Σύγκριση Ταξινομητών Για Κατηγοριοποίηση Ουράνιων Σωμάτων

Αθανάσιος Μιχαηλούδης

28 Φεβρουαρίου 2021

Σύνοψη Προβλήματος

Η πληθώρα πληροφορίας από παρατηρήσεις καθώς και οι τελευταίες εξελίξεις στον τομέα της αστρονομίας και της αστροφυσικής (βαρυτικά κύματα, φωτογραφίας μελανής οπής) καθιστούν σαφή την ανάγκη για ανάπτυξη μεθόδων ταχείας και ακριβούς μελέτης δεδομένων που θα ελαφρύνουν το βεβαρυμένο φόρτο εργασίας των ερευνητών. Με αυτή τη σκέψη, σκοπός της εργασίας είναι η ανάπτυξη κώδικα για την επεξεργασία δεδομένων που συλλέγονται από όργανα παρατήρησης (πχ τηλεσκόπια) και την κατηγοριοποίση των σωμάτων από τα οποία προέρχονται σύμφωνα με τα δεδομένα αυτά.

Δημοσιεύσεις/Αναφορές

Το πρόβλημα της κατηγοριοποίησης στην αστρονομία δεν είναι καινούριο αφού παρέχει σημαντική βοήθεια και εξοικονώμηση χρόνου και πόρων και συνεπώς έχει μελετηθεί εκτενώς, ειδικότερα τα τελευταία χρόνια.
Έχει χρησιμοποιηθεί πληρθώρα μεθόδων τόσο επεξεργασίας δεφομένων, όσο και ταξινόμησης πληροφοριών, με τις πρώτες να περιλαμβάνουν ανάλυση και καθαρισμό σημάτων στα διαφορετικά μήκη κύματος του ηλεκτρομαγνητικού φάσματος, την εξαγωγή μορφολογικών[1], χρωματικών κ.ά. χαρακτηριστικών[2] από φωτογραφίες και τις τελευταίες να περιλαμβάνουν την εφαρμογή μοντέλων μηχανικής μάθησης, επιβλεπόμενης και μη[3], όπως οι SVM, τα Random Forests (RF) ή τα Νευρωνικά Δίκτυα[4].

Πειραματική Διαδικασία

Στόχος της εργασίας

Στόχος της εργασίας είναι η σύγκριση και η εκπαίδευση μοντέλων SVM και RF και η εφαρμογή τους σε ένα σύνολο δεδομένων αποτελούμενων από εικόνες γαλαξιών και αστέρων για την αναγνώριση και ορθή κατηγοριοποήση των φωτογραφιών των δύο ουράνιων σωμάτων στην αντίστοιχη κλάση, μετά την κατάλληλη προεπεξεργασία των συνόλων δεδομένων. Αυτή η επιλογή σωμάτων έγινε λόγω της ομοιώτητας που φέρουν, εκ πρώτης όψεως, αφού, με γυμνό μάτι, αμφότερα φαίνονται ως δύο στρογγυλές, φωτεινές κουκίδες στο νυχτερινό ουρανό. Ωστόσο, με μια πιο προσεκτική ματιά, εύκολα διακρίνει κανείς τις διαφορές τους, όπως το πιο δισκοειδές σχήμα των γαλαξιών με τη μεγαλύτερη μάζα συγκεντωμένη στο κέντρο τους, σε αντίθεση με το πιο συμπαγές σχήμα των αστέρων και το πιο ¨αραιό' και διάχυτο φως των γαλαξιών, εφάμιλλο ενός φακού για τα γήινα δεδομένα, σε αντίθεση με το πιο συγκεντρωμένο κι

έντονο φως των αστέρων, που μπορεί να παρομοιαστεί με ακτίνα laser. Οι διαφορές αυτές φαίνονται και στις παρακάτω δύο εικόνες:
Το dataset που επιλέχθηκε για την εκπόνηση της εργασίας είναι το ***Stars and Galaxies:*** *https://www.kaggle.com/siddharthchaini/stars-and-galaxies* και αποτελείται συνολικά από 10000 εικόνες γαλαξιών και αστέρων (5000 +5000) εκ των οποίων 2000 (1000 + 1000) χρησιμοποιούνται σαν σύνολο δοκιμής.

