Tecniche per il riconoscimento morfologico di una galassia

Carlo Cabras

25 agosto 2018

Sommario

Il riconoscimento morfologico di una galassia è uno degli elementi fondamentali per comprendere come le galassie si formino e si evolvano.

Ho testato diverse tecniche per cercare di determinare in automatico la forma di una galassia, distinguendo tra ellittica, irregolare ed a spirale. Per fare ciò, ho estratto feature dalle immagini sia tramite BoVW che tramite CNN, usando tecniche supervisionate come kNN e SVM, dove il risultato migliore riguarda l'utilizzo di un dataset contenente immagini aumentate e la classificazione delle tre categorie con un'accuratezza pari a 0,9796.

Nella sez. 1 introduco la classificazione morfologica delle galassie, nella sez. 2 discuto brevemente dei risultati presenti in letteratura, nella sez. 3 descrivo il dataset usato, nelle sez. 4 e 5 espongo le tecniche utilizzate per l'estrazione delle feature ed i classificatori usati, nella sez. 6 parlo del lavoro svolto ed infine nella sez. 7 sono presenti le conclusioni.

La repository contenente il codice scritto da me la si può trovare al seguente link.

1 Introduzione

Le galassie sono tra gli oggetti più importanti dell'intero universo. Alcune di esse contengono semplicemente stelle e altre contengono sia stelle che gas neutro o ionizzato, polvere, raggi cosmici e campi magnetici. Le galassie possono raggrupparsi formando dei *cluster* ed il centro galattico è caratterizzato da un'intensa luminosità[2].

Per comprendere al meglio una galassia, è importante poterla classificare in base alla sua forma, in quanto ci può dare informazioni sui processi interni ed esterni che hanno portato alla sua modellazione[3]. Si possono classificare in quattro macrogruppi:

- ellittiche: si presentano come un gruppo concentrato di stelle a forma di ellisse, dove la densità di stelle diminuisce man mano che ci si allontana dal centro (Fig. 1a);
- a spirale: la loro caratteristica è la presenza di bracci che partono dal centro e si estendono verso l'esterno, a forma di spirale (Fig. 1b);
- a spirale barrata[1]: come le precedenti, ma il centro galattico è caratterizzato da una "barra" (Fig. 1c);
- irregolari: senza forma precisa (Fig. 1d).

(a) Elittica. (b) A spirale. (c) A spirale barrata. (d) Irregolare.

Figura 1: Tipi di galassie.

È importante notare che le informazioni sulla forma dipendono dalla distanza, dalla luminosità e dalla grandezza della galassia e non sempre si hanno immagini che permettono di distinguere la forma oppure può apparire puntiforme, come nel caso dei quasar.

[1]Si ritiene che la nostra galassia sia così.

2 Altri lavori

Le prime parti del lavoro riguardavano la ricerca dei lavori presenti in letteratura e di un dataset. Espongo qua i diversi lavori che ho trovato.

Schutter e Shamir [5] analizzano una serie di similarità tra i tipi morfologici ed automaticamente deducono una sequenza morfologica di galassie. L'analisi è basata su tecniche di computer vision che calcolano le similarità visuali tra i diversi tipi morfologici.

Shamir [6] utilizza Nearest Neighbour su insieme di feature estratte dalle immagini, cercando di distinguere tra ellittica, a spirale e "vista di taglio", con un'accuratezza del $\sim 90\%$ rispetto a quanto rilevato dall'autore.

Sanchez et. al. [7] tramite una CNN, catalogano un database di 670'000 galassie, fornendo il più grande catalogo esistente e raggiungendo un'accuratezza ($> 90\%$).

Dieleman et. al. [8] con un'accuratezza quasi perfetta ($> 99\%$) riescono a distinguere una grande collezione di immagini tramite il deep learning.

Elfattah et. al. [9] Utilizzano momenti statistici invarianti a rotazione e posizione della galassia rispetto all'immagine, dimostrandosi delle buone proprietà e risultando in un'accuratezza di circa 90%.

El Aziz et. al. [10] con questo metodo riescono a trovare l'immagine più simile a quella in esame, confrontando poi la similarità tra le feature delle due immagini, ottenendo risultati migliori rispetto ad altre tecniche da loro confrontate.

Dhami et. al. [11] propongono una pipeline di estrazione feature e classificazione, mostrando buoni risultati.

3 Dataset

Il dataset usato è EFIGI (Extraction de Formes Idealisées de Galaxies en Imagerie)[4], composto da 4458 immagini di galassie dalla risoluzione di 255x255. Contiene immagini scattate in diverse lunghezze d'onda, che vanno dall'infrarosso all'ottico all'ultriavioletto; ho usato le immagini scattate nell'ottico, quindi 255x255 RGB. Di ogni immagine sono disponibili tante informazioni, tra cui la classificazione morfologica: è stato così possibile allenare gli algoritmi con questo dataset.

Inizialmente il dataset è stato diviso in 4 categorie, ovvero: 741 Barred Spiral, 422 Elliptic, 348 Irregular e 2947 Spiral. Successivamente è stato diviso in 3, cioè: 1104 Elliptic, 338 Irregular e 3014 Spiral.

4 Tecniche

4.1 Bag of Visual Words

Questa tecnica è una rivisitazione della "Bag of Words" dove un testo è rappresentato come una collezione non ordinata di parole e viene categorizzato in base a quelle che contiene. Ci sono tre elementi: un vocabolario, un istogramma normalizzato che rappresenta i documenti ed un classificatore.

Nella Bag of Visual Words un'immagine viene trattata come un documento e le feature estratte da essa sono considerate come *parole visuali*, così l'immagine viene rappresentata come una collezione di esse. I passi prevedono il rilevamento delle feature ed il calcolo dei descrittori, la creazione di un vocabolario visuale e la rappresentazione dell'immagine.

