Relazione Progetto AI

Francesco Fantechi

3 Gennaio 2021

1 Obbiettivo:

L'obbiettivo di questo esperimento é quello di testare e confrontare i classificatori Random Forest e Naive Bayes nel riconoscimento di pelle presente in dei video.

2 Caratteristiche degli algoritmi:

2.1 Random Forest:

Random Forest è un classificatore che utilizza un voto di maggioranza per predire una decisione. Dato un insieme di N esempi con p caratteristiche ciascuno, l'algoritmo genera B Alberi di Decisione ognuno dei quali addestrati su un sottoinsieme con $n \leq N$ esempi presi casualmente e con possibili ripetizioni dall'insieme di partenza. Inoltre, per ogni nodo terminale vengono scelte casualmente $m \ll p$ caratteristiche dalle quali estrarre la migliore per poterlo dividere ulteriormente. Per classificare un nuovo oggetto da un vettore in input, l'algoritmo sottopone tale vettore a tutti gli alberi della foresta. Ognuno di essi esprime una propria decisione e l'algoritmo sceglie poi come classificazione quella maggiormente votata.
Addestrando gli alberi su un numero di esempi e attributi diversi, Random Forest riduce la correlazione fra gli alberi e quindi l'impatto che la varianza ha solitamente su di essi e il rischio di overfitting sui dati.

2.2 Naive Bayes:

Naive Bayes è un classificatore basato sul Teorema di Bayes. Viene definito "ingenuo" per la sua assunzione che ogni caratteristica sia indipendente dalle altre data la classe. Quindi, data una classe Y e un esempio costituito su n attributi $x_1 \dots x_n$, la probabilitá della classe dato l'esempio si riduce a:
$P(Y \mid x_1 \dots x_n) = \prod_j P(x_j \mid Y) \cdot P(Y)$.
Gaussian Naive Bayes è un'estensione del classificatore Naive Bayes usata maggiormente quando gli attributi dei dati hanno valori appartenenti ad un dominio continuo. Come intuibile dal nome, esso considera tali valori continui come distribuiti secondo una distribuzione normale. Utilizza quindi la funzione di densitá

di probabilitá gaussiana, attraverso la media e la deviazione standard per ogni attributo e per ogni classe, per calcolare i vari likelihood $P(x_j \mid Y)$.

3 Procedimento sperimentale:

Per addestrare e testare i due classificatori, il training e il test set sono stati generati a partire da alcuni video presi dal dataset online di Julian Stöttinger, $https://feeval.org/Data-sets/Skin\_Colors.html$. Questo dataset contiene sia video grezzi che i corrispettivi video giá elaborati dove la pelle é evidenziata in bianco e il restante contenuto in nero. Per ogni coppia di video elaborato/grezzo ne sono stati salvati i frame e, per ognuno di essi, ne sono stati elaborati i pixel per generare i due set. Per creare gli esempi, i pixel grezzi sono stati portati dallo spazio colore RGB in scala $[0, 255]$ prima in scala $[0, 1]$ e poi nello spazio colore IHLS attraverso le opportune trasformazioni. Ogni esempio associato ad un pixel risulta quindi una tupla contenete le $p = 3$ variabili a valori continui di tonalitá, luminanza e saturazione. I corrispettivi pixel elaborati sono stati inseriti nei set come classe "1" se identificati come pelle e quindi bianchi e come classe "0" altrimenti. Nel nostro esperimento sono state utilizzate 3 coppie di video ("3.avi", "3_gt_202frames.avi", "4.avi", "4_gt_108frames.avi", "6.avi", "6_gt_213frames.avi") e gli esempi da essi generati sono stati ripartiti il 30% nel test set e i restanti nel training set ottenendo $N = 28152320$ esempi per l'addestramento.
Nel nostro esperimento Random Forest esegue il bootstrap generando 10 alberi di decisione ognuno dei quali addestrato su un insieme di N esempi presi casualmente dall'intero set di training e prendendo $m = \sqrt{p}$ attributi per meglio dividere ogni nodo terminale. Gli alberi della foresta utilizzano come misura di impuritá il criterio di Gini.
Dopo essere stati addestrati, i classificatori Random Forest e Gaussian Naive Bayes sono stati testati predicendo le possibili classi degli esempi appartenenti al test set. Per ogni classificatore é stata poi generata la matrice di contingenza corrispondente confrontando la propria predizione con le vere classificazioni degli esempi di test. Infine, con i dati di tali matrici sono stati calcolati i valori di precisione, di richiamo e di accuratezza dei due classificatori e il loro valore di F-score. I risultati dell'esperimento sono riportati nella sezione "Tabelle e Grafici dei risultati".

