure di *digital image processing* allo scopo di favorire una visione ottimale del documento e del contenuto (aumento della luminosità, del contrasto, ingrandimenti prestabiliti o sequenziali, ecc.), funzioni che si rivelano molto utili quando si è in presenza di fonti in cattivo stato di conservazione.

Alla destra dell'immagine un programma di videoscrittura mette a disposizione un *file* ove si effettua la trascrizione del testo in essa contenuto. Il testo, se già trascritto, può essere importato da dischetto, da CD o da altro supporto magnetico o digitale.

Nel caso in cui si tratti di opere redatte con alfabeti diversi da quello latino, viene offerta la possibilità di selezionare il font desiderato su una tastiera virtuale, seguendo un criterio simile, sia pure semplificato, a quello che abbiamo descritto a proposito dei testi demotici. Questa soluzione rende

[3] Il programma è stato adoperato sperimentalmente su varie edizioni a stampa antiche del *Contradicentium medicorum* di Gerolamo Cardano, in vista della produzione di un CD da parte della sezione milanese dell'Istituto per la storia del pensiero filosofico e scientifico moderno del CNR. La preparazione dei dati è avvenuta a cura di Guido Canziani e Maria Luisa Baldi. Ottimi risultati si sono avuti anche nel settore della filologia romanza (testi occitanici di argomento medico-farmaceutico seguiti da Maria Sofia Corradini dell'Università di Pisa) e della papirologia greca, con l'utilizzo di un campione di frammenti di ricette mediche conservate presso l'Istituto Papirologico "G. Vitelli" di Firenze (con la consulenza di Isabella Andorlini). Per maggiori informazioni cfr. Bozzi 2003.

possibile la fase della trascrizione anche da parte di personale di archivio e/o di biblioteca non particolarmente esperto: in molti casi è sufficiente che essi siano in grado di interpretare la scrittura con la quale i documenti sono stati redatti.

5. Il problema degli standard e della codifica dei dati

La registrazione di dati in formato elettronico assume un significato molto importante se a beneficiarne non sono solo coloro che li producono, ma tutta una comunità di utenti che la rete Internet rende innumerevoli e non facilmente quantificabili. Per tale motivo è necessario adottare, fin dove è possibile, tutti gli strumenti che sono oggi disponibili al fine di adeguare il lavoro prodotto agli standard internazionali, gli unici in grado di garantire ad una banca dati una reale compatibilità con le altre dello stesso tipo. Tale compatibilità si realizza su due livelli: a livello di sistema di catalogazione dei documenti (compatibilità dei metadati), e a livello dei dati stessi (immagini e, soprattutto, testi).

Per quanto riguarda il primo aspetto, il sistema di catalogazione da noi adottato non è ancora del tutto compatibile con gli standard del settore. Ciò è dovuto a due fatti: in primo luogo le informazioni messe a disposizione degli Istituti culturali che hanno collaborato al progetto[4] appartengono a categorie molto differenziate (libri, manoscritti, illustrazioni su cartoline e manifesti, documenti archivistici, oggetti museali). Inoltre, ogni Istituzione aveva organizzato i propri archivi in formato elettronico secondo sistemi molto diversi (TinLib, CDS/ISIS, Guarini Archivi, FileMaker) che dovevano venire armonizzati in un ambiente valido per tutti e capace di garantire un accesso integrato all'insieme dei dati.

I testi che sono stati trascritti e che sono consultabili nella rete locale, invece, hanno ricevuto una codifica basata sul linguaggio XML e nel rispetto delle norme stabilite dal comitato internazionale della TEI (*Text Encoding Initiative*). Ciò significa, per esempio, che tutte le parole del testo, tutti i segni di punteggiatura e tutte le cifre numeriche ricevono una codifica differenziata che viene attribuita automaticamente dal sistema. L'utente, poi, può