Una volta costruito il vocabolario visuale (la tecnica più utilizzata è quella del clustering attraverso l'algritmo kmeans) si costruisce l'istogramma (che rappresenterà l'immagine) dove per ogni feature si trova la parola visuale più vicina e si contano il numero di occorrenze, dopodiché si normalizza.

4.2 Convolutional Neural Networks

Una rete neurale è un sistema di calcolo concepito per simulare il cervello umano, cioè un'organizzazione con nodi (neuroni) che funzionano come unità di calcolo collegate fra loro tramite collegamenti pesati che indicano quanto è forte la connessione. La rete apprende variando l'intensità dei pesi. L'unità base è il *percettrone* che calcola la somma pesata del vettore di input e come

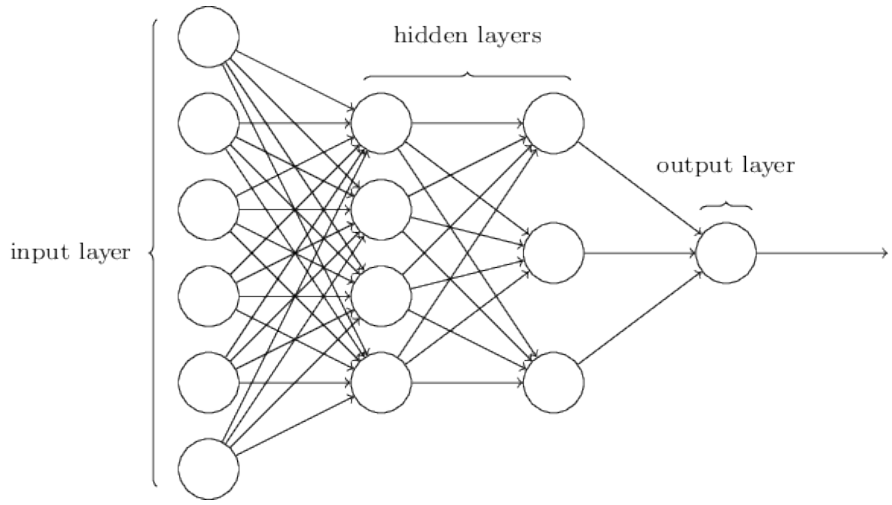


Figura 2: Rete neurale multistrato.

output dà una risposta binaria. Esistono poi le *reti neurali multistrato* che sono una generalizzazione del percettrone: esse sono formate da più livelli (fig. 2) permettono di avere una frontiera decisionale non lineare oltre a permettere una modellazione gerarchica dei livelli e l'estrazione di feature più complesse e ad alto livello.

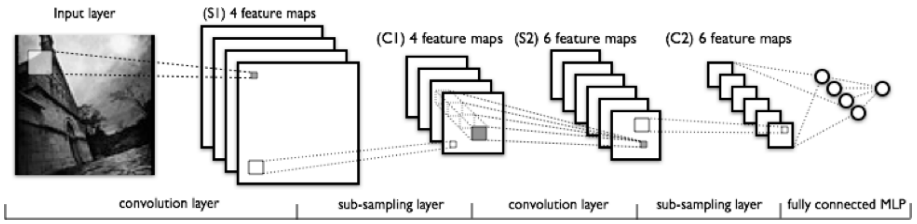


Figura 3: Rete neurale convoluzionale.

Le reti neural convoluzionali (fig. 3) sono un tipo di reti neurali specializzate nell'analizzare dati sotto forma di griglia, come ad esempio un'immagine 2D. Il termine "convoluzionale" deriva dal fatto che in qualche livello della rete è presente l'operazione di convoluzione anziché una semplice moltiplicazione matriciale. In una rete neurale normale, si collegherebbe ogni singolo pixel ad ogni singolo neurone, senza tener conto del fatto che la prossimità dei pixel è fortemente collegata alla loro similarità, perciò quello che si fa in una rete neurale convoluzionale è eliminare le connessioni tra pixel che sono poco significative (ad esempio connessioni tra pixel lontani tra loro), perciò ogni neurone è collegato ad una regione dell'immagine anziché al singolo pixel, rendendo il carico meno pesante[12]. Con l'avanzare dei livelli si applica la convoluzione e si fa un campionamento. Un modo per interpretare il passaggio da un livello all'altro

Sono utilizzate sia nel riconoscimento di immagini che in altri campi come ad esempio il processing del linguaggio naturale.

5 Classificatori

5.1 kNN

Classificatore basato sulle istanze, non induce un modello ma memorizza l'intero training set ed effettua una classificazione per analogia. Quando bisogna classifcare una nuova istanza, cerca un esempio il più possibile simile, tramite il concetto di *distanza*: si prendono i k esempi più vicini ed al nuovo record è assegnata la classe prevalente.

5.2 SVM

L'idea di base che sta dietro il Support Vector Machine è quella di trovare un iperpiano che meglio separi le due classi di un problema binario. Per classifcare un'istanza, si valuta la sua posizione

rispetto alla frontiera decisionale. Se la frontiera non è linare, si utilizza un kernel che simuli una trasformazione in uno spazio tale che le due classi possano essere linearmente separabili. Questo problema che di natura è binario, è facilmente estensibile al caso multiclasse decompondolo in sottoproblemi binari del tipo one-versus-rest o one-versus-one.

6 Svolgimento

6.1 handsonbow

La prima tecnica che ho testato, è stata "handsonbow", che utilizza l'approccio Bag of Visual Words ed una serie di classificatori Nearest Neighbours e Support Vector Machine.

I primi test sono stati effettuati sulle immagini originali cercando di distinguere tra quattro categorie: ellittiche, irregolari, a spirale ed a spirale barrata. Ho lanciato il programma dopo alcune modifiche per poter leggere i file `.png` e dopo aver separato le immagini per categoria in diverse sottocartelle. Estraendo feature DSIFT i risultati sono stati un'accuratezza del 43% per quanto riguarda il classificatore NN e 56,5% per il SVM.