4 Tabelle e Grafici dei risultati:

Prediction/True_Class	Skin	Background
Skin	929126	1107474
Background	659928	9368752

Figure 1: Matrice di Contingenza del classificatore Gaussian Naive Bayes.

Prediction/True_Class	Skin	Background
Skin	1104243	321610
Background	484811	10154616

Figure 2: Matrice di Contingenza del classificatore Random Forest.

Classifier	Precision	Recall	Accuracy
Gaussian_Naive_Bayes	0.456	0.585	0.854
Random_Forest	0.774	0.695	0.933

Figure 3: Statistiche dei classificatori.

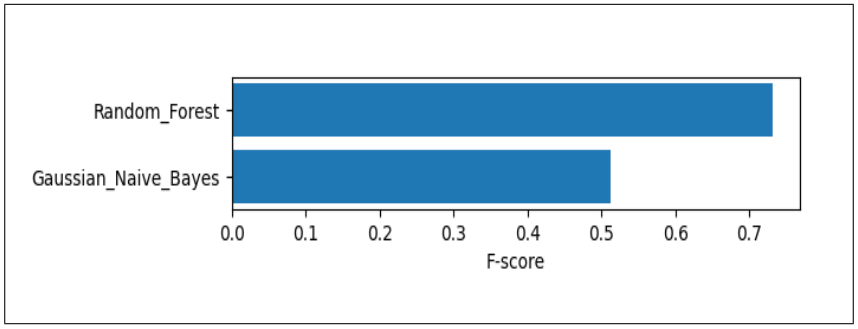

Figure 4: F-score dei classificatori.

5 Analisi dei risultati:

Guardando i risultati ottenuti nella figura 3 notiamo che il classificatore Random Forest é risultato piú accurato nel classificare gli esempi di test rispetto al classificatore Gaussian Naive Bayes. Questo é probabilmente dovuto al voto di maggioranza che esso utilizza per prendere una decisione e al fatto di non fare assunzioni "ingenue" sull'indipendenza degli attributi data la classe.
Nonostante ció, anche l'accuratezza ottenuta con il classificatore Gaussian Naive Bayes risulta piuttosto elevata. Come si puó notare in figura 1, questo risultato é dovuto all'alto numero di veri negativi da esso individuati rispetto agli altri parametri. Lo stesso non si puó dire peró dei veri positivi predetti che risultano nettamente inferiori rispetto ai falsi positivi e dello stesso ordine di grandezza rispetto ai falsi negativi con conseguenti valori di precisione e di richiamo ridotti.
Come é possibile apprezzare in figura 2, Random Forest presenta invece un alto numero di veri positivi e di veri negativi predetti rispetto agli altri parametri, ottenendo infatti valori di precisione e di richiamo modesti. Questo divario fra i due classificatori si puó notare maggiormente se si confrontano i loro F-score come riportato nell'istogramma di figura 4.
In conclusione, nonostante saremmo riusciti forse ad ottenere risultati piú dettagliati aumentando il numero di esempi per addestrare i nostri classificatori, possiamo affermare che Random Forest risulti un classificatore superiore rispetto a Naive Bayes predicendo un buon numero di classi vere e un numero ridotto di classi false.