Εξαγωγή Χαρακτηριστικών

Η εξαγωγή χαρακτηριστικών έγινε στοχεύοντας στις διαφορές των δύο ουράνιων σωμάρων, οι οποίες αναφέρθηκαν παραπάνω. Έτσι, αφού μειώθηκαν οι διαστάσεις των αρχικών εικόνων από 512×512 σε 128×128 για εξοικονόμηση επεξεργαστικής ισχύς του υπολογιστή, τα χαρακτηριστικά που εξήχθησαν από κάθε είκόνα ήταν δύο: 1) Το χρώμα του αντικειμένου, μέσω του ιστογράμματος χρώματος και 2) το σχήμα του ατικειμένου, μέσω των "hu momnents", που δίνει τα μορφολογικά χαρακρηριστικά του. Εκ των δύο, το χαρακτηριστικό που προβλέπεται ως σημαντικότερο είναι το χρώμα, καθώς σε αυτό περικλειται και η ένοια της έντασης του εκπεμπόμενου φωτός. Με το πέρας της ανωτέρω διαδικασίας, σε κάθε εικόνα πλέον αντιστοιχούν 512 χαρακτηριστικά χρώματος από το ιστόγραμμα και 7 από τις "hu momnents", ενώ εφαρμόστηκε και το κατάλληλο scaling στο σύνολο των χαρακτηριστικών, για την ευκολότερη εκπαίδευση των αλγορίθμων μάθησης. Αυτό έγινε με τον MinMaxScaler με σκοπό τη ¨δημιουργία' όσο το δυνατόν περισσότερων μηδενικών στο σύνολο των δεδομένων, αποσκοπόντας στην βέλτιστη εκπαιδευση του αλγόριθμου SVM.
Και με αυτή την ενέργεια δημιουργείθηκαν τα τελικά dataframes (τόσο εκπαίδευσης όσο και δοκιμής), τα οποία ήταν έτοιμα προς χρήση από τους ταξινομητές.

Εκπαίδευση

Ξεκινώντας την εκπαίδευση διεξήχθησε ένα benchmark τεστ για να εκτιμηθεί η απόδοση των δύο αλγορίθμων στο τεστ δεδομένων εκπαίδευσης. Έτσι, χρησιμοποιόντας την εντολή *TrainTestSplit* για τη δημιουργεία ενός σετ επικύρωσης ίσο με 20% του συνολικού τέστ εκπαίδευσης και τις προκαθορισμένες παραμέτρους των δύο μοντέλων (RF: n_estimators=100, criterion='gini' και SVM: C=1.0, kernel= 'rbf') προέκυψαν τα εξής αποτελέσματα:

- *RF:* Validation Accuracy = 92.1%
- *SVM:* Validation Accuracy = 92.2%

Από αυτό έγινε αντιληπτό ότι τα μοντέλα είναι ικανά να παράσχουν ικανοποιητική ακρίβεια πρόβλεψης, αλλά και απόδοση γενικότερα, οπότε αυτό που έγινε στη συνέχεια ήταν η αναζήτηση του αποδοτικότερου εκ των δύο με τις καταλληλότερες παραμέτρους.

(α) Γαλαξίας

(β)Αστέρας

Σχήμα 1: (α) Γαλαξίας: Διάχυτο φως, μικρή ένταση, εξασθενεί προς τα όρια. (β) Αστέρας: Συγκεντρωμένη ακτινοβολία, σχετικά σταθερή σε όλο το εύρος του.

GridSearch και Επικύρωση