Il prossimo passo è stato quello di filtrare le immagini ed ho provato diversi filtri, nella tabella 1 è mostrata l'accuratezza in base ai diversi filtri e classificatori usati, usando 30 immagini per il training set e 50 per il test set.

	nessun filtro	median 3x3	median 4x4	minimum 3x3
NN L2	0,43	0,495	0,485	0,42
SVM linear	0,565	0,585	0,56	0,52
SVM intersection kernel	/	/	/	0,555
SVM chi2 kernel	/	/	/	0,58

Tabella 1: Accuratezza in base ai diversi filtri e classificatori usati. Le feature estratte sono le DSIFT e sono state usate 30 immagini per il training e 50 per il test, come da default.

Il filtro che dava risultati migliori è il media 3x3, quindi d'ora in avanti tutti i test saranno condotti su immagini filtrate con quel filtro. Nella tabella 2 sono mostrati i risultati relativi a diversi classificatori testati su un diverso numero di immagini, estraendo le feature DSIFT.

	100 training 50 test	80% training 20% test
NN L2	0,545	0,494
NN chi2	0,565	0,512
SVM linear	0,59	0,4851
SVM linear LLC	0,62	0,493
SVM linear kernel	0,59	0,4851
SVM intersection kernel	0,585	0,5352
SVM chi2 kernel	0,61	0,5384

Tabella 2: Accuratezza in base al numero di immagini usate ed ai diversi classificatori. Le feature estratte sono le DSIFT.

Ho poi testato le feature SIFT e MSDSIFT mostrate in tabella 3, queste ultime ho avuto difficoltà a portare a termine l'operazione per l'elevata memoria utilizzata, tanto da saturare la RAM e rendere impossibile l'estrazione delle feature per tutte le immagini.

Visti i bassi risultati l'idea era quella di aumentare artificialmente il dataset, però dati i problemi di memoria e la relativa complessità del codice, sono passato all'utilizzo delle CNN.

	SIFT 30 training 50 test	MSDSIFT 30 training 50 test	SIFT 80% training 20% set	MSDISFT 20% training 5% test
NN L2	0,29	0,51	0,5084	0,6396
NN chi2	0,305	0,54	0,4972	0,6126
SVM linear	/	0,585	/	0,7117
SVM linear LLC	/	0,585	/	0,6081
SVM linear kernel	0,315	0,59	/	0,7117
SVM intersection kernel	0,43	0,62	/	0,6847
SVM chi2 kernel	0,435	0,645	/	0,7117

Tabella 3: Accuratezza riguardante le feature SIFT e MDSIFT, nelle prime due colonne le immagini sono state scelte casualmente. Le immagini sono le filtrate con median 3x3.

6.2 CNNs

Il prossimo passo è stato quello di testare le feature estratte tramite CNN, in particolare quelle estratte tramite AlexNet, usando il codice e la rete già allenata presente in MATLAB. Ho usato il codice presente nel tutorial[2]. Inizialmente però non mi ero accorto che la rete presente fosse già allenata a distinguere tra categorie di immagini, pertanto provai ad allenarne una io ma ottenendo scarsi risultati. Quindi una volta avviata correttamente la rete AlexNet ho ottenuto i primi risultati, usando 348 immagini scelte casualmente tra le categorie e 30% per il training set, SVM lineare come classifcatore. I risultati sono riportati in tabella 4.

		attuale			
		Barred Spiral	Elliptic	irregular	Spiral
predetta	Barred Spiral	0.4016	0	0.0820	0.5164
	Ellitptic	0.0123	0.2623	0.0246	0.7008
	Irregular	0.1434	0.0041	0.7500	0.1025
	Spiral	0.3361	0.0123	0.0820	0.5697

accuratezza = 0,4959

Tabella 4: Primi risultati con la CNN, sono state usate 348 immagini per categoria filtrate con median 3x3 di cui il 30% per il training set.

Si nota subito come non riesca a distinguere correttamente tra spirale e spirale barrata e molte ellittiche sono classificate come a spirale.

Dato che le immagini sono scelte casualmente, ho lanciato una seconda volta l'algoritmo, i risultati sono mostrati in tab. 5.

		attuale			
		Barred Spiral	Elliptic	irregular	Spiral
predetta	Barred Spiral	0.7500	0.0410	0.1721	0.0369
	Ellitptic	0.0451	0.9262	0.0246	0.0041
	Irregular	0.1352	0.0246	0.8197	0.0205
	Spiral	0.5410	0.2172	0.1475	0.0943

accuratezza = 0,6475

Tabella 5: Seconda esecuzione della CNN, si notano risultati diversi rispetto alla prima (mostrata in tab 4), mostrando una forte dipendenza dalle immagini scelte.

Si hanno dei risultati migliori per quanto riguarda la classificazione delle spirali barrate, ellittiche ed irregolari. Molto male quelle a spirale dove solo il 9% sono classificate correttamente e più della metà sono classificate come spirali barrate.

L'ho poi lanciato usando tutte le immagini del dataset, risultando così in un dataset sbilanciato (741 Barred Spiral, 422 Elliptic, 348 Irregular, 2947 Spiral). Il dataset è ancora diviso usando il 30% per il training set. Ho eseguito due volte l'algoritmo ma i risultati erano molto simili, pertanto in tab. 6 sono mostrati soltanto quelli della prima esecuzione.

[2] `https://it.mathworks.com/help/vision/examples/image-category-classification-using-deep-learning.html`

		attuale		
	Barred Spiral	Elliptic	irregular	Spiral
Barred Spiral	0.0366	0	0.0193	0.9441
Ellitptic	0	0.0305	0.0237	0.9458
Irregular	0	0	0.4098	0.5902
Spiral	0.0199	0.0024	0.0233	0.9544

predetta

accuratezza = 0,3578

Tabella 6: CNN usando tutte le immagini del dataset, risultando così sbilanciato, di cui il 30% per il training set.

