Features extraction dalle immagine di cellule della mucosa nasale

Vito Proscia✉

Abstract

Uno degli step determinanti dell'image analysis, specialmente in ambito biomedico, è l'estrazione delle features da immagini cellulari, la Citologia Nasale è una tecnica clinica utile alla diagnostica in ambito rinoallergologico, essa si basa sull'analisi delle immagini della mucosa nasale, in questo report presentremo tecniche di features extraction per meglio estrarre informazioni da queste immagini cellulari usando varie metodologie quali morphological extraction, texture extraction e color extraction, per poi testare i dati estratti allenando un classificatore con l'obiettivo di meglio catalogare le singole cellule fornendo un supporto al lavoro dei citologi.

1 Introduzione

1.1 Background

Le cellule sono l'unità strutturale e funzionale degli organismi viventi, il loro aspetto e la loro morfologia possono rispecchiare la natura biologica dell'organo e persino del corpo, in particolare la citologia nasale è una diagnostica dello studio delle rinopatie, malattie del naso e delle fosse nasali caratterizzato da ostruzione respiratoria, rinorrea (scolo di muco), prurito e starnuti.
I citologi sono coloro i quali si occupano del processo di estrazione e di analisi delle cellule con l'obiettivo di diagnosticare condizioni infiammatorie e agenti patogeni.

Con il recente sviluppo delle tecniche di image analysis è stato possibile contribuire al lavoro dei suddetti aiutandoli sia in fase di rilevazione che di classificazione delle singole cellule, la correttezza di questi modelli è in larga parte determinata dal processo di features extraction che, in questo caso, consiste nel carpire le informazioni rilevanti dalle immagini delle cellule.

1.2 Cellule della mucosa nasale

1.2.1 Estrazione

Inizialmente si dilata la narice con uno *Speculum* nasale, successivamente si preleva con il Rhinoprobe una quantità di materiale cellulare dalla mucosa nasale, infine si striscia il materiale su un vetrino e lo si colora, essendo trasparente, con *May Grumwald Giemsa*, un colorante tendente al viola.

1.2.2 Tipologie di cellule

In una mucosa nasale si possono trovare, nelle giuste proporzioni, diverse popolazioni di cellule, tra quelle che compongono il dataset fornito possiamo distinguere:

- *Ciliate*: cellule epiteliali più comuni, responsabili del movimento del muco, sono caratterizzate dal nucleo situato ad un'altezza variabile dal fondo della cellula e dalle "ciglia" situate in alto;
- *Mucipare*: responsabili della produzione di muco, l'insieme di queste cellule ha la funzione di creare un flusso costante di fluido che si scarica nella faringe;
- *Neutrofili*: caratterizzate da più nuclei legati da sottilissimi filamenti di materiale nucleare all'interno del citoplasma;
- *Mastociti*: cellule immunitarie rotondeggianti od ovali composti da granuli circondati da membrana, intervengono nella genesi delle reazioni allergiche, di ipersensibilità e anafilattiche;
- *Linfociti*: analizzano i germi che penetrano attraverso il naso memorizzandoli per poi preparare la produzione di anticorpi specifici, caratterizzate dall'avere un nucleo molto grande che ricopre quasi interamente la superficie della cellula stessa;
- *Granulociti eobsinofili*: tipo di globuli bianchi che hanno funzione di difesa dell'organismo, caratterizzati dall'avere due nuclei collegati tra loro;
- *Metaplastiche*: espressione di infiammazione cronica;
- *Batteri*: indice di un'infiammazione batterica in corso.

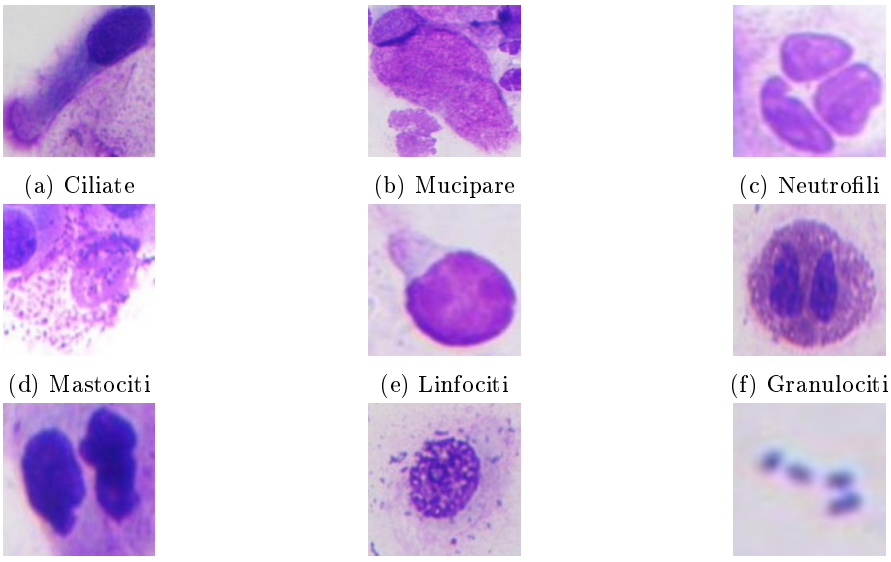
2 Materiali e metodi

In questa sezione approfondiremo come si è andati ad elaborare il dataset ed i successivi metodi di features extraction, per poi effettuare un test delle features ottenute andando ad allenare, su queste, un classificatore con l'obiettivo di discriminare il tipo delle cellule. [1]

2.1 Image processing

Originariamente il dataset ricevuto era composto da una serie di immagini di cellule, ogni immagine contava più cellule insieme, tutte annotate con le rispettive coordinate e classi di appartenenza.
Per i nostri scopi si è andati ad isolare ogni singola cellula dalle immagini originali, per poi prodourre un file che contenesse i nomi delle immagini delle singole cellule e le loro classi corrispondenti.