[4] Si tratta del progetto FAD (Fondi ed Archivi Digitali), finanziato dal Ministero per i Beni e le Attività Culturali nell'ambito della Legge 21/12/1999 n. 513 ("Interventi straordinari nel settore dei beni e delle attività culturali"). Proposto dalla "Fondazione Primo Conti onlus" di Fiesole, ha visto coinvolti altri tre istituti culturali (il Gabinetto G.P. Vieusseux e l'Istituto Papirologico "G. Vitelli" di Firenze nonché la "Fondazione Nello e Carlo Rosselli" di Torino) e un istituto di ricerca del CNR (Istituto di Linguistica Computazionale di Pisa) il cui dirigente di ricerca, Andrea Bozzi, ne ha assunto la responsabilità scientifica. Il coordinamento tecnico e lo sviluppo del software è stato curato da Alberto Raggioli della M.E.T.A. di Lucca. I risultati del progetto sono stati presentati in un convegno a Firenze il 14 maggio 2004; una descrizione delle attività si legge all'indirizzo www.fadnet.org oppure nel contributo di Bozzi, Raggioli 2003.

ulteriormente precisare la funzione espressa da ciascuna di queste classi: per esempio, egli può distinguere le cifre numeriche che si riferiscono ad una data da quelle che, invece, sono relative a peso o distanza. Questa struttura di classificazione consente di ottenere una pluralità di selezioni che, in effetti, si traduce in una produzione di indici effettivamente utili all'interpretazione linguistica, stilistica e filologica della documentazione stessa.

La sperimentazione della postazione di lavoro ha ormai superato la fase prototipale dal momento che sono state allestite tante reti locali quante sono le Istituzioni partecipanti per un totale di circa 300.000 documenti. Essi, oltre ad essere catalogati secondo il sistema omogeneo del quale abbiamo detto sopra, sono rappresentati da immagini digitali ad alta risoluzione; le fasi di trascrizione sono state avviate nell'ambito del progetto e continueranno secondo criteri di priorità ancora da stabilire.

6. Conclusioni

Le applicazioni descritte in questo contributo sono frutto di una ipotesi di lavoro che vede, nell'integrazione fra fonti archivistico-bibliotecarie e tecnologia informatico-telematica una grande opportunità di rinnovamento di metodi e discipline che oggi risentono di una notevole crisi di identità. Se da un lato, infatti, sentiamo parlare sempre più spesso dell'importanza delle Biblioteche Digitali con un forte richiamo ai vantaggi di tipo turistico-economico che esse comportano[5], sempre più debole, invece, è il sostegno effettivo allo sviluppo degli strumenti veramente specialistici che si devono accompagnare a qualsiasi attività in questo settore affinché vengano prodotti risultati duraturi. Per quanto concerne la fruizione dei beni librari mediante tecnologia digitale nella prospettiva di rinnovamento delle discipline filologiche si sono fatti notevoli passi avanti, nonostante le difficoltà di reperire fonti di finanziamento adeguate. L'interesse manifestato a livello internazionale su questo tema sta crescendo e la Commissione Europea, unitamente alla Fondazione Europea di Strasburgo[6], promuovono ormai da un decennio alcuni

[5] Ci si riferisce qui al Network Turistico Culturale telematico, rappresentato da una serie di iniziative volte a promuovere l'innovazione tecnologica digitale, approvato nell'agosto scorso dal Comitato dei Ministri per la Società dell'Informazione su proposta avanzata dal ministro per i Beni e le Attività Culturali, per avviare la valorizzazione, mediante la produzione di contenuti digitali per le reti globali, di quello che viene definito come il nostro "petrolio".

[6] Le prime iniziative in tal senso furono avviate nell'ambito del 3° Programma Quadro di ricerca e sviluppo e sono terminate negli anni 1996/1997. Ricordiamo, a titolo di esempio, il progetto *BAMBI – Better Access to Manuscript and Browsing of Images – LIB 3114*) per il quale cfr. Bozzi 1997. La Fondazione Europea della Scienza ha di recente (settembre 2003) co-finanziato, unitamente a CNRS di Parigi e Regione Toscana, una conferenza internazionale (ESF-EURESCO) dal titolo "Philological Disciplines and Digital Technology. Computational Philology: tradition versus innovation" (Bozzi, Cignoni, Lebrave c.s.).

progetti che si muovono in questa direzione. La strada appare tracciata: è necessario ora proseguire con convinzione per consolidare quello che è stato ottenuto fino ad oggi.