Lanciandolo invece usando l'80% delle immagini come training set, si hanno risultati leggermente migliori e mostrati in tab. 7.

		attuale		
	Barred Spiral	Elliptic	irregular	Spiral
Barred Spiral	0.0405	0.0270	0.1216	0.8108
Ellitptic	0	0.8452	0	0.1548
Irregular	0	0.0143	0.7143	0.2714
Spiral	0.0085	0.0951	0.0611	0.8353

predetta

accuratezza = 0,6088

Tabella 7: CNN usando 80% delle immagini per il training set.

Ho poi provato ad aumentare artificialmente il dataset, creando una copia di ogni immagine e ruotandola casualmente. I risultati (mostrati in tab. 8) sono stati pessimi, dove la dominanza della categoria spiral si nota ancora di più, visto lo sbilanciamento del dataset.

		attuale		
	Barred Spiral	Elliptic	irregular	Spiral
Barred Spiral	0.0034	0	0.0135	0.9831
Ellitptic	0	0	0.0059	0.9941
Irregular	0.0072	0	0.2374	0.7554
Spiral	0	0	0.0085	0.9915

predetta

accuratezza = 0,3081

Tabella 8: CNN usando un dataset aumentato artificialmente, creando una copia di ogni immagine (quindi ottenendo il doppio delle immagini per categoria) ed usando 80% per il training set.

Al fine di diminuire la variabilità dei risultati, ho iniziato ad adottare l'approccio leave-one-out cross-validation ed ho provato il classificatore kNN sia con distanza euclidea che con distanza chi2, applicato al dataset contenenti immagini aumentate.

Per la scrittura del codice kNN con approccio leave-one-out ho calcolato una matrice delle distanze tra le immagini, in modo che in posizione (i,j) avessi la distanza tra l'immagine di test i e l'immagine di training j. Ho poi ordinato queste distanze e preso gli indici delle k immagini più vicine: ciò è possibile perché gli indici sono gli stessi usati nella struttura `imds` di tipo `ImageDatastore`. Quindi di questi k ho estratto la classe prevalente e creato la matrice di confidenza. Nelle tabelle che seguono sono riportati i risultati ottenuti tramite kNN leave-one-out cross-validation.

Nella tab. 9 mostro i risultati ottenuti usando la distanza euclidea tra le feature estratte tramite la CNN AlexNet e k = 30.

Nelle tabelle 10, 11, 12 mostro i risultati usando la distanza chi2 con k=30, 10, 5. Con k alti si nota la dominanza della classe spiral, dovuta allo sbilanciamento del dataset. Al decrescere di k si hanno predizioni più varie ma accuratezza più bassa; con k alti si predicono soltanto classi spiral.

Al fine di eliminare lo sbilanciamento del dataset, ho provveduto a scegliere le 696 immagini più vicine al centroide e ad effettuare i test leave-one-out cross-validation con quelle immagini. Prima di testare le immagini kmeans, ho effettuato il test su 696 immagini casuali, in modo da porre in evidenza le differenza. Nelle tabelle 13 e 14 sono riportati i risultati di kNN e SVM nel caso di 696 immagini random mentre nelle tabelle 15 e 16 i risultati di kNN e SVM nel caso di 696 immagini kmeans.

Si nota subito come nel caso delle immagini kmeans i risultati siano migliori.

		attuale		
	Barred Spiral	Elliptic	irregular	Spiral
Barred Spiral	24	0	0	13
predetta Ellitptic	15	416	6	294
Irregular	23	5	234	70
Spiral	1420	423	456	5517

accuratezza = 0,6944

Tabella 9: risultati ottenuti tramite kNN leave-one-out cross-validation usando la distanza euclidea e k = 30, il migliore tra 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 100, 200. Feature estratte tramite AlexNet.

		attuale		
	Barred Spiral	Elliptic	irregular	Spiral
Barred Spiral	0	0	0	0
predetta Ellitptic	0	0	0	0
Irregular	0	0	0	0
Spiral	1482	844	696	5894

accuratezza = 0,6611

Tabella 10: kNN con k=30 e distanza chi2 applicato alle feature delle immagini appartenti al dataset aumentato.

		attuale		
	Barred Spiral	Elliptic	irregular	Spiral
Barred Spiral	61	24	28	203
predetta Ellitptic	6	1	6	29
Irregular	13	7	2	35
Spiral	1402	812	660	5627

accuratezza = 0,6383

Tabella 11: kNN con k=10 e distanza chi2 applicato alle feature delle immagini appartenti al dataset aumentato.

		attuale		
	Barred Spiral	Elliptic	irregular	Spiral
Barred Spiral	167	87	99	651
predetta Ellitptic	52	17	30	219
Irregular	65	35	15	197
Spiral	1198	705	552	4827

accuratezza = 0,5637

Tabella 12: kNN con k=5 e distanza chi2 applicato alle feature delle immagini appartenti al dataset aumentato.

		attuale		
	Barred Spiral	Elliptic	irregular	Spiral
Barred Spiral	377	12	67	238
predetta Ellitptic	93	666	42	181
Irregular	92	12	533	97
Spiral	134	6	54	180

accuratezza = 0,6307

Tabella 13: kNN con k=10 (il migliore tra 5:5:50) e distanza euclidea effettuato sulle feature estratte da AlexNet e sulle 696 immagini random.

		attuale		
	Barred Spiral	Elliptic	irregular	Spiral
Barred Spiral	369	46	94	228
predetta Ellitptic	30	552	14	109
Irregular	98	19	489	102
Spiral	199	79	99	257

accuratezza = 0,5988

Tabella 14: Linear SVM effettuato sulle feature estratte dalla CNN e sulle 696 random.

		attuale		
	Barred Spiral	Elliptic	irregular	Spiral
Barred Spiral	303	4	68	66
Ellitptic	29	642	53	52
Irregular	48	1	511	3
Spiral	316	49	64	575

predetta

accuratezza = 0,7295

Tabella 15: kNN con k=20 (il migliore tra 5:5:50) e distanza euclidea effettuato sulle feature estratte dalla CNN e sulle 696 immagini più vicine al centroide trovato tramite kmeans.

		attuale		
	Barred Spiral	Elliptic	irregular	Spiral
Barred Spiral	370	22	52	218
Ellitptic	26	606	15	69
Irregular	88	6	587	29
Spiral	212	62	42	380

predetta

accuratezza = 0,6979

Tabella 16: Linear SVM effettuato sulle feature estratte dalla CNN e sulle 696 immagini più vicine al centroide trovato tramite kmeans.