[1] Link sperimentazione: https://github.com/Giut0/Nasal-Cell-Feature-Extraction



(a) Ciliate (b) Mucipare (c) Neutrofili

(d) Mastociti (e) Linfociti (f) Granulociti

(g) Metaplastiche (h) Epiteliali (i) Batteri

Figure 1: Differenti tipologie di cellule nel dataset

2.2 Features extraction

Per caratterizzare ogni singola cellula isolata è necessario estrarre, da queste, delle informazioni, per poi usare il tutto allenando un classificatore.
In questo caso si è pensato di combinare tre differenti tipologie di features:

2.2.1 Feature morfologiche

Le caratteristiche morfologiche includono dimensioni, forme e bordi delle cellule, si sono ottenute convertendo ogni immagine in bianco e nero, successivamente trovando i contorni (massimo 2 in questo caso essendo cellule isolate quindi cellula stessa e nucleo) ed infine per ogni contorno trovato si salvano sia l'area che il perimetro.

2.2.2 Feature di texture

Le caratteristiche della texture descrivono le variazioni nei livelli di grigio all'interno dell'immagine, per questo scopo si sono usate le matrici di co-occorrenza a livello di grigio (GLCM) [1], strumento per l'analisi delle texture di un'immagine in scala di grigi, misura la frequenza di coppie di pixel con valori d'intensità specifici (livelli di grigio) ad una data distanza e angolo (nel nostro caso specifico la distanza di valutazione è stata impostata a 1 e l'angolo a 0°), le informazioni ottenibili con questo strumento sono :

- **Contrasto**: indica quanto variano i livelli di grigio tra i pixel vicini;

$$\sum_{i,j=0}^{levels-1} P_{i,j}(i-j)^2$$

- **Dissimilarità**: simile al contrasto, ma tiene conto solo della differenza assoluta tra i livelli di grigio dei pixel vicini, senza elevarla al quadrato;

$$\sum_{i,j=0}^{levels-1} P_{i,j}|i-j|$$

- **Omogeneità**: misura quanto i pixel vicini sono simili in termini di intensità;

$$\sum_{i,j=0}^{levels-1} \frac{P_{i,j}}{1+(i-j)^2}$$

- **Energia**: misura la quantità di ordine e ripetitività nella texture, valori elevati di energia indicano una texture altamente strutturata e ripetitiva;

$$\sqrt{\sum_{i,j=0}^{levels-1} P_{i,j}^2}$$

- **Correlazione**: misura quanto una coppia di pixel varia in modo correlato rispetto alla media.

$$\sum_{i,j=0}^{levels-1} P_{i,j} \left[\frac{(i-\mu_i)\ (j-\mu_j)}{\sqrt{(\sigma_i^2)(\sigma_j^2)}} \right]$$

2.3 Feature di colore

Le caratteristiche del colore sono importanti per descrivere le proprietà visive delle immagini delle cellule, in questo, caso per ogni immagine, si è andati a salvare l'istogramma dei colori [2].

2.4 Classificazione

Le features estratte sono state successivamente testate allenando il modello **Random Forest**, modello *ensambled* di apprendimento automatico supervisionato che combina molteplici alberi decisionali, ognuno dei quali addestrato su un subset casuale dei dati, per migliorare le previsioni.

Dopo varie sperimentazioni si riporta la migliore configurazione del modello:

1. `n_estimators = 200`, rappresenta il numero di alberi su cui lavorare;
2. `criterion = "gini"`, criterio di *split* degli alberi, in particolare il criterio punta a minimizzare l'impurità dei nodi, quindi per scegliere l'attributo che meglio divide i dati in base alla target feature per ogni nodo si calcola:

$$GINI(v) = 1 - \sum_{i=1}^{|C|} p_i^2$$

dove:

- v è il nodo in esame;
- C rappresenta l'inseme delle classi;
- p_i è la probabilità che un campione nel nodo v appartenga alla classe i.

infine vene scelto il nodo con minor impurità.

L'indice `gini` è stato scelto per la sua efficenza, infatti la complessità per il calcolo dell'indice per uno nodo è dell'ordine di $O(c)$, dove c è il numero di classi.

3 Risultati

Per andare a valutare il modello ottenuto si sono scelte le metriche più comuni:

Metrica	**Valore**
Accuracy	0.800
Precision	0.782
Recall	0.800

Table 1: Risultati della classificazione

4 Conslusioni

L'obiettivo principale della sperimentazione è l'estrazione di features da immagini cellulari della mucosa nasale, in primo luogo siamo andati ad isolare le singole cellule dalle immaigni utilizzando le coordinate fornite, successivamente c'è stata la fase di feature extraction andando a combinare tre differenti metodologie quali morphological extraction, texture extraction e color extraction ed infine abbiamo provato ad allenare un classificatore sui dati ottenuti raggiungendo dei buoni risultati.
Tuttavia, ci sono ancora ampi margini di miglioramento che possono aiutare il modello ad effettuare delle migliori predizioni, magari provando a combinare il tutto con altre tecniche come la trasformata di Fourier, oppure intraprendendo la strada del deep learning.

References

[1] Qing Liu and Xiping Liu. Feature extraction of human viruses microscopic images using gray level co-occurrence matrix. In *2013 International Conference on Computer Sciences and Applications*, pages 619–622. IEEE.

[2] Khalid Baker, Rakan Rashid, Nashat Salih, Abdulkarim Alsandi, Omar Farook Mohammad, Baker, and Nashat Alsandi. Classification of image blood cancer by using multi-training rnn. *Turkish Online Journal of Qualitative Inquiry*, 12:1065–10475, 07 2021.