Edda Bresciani, Angiolo Menchetti
Dipartimento di Scienze Storiche del Mondo Antico
Università degli Studi di Pisa

Andrea Bozzi
Istituto di Linguistica Computazionale
CNR – Pisa

Giuseppe Fedele
Dipartimento di Elettronica, Informatica e Sistemistica
Università degli Studi della Calabria

BIBLIOGRAFIA

Bennani Y. 1999, *Adaptive weighting of pattern features during learning*, in *Proceeding of the IEEE*, 87, 3008-3013.

Bokser M. 1992, *Omnidocument technologies*, in *Proceeding of the IEEE*, 80, 7, 1066-1078.

Bozzi A. 1997, *Better Access to Manuscripts and Browsing of Images. Aims and Results of an European Research Project in the Field of Digital Libraries (BAMBI LIB-3114)*, Bologna, CLUEB.

Bozzi A. 2000, *Computer-aided Recovery and Analysis of Damaged Text Documents*, Bologna, CLUEB.

Bozzi A. 2003, *Digital documents and computational philology: the Digital Philology System (DIPHILOS)*, in M. Veneziani (ed.), *Informatica e Scienze umane. Mezzo secolo di studi e ricerche*, Firenze, Olschki, 175-201.

Bozzi A., Cignoni L., Lebrave J.L. 2004, *Digital Technology and Philological Disciplines*, Pisa-Roma, IEPI.

Bozzi A., Raggioli A. 2003, *Tecnologia digitale negli Istituti culturali: un case study*, in S. Zoppi (ed.), *Itinerari multimediali umanistici*, Alessandria, Edizioni dell'Orso, 23-42.

Bresciani E., Pernigotti S., Betrò M.C. 1983, *Ostraka demotici da Narmuti*, I, Pisa, ETS.

Fedele G. 2002, *Restauro di documenti a stampa antichi per il riconoscimento automatico dei caratteri*, in *Escuela Interlatina de Altos Estudios en Linguistica Aplicada: Matemàticas y Tratamiento de Corpus (San Millàn de La Cogolla, 2000)*, Logroño, Fundación San Millàn de La Cogolla, 289-308.

Gallo P. 1997, *Ostraka demotici e ieratici dall'archivio bilingue di Narmouthis*, II, Pisa, ETS.

Glaubermann M.H. 1956, *Character recognition for business machine*, «Electronics», 29, 132-136.

Kohonen T. 1990, *The self-organising maps*, in *Proceeding of the IEEE*, 78, 9, 1464-1480.

Lo Z.P., Yu Y. 1992, *Derivation of learning vector quantization algorithm*, in *Proceeding of the IEEE*, 80, 561-566.

Pratt W.K. 1991, *Digital Image Processing – Second Edition*, Wiley-Interscience Publication, New York, John Wiley & Sons, Inc.

ABSTRACT

A project for a Demotic Inscriptions on Ostraka Database is being carried out in collaboration between ILC/CNR (Pisa), the Department of Electronic Engineering (Calabria University) and the Department of the Ancient World History (Egyptological section, Pisa University). The aim of the project is to analyse the digital colour images of demotic texts on ostraka (Medinet Madi, in Fayyum region) with the aid of computational tools. The module described in the paper is a neural component able to learn the graphical features of each demotic symbol, which has been previously segmented in the images thanks to a semiautomatic procedure. A specific neural network tries to recognize the text written in the images linking the symbols segmented within the ostraka images database to the correspondent symbols available on a virtual keyboard. The graphical interface is particularly useful for teaching and research activities on this type of archaeological documentation.