6.3 Divisione del dataset in tre categorie

Dato che i risultati non sono stati particolarmente soddisfacenti e le categorie Spiral e Barred Spiral continuano ad essere confuse, sono passato al riconoscimento di sole tre categorie: Elliptic, Irregular e Spiral, quest'ultima ottenuta dall'unione delle precedenti Barred Spiral e Spiral. Le immagini usate per il testing sono state ulteriormente processate prendendo soltanto la porzione centrale di esse, quella più significativa. Pertanto ora il dataset è composto da immagini filtrate con median 3x3 e presa soltanto la parte centrale di dimensioni 129x129 rispetto alle originali 255x255.

Nella ricerca – non andata a buon fine – di un nuovo dataset che avesse più immagini e che fosse bilanciato, mi sono accorto di un errore commesso durante la separazione delle immagini per categoria. Questo errore riguarda il tipo S0, un tipo intermedio tra ellittica ed a spirale, come mostrato in fig. 4.

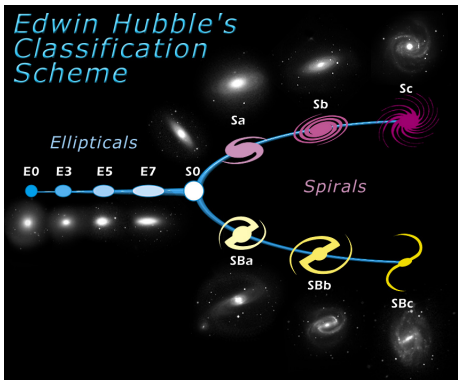


Figura 4: Classificazione delle galassie secondo Hubble.

Inizialmente le classificavo come Spirale, sarebbe invece più corretto classifcarle come Ellittiche. Ho così fatto i soliti test, cioè leave-one-out cross-validation sulle feature estratte tramite AlexNet e con i classificatori kNN (dove mostrerò il k migliore scelto tra 5:5:50) e linear SVM. È stato stato usato l'intero dataset. I risultati sono mostrati nelle tabelle 17 e 18, dove si può notare come i risultati migliorino.

		attuale		
		Elliptic	irregular	Spiral
predetta	Ellitptic	651	2	328
	Irregular	1	113	37
	Spiral	150	223	2951

accuratezza = 0,8337

Tabella 17: kNN su intero dataset, k=25, feature estratte tramite CNN da immagini filtrate con median 3x3 e presa la parte centrale 129x129.

		attuale		
		Elliptic	irregular	Spiral
predetta	Ellitptic	584	4	200
	Irregular	14	170	100
	Spiral	204	164	3016

accuratezza = 0,8461

Tabella 18: SVM su intero dataset, feature estratte tramite CNN da immagini filtrate con median 3x3 e presa la parte centrale 129x129.

Dopodiché, tramite un algoritmo kmeans, sono passato ad estrarre per ogni categoria le immagini più vicine al rispettivo centroide, in modo da usare le più significative. Nella tab. 19 mostro i risultati ottenuti con kNN nel caso di 338 immagini random, in modo da poter essere confrontati con quelli nel caso di 338 immagini kmeans delle tabelle 20 e 21.

I risultati mostrati in tab. 21 sono i migliori finora ottenuti.

Sono poi passato ad usare il dataset con immagini aumentate, stavolta soltanto relative alla classe Irregular che è quella che ne contiene meno, 338, passando quindi a 676 immagini Irregular.

		attuale		
		Elliptic	irregular	Spiral
predetta	Ellitptic	228	124	92
	Irregular	98	199	43
	Spiral	12	15	203

accuratezza = 0,6213

Tabella 19: kNN con 338 immagini random tra categorie, k = 40.

		attuale		
		Elliptic	irregular	Spiral
predetta	Ellitptic	331	13	24
	Irregular	0	278	0
	Spiral	7	47	314

accuratezza = 0,9103

Tabella 20: kNN su 338 immagini kmeans per categoria, k=15.

		attuale		
		Elliptic	irregular	Spiral
predetta	Ellitptic	325	5	13
	Irregular	0	321	5
	Spiral	13	12	320

accuratezza = 0,9527

Tabella 21: SVM su 338 immagini kmeans per categoria.

Per rimuovere il bias dovuto alla classificazione di un'immagine basata sulla propria copia, ho rimosso la corrispondete immagine aumentata dal training set.

Inizialmente, nelle tabelle 22 e 23 mostro i risultati usando 676 immagini kmeans per Elliptic e Spiral e 338 per Irregular.

		attuale		
		Elliptic	irregular	Spiral
predetta	Ellitptic	640	15	69
	Irregular	0	259	1
	Spiral	36	64	606

accuratezza = 0,8905

Tabella 22: kNN, 676 immagini kmeans per elliptical e spiral, 338 per irregular, k=10.

		attuale		
		Elliptic	irregular	Spiral
predetta	Ellitptic	619	6	46
	Irregular	5	309	19
	Spiral	52	23	611

accuratezza = 0,9107

Tabella 23: SVM, 676 immagini kmeans per elliptical e spiral, 338 per irregular.

I risultati sono leggermente peggiori rispetto al caso del semplice 338 kmeans per ogni categoria.

In tab. 24 mostro i risultati ottenuti usando 676 immagini casuali senza rimuovere dal training set la corrispondente immagine aumentata.

In tab. 25 mostro i risultati usando tutte le 676 della classe Irregular ed usando 676 kmeans delle classi Elliptic e Spiral.

Nelle tabelle 26 e 27 mostro i risultati basati sulle 676 immagini kmeans ma rimuovendo le corrispondenti immagini aumentate. In questo caso si ha un'accuratezza leggermente minore ma è risultato non affetto da bias.

Ho notato che alcune galassie ellittiche sono classificate come spirali e viceversa, mi sono accorto che erano classificate come S0: esse non sono né ellittiche né a spirale, ma una via di mezzo. Si

		attuale		
		Elliptic	irregular	Spiral
predetta	Ellitptic	632	138	153
	Irregular	21	457	153
	Spiral	23	81	370

accuratezza = 0,7194

Tabella 24: kNN, 676 immagini random per elliptical e spiral, tutte le 676 per irregular, k = 10.

		attuale		
		Elliptic	irregular	Spiral
predetta	Ellitptic	648	21	35
	Irregular	0	620	1
	Spiral	28	35	302

accuratezza = 0,9290

Tabella 25: kNN, 676 immagini dove elliptic/spiral prese con kmeans ed irregular augmented x2, senza rimuovere dal training la corrispondente immagine augmented (e viceversa), k=5.

		attuale		
		Elliptic	irregular	Spiral
predetta	Ellitptic	640	24	70
	Irregular	0	565	3
	Spiral	36	87	603

accuratezza = 0,8915

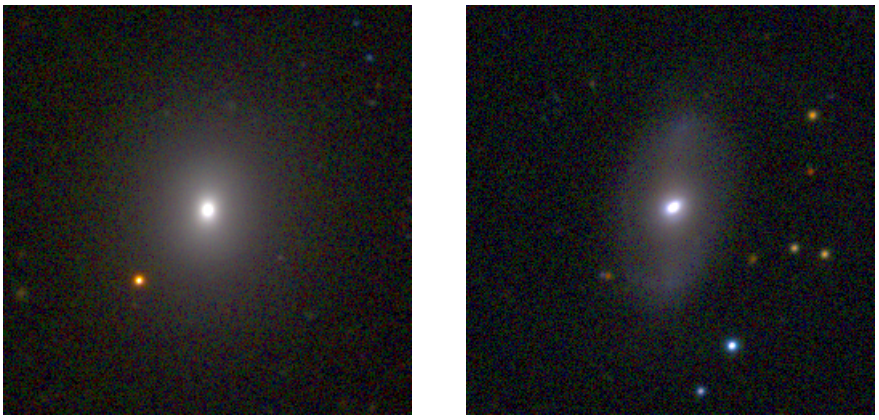
Tabella 26: kNN, 676 kmeans per elliptic/spiral e tutte le 676 irregular ma rimuovendo dal training la corrispondente immagine augmented (e viceversa), k = 10.

		attuale		
		Elliptic	irregular	Spiral
predetta	Ellitptic	620	11	57
	Irregular	6	635	28
	Spiral	50	30	591

accuratezza = 0,9103

Tabella 27: SVM, 676 kmeans per elliptic/spiral e tutte le 676 irregular ma rimuovendo dal training la corrispondente immagine augmented (e viceversa).

presentano come ellttiche, cioè con una forte prominenza centrale, per poi avere un disco di stelle e talvolta qualche struttura simile ad un braccio, come mostrato in fig. 5. Questa categoria inizialmente la classificavo come Spiral, mentre ora la classifico come Elliptic.



(a) Galassia S0 simile ad una galassia ellittica.

(b) Galassia S0 simile ad una galassia a spirale.

Figura 5: Due galassie S0, la seconda può essere confusa con una a spirale nonostante non presenti dei bracci evidenti.

Ho quindi condotto dei test senza prendere in considerazione questo genere di galassie. Nelle tabelle 28 e 29 mostro i risultati ottenuti tramite kNN e SVM sulle 287 immagini ottenute tramite kmeans rispetto alle feature estratte da AlexNet[3].

		attuale		
		Elliptic	irregular	Spiral
predetta	Ellitptic	278	2	0
	Irregular	9	280	0
	Spiral	0	5	287

accuratezza = 0,9814

Tabella 28: kNN sulle feature estrate tramite AlexNet, senza tenere in considerare le immagini S0, 287 immagini kmeans per categoria, k=5.

		attuale		
		Elliptic	irregular	Spiral
predetta	Ellitptic	278	3	0
	Irregular	9	284	0
	Spiral	0	0	287

accuratezza = 0,9861

Tabella 29: SVM sulle feature estrate tramite AlexNet, senza tenere in considerare le immagini S0, 287 immagini kmeans per categoria.

Ho effettuato test utilizzando le feature estratte dalla rete ResNet. Nelle tabelle 30 e 31 mostro i risultati ottenuti usando kNN e SVM per classificare le feature estratte da tutto il dataset.

		attuale		
		Elliptic	irregular	Spiral
predetta	Ellitptic	920	7	180
	Irregular	2	116	34
	Spiral	182	215	2800

accuratezza = 0,8609

Tabella 30: kNN su feature estratte tramite ResNet, dataset intero, k=25.

		attuale		
		Elliptic	irregular	Spiral
predetta	Ellitptic	899	6	183
	Irregular	12	176	132
	Spiral	193	156	2699

accuratezza = 0,8469

Tabella 31: SVM su feature estratte tramite ResNet, dataset intero.

Nelle tabelle 32 e 33 i risultati relativi al dataset con 336 immagini ottenute tramite kmeans; essi sono relativamente buoni ma inferiori a quelli ottenuti tramite feature estratte da AlexNet.

		attuale		
		Elliptic	irregular	Spiral
predetta	Ellitptic	336	4	2
	Irregular	0	238	19
	Spiral	0	94	315

accuratezza = 0,8819

Tabella 32: kNN su feature estratte da ResNet, dataset composto da 336 immagini kmeans, k=5.

[3]Durante i primi test in cui caricavo il dataset senza le immagini S0, non rimuovevo correttamente tutte le S0 per via di un bug nello script. Riporterò quindi soltanto i risultati corretti e ripetuti in un secondo momento.

		attuale		
		Elliptic	irregular	Spiral
predetta	Ellitptic	335	4	1
	Irregular	0	278	45
	Spiral	1	54	290

accuratezza = 0,8958

Tabella 33: SVM su feature estratte da ResNet, dataset composto da 336 immagini kmeans.

Infine nelle tabelle 34 e 35 i risultati col dataset senza immagini S0, 287 kmeans. Stavolta i classificatori non si confondono tra Elliptic e Spiral, ma bensì tra Irregular e Spiral, dove talvolta una galassia irregolare può presentare strutture simili a dei bracci ma non possono essere categorizzate come a spirale in quanto non sono "ordinati".

		attuale		
		Elliptic	irregular	Spiral
predetta	Ellitptic	273	1	0
	Irregular	9	210	20
	Spiral	5	76	267

accuratezza = 0,8711

Tabella 34: kNN su feature estratte tramite ResNet, dataset senza immagini S0 e 287 kmeans, k=10.

		attuale		
		Elliptic	irregular	Spiral
predetta	Ellitptic	275	3	0
	Irregular	7	237	46
	Spiral	5	47	241

accuratezza = 0,8746

Tabella 35: SVM su feature estratte tramite ResNet, dataset senza immagini S0 e 287 kmeans

Un altro tipo di test che ho effettuato è quello su BoVW su feature SURF, approccio presente in un tutorial di MATLAB[4]. Nelle tabelle 36 e 37 mostro i risultati ottenuti usando kNN e SVM su un dataset composto da 336 immagini kmeans.

		attuale		
		Elliptic	irregular	Spiral
predetta	Ellitptic	332	12	125
	Irregular	0	297	0
	Spiral	4	27	211

accuratezza = 0,8333

Tabella 36: kNN su feature SURF estratte tramite BoVW, dataset composto da 336 immagini kmeans per categoria, k=5.

		attuale		
		Elliptic	irregular	Spiral
predetta	Ellitptic	329	3	56
	Irregular	0	318	1
	Spiral	7	15	279

accuratezza = 0,9187

Tabella 37: SVM su feature SURF estratte tramite BoVW, dataset composto da 336 immagini kmeans per categoria.

Prendendo in mano i classificatori usati finora, ho testato anche sui vettori bof generati dal codice handsonbow, con feature sift e dsift: una volta estratti i vettori in base al vocabolario, ho

[4] `https://it.mathworks.com/help/vision/examples/image-category-classification-using-bag-of-features.html`

preso le 338 immagini kmeans per categoria più vicine al rispettivo centroide ed ho nuovamente estratto i descrittori, così poi da poterli analizzare con kNN e SVM.

Nelle tabelle 38 e 39 mostro i risultati relativi ai classificatari kNN e SVM applicati ai vettori estratti tramite BoVW su feature SIFT, il dataset è composto dalle 336 immagini kmeans per categoria.

		attuale		
		Elliptic	irregular	Spiral
predetta	Ellitptic	325	77	98
	Irregular	0	208	27
	Spiral	11	51	211

accuratezza = 0,7381

Tabella 38: kNN su feature SIFT estratte tramite BoVW dal codice handsonbow, dataset composto da 336 immagini kmeans per categoria, k=30.

		attuale		
		Elliptic	irregular	Spiral
predetta	Ellitptic	314	12	15
	Irregular	3	253	40
	Spiral	19	71	281

accuratezza = 0,8413

Tabella 39: SVM su feature SIFT estratte tramite BoVW dal codice handsonbow, dataset composto da 336 immagini kmeans per categoria.

Ho ripetuto i test per le feature DSIFT, i cui risultati sono mostrati nelle tabelle 40 e 41 e risultano migliori rispetto alle feature SIFT.

		attuale		
		Elliptic	irregular	Spiral
predetta	Ellitptic	336	27	1
	Irregular	0	220	15
	Spiral	0	89	320

accuratezza = 0,8690

Tabella 40: kNN su feature DSIFT estratte tramite BoVW dal codice handsonbow, dataset composto da 336 immagini kmeans per categoria, k=5.

		attuale		
		Elliptic	irregular	Spiral
predetta	Ellitptic	335	10	0
	Irregular	1	280	19
	Spiral	0	46	317

accuratezza = 0,9246

Tabella 41: SVM su feature DSIFT estratte tramite BoVW dal codice handsonbow, dataset composto da 336 immagini kmeans per categoria.

Per queste feature DSIFT, ho effettuato i test anche sul dataset senza S0: stavolta però ho ridovuto ricalcolare tutti i vettori finali in quanto si basavano sul dizionario formato dall'intero dataset. Il dataset usato è formato da 287 immagini kmeans per categoria, risultati mostrati nelle tabelle 42 e 43.

Per il dataset composto da immagini senza S0, ho potuto anche estrarre le feature MSDSIFT. È stato possibile farlo solo per questo dataset perché conteneva qualche immagine in meno ed il computer ha retto il carico di lavoro. Anche qua il test è stato effettuato tramite 287 immagini kmeans per categoria, risultati mostrati nelle tabelle 44 e 45. Essi sono buoni ma rimangono sulla linea dei precedenti buoni risultati ottenuti finora.

Infine ho aumentato il dataset fino a prendere in considerazione 3016 immagini per categorie: usando sempre l'approccio leave-one-out cross-validation, quando testavo un'immagine, nel training

		attuale		
		Elliptic	irregular	Spiral
predetta	Ellitptic	276	10	58
	Irregular	6	269	10
	Spiral	5	8	219

accuratezza = 0,8873

Tabella 42: kNN su feature DSIFT estratte tramite BoVW dal codice handsonbow, dataset composto da 287 immagini kmeans per categoria e senza immagini S0, k=5.

		attuale		
		Elliptic	irregular	Spiral
predetta	Ellitptic	273	2	7
	Irregular	6	275	5
	Spiral	8	10	275

accuratezza = 0,9559

Tabella 43: SVM su feature DSIFT estratte tramite BoVW dal codice handsonbow, dataset composto da 287 immagini kmeans per categoria e senza immagini S0.

		attuale		
		Elliptic	irregular	Spiral
predetta	Ellitptic	277	24	37
	Irregular	4	255	3
	Spiral	6	8	247

accuratezza = 0,9048

Tabella 44: kNN su feature MSDSIFT estratte tramite BoVW dal codice handsonbow, dataset composto da 287 immagini kmeans per categoria e senza immagini S0, k=5

		attuale		
		Elliptic	irregular	Spiral
predetta	Ellitptic	268	5	3
	Irregular	10	270	3
	Spiral	9	12	281

accuratezza = 0,9512

Tabella 45: SVM su feature MSDSIFT estratte tramite BoVW dal codice handsonbow, dataset composto da 287 immagini kmeans per categoria e senza immagini S0.

set di riferimento non prendevo in considerazione le corrispondenti immagini volte ad aumentare il dataset, in modo da evitare eventuali bias. In questo modo ho ottenuto i risultati migliori, mostrati nelle tabelle 46 e 47 dove ho testato i classificatori kNN e SVM sulle feature estratte da AlexNet.

		attuale		
		Elliptic	irregular	Spiral
predetta	Ellitptic	2961	176	39
	Irregular	48	2832	11
	Spiral	7	8	2966

accuratezza = 0,9681

Tabella 46: kNN su feature estratte da AlexNet, dataset aumentato fino ad ottenere 3016 immagini per categoria, k=10

		attuale		
		Elliptic	irregular	Spiral
predetta	Ellitptic	2941	98	6
	Irregular	73	2916	4
	Spiral	2	2	3006

accuratezza = 0,9796

Tabella 47: SVM su feature estratte da AlexNet, dataset aumentato fino ad ottenere 3016 immagini per categoria.

Nella tab. 48 mostro la matrice di confusione estratte dalle predizioni effettuate da SVM nell'ultimo test ma relative soltanto alle immagini non aumentate.

		attuale		
		Elliptic	irregular	Spiral
predetta	Ellitptic	1078	14	6
	Irregular	25	323	4
	Spiral	1	1	3006

accuratezza = 0,9886

Tabella 48: SVM relativo al dataset aumentato, matrice di confusione creata prendendo soltanto le immagini non aumentate.

7 Conclusioni

I risultati migliori sono stati quelli ottenuti analizzando le feature estratte tramite la rete neurale AlexNet, utilizzando un dataset dove sono presenti immagini aumentate fino ad avere 3016 immagini per ognuna delle tre categorie (Elliptic, Irregular e Spiral), un classificatore SVM con l'approccio leave-one-out cross-validation dove nel training set non sono state prese in considerazione le corrispondenti immagini aumentate del test set, ottenendo un'accuratezza pari a 0,9796, che diventa 0,9886 se prendiamo in considerazione la matrice di confusione estratta dalle predizioni delle immagini originali, cioè non aumentate.

Il lavoro si è trattato principalmente di provare nuove tecniche e cercare di muoversi verso una direzione che portasse a risultati migliori, analizzando quelli ottenuti.

In conclusione posso ritenermi soddisfato del lavoro effettuato ed ho capito come avviene un lavoro di questo tipo ed appreso nuove metodologie.

Elenco delle tabelle

Elenco delle figure

Riferimenti bibliografici

[1] Repository `https://github.com/carlocabras21/galaxy_morphology_recognition`

[2] H. Karttunen et.al. (2007) Fundamental Astronomy, fifth edition, ISBN 978-3-540-34143-7 5th Edition, Springer Berlin Heidelberg New York.

[3] Buta et. al. (2015) A Classical Morphological Analysis of Galaxies in the Spitzer Survey of Stellar Structure in Galaxies (S4G). arXiv:1501.00454v2 [astro-ph.GA]

[4] Baillard et. al. (2011) The EFIGI catalogue of 4458 nearby galaxies with detailed morphology arXiv:1103.5734v3 [astro-ph.CO]

[5] Andrew Schutter, Lior Shamir (2015) Galaxy morphology - an unsupervised machine learning approach. arXiv:1505.04876v2 [astro-ph.IM]

[6] Lior Shamir (2009) Automatic morphological classification of galaxy images. arXiv:0908.3904v1 [astro-ph.IM]

[7] H. Domínguez Sánchez, M. Huertas-Company, M. Bernardi, D. Tuccillo, J. L. Fischer (2017) Improving galaxy morphologies for SDSS with Deep Learning. arXiv:1711.05744v2 [astro-ph.GA]

[8] Sander Dieleman, Kyle W. Willett, Joni Dambre (2015) Rotation-invariant convolutional neural networks for galaxy morphology prediction. arXiv:1503.07077v1 [astro-ph.IM]

[9] Mohamed Abd Elfattah, Mohamed A. Abu ELsoud, Aboul Ella Hassanien, Tai-hoon Kim (2012) Automated Classification of Galaxies Using Invariant Moments. Future Generation Information Technology, pp. 103–111, Springer

[10] Mohamed Abd El Aziz, I. M. Selim, Shengwu Xiong (2017) Automatic Detection of Galaxy Type From Datasets of Galaxies Image Based on Image Retrieval Approach. Scientific Reports volume 7, Article number: 4463

[11] Devendra Singh Dhami, David Leake, Sriraam Natarajan (2017) Knowledge-Based Morphological Classification of Galaxies from Vision Features. The AAAI-17 Workshop on Knowledge-Based Techniques for Problem Solving and Reasoning. WS-17-12

[12] How Convolutional Neural Networks Accomplish Image Recognition? visitato il 25/08/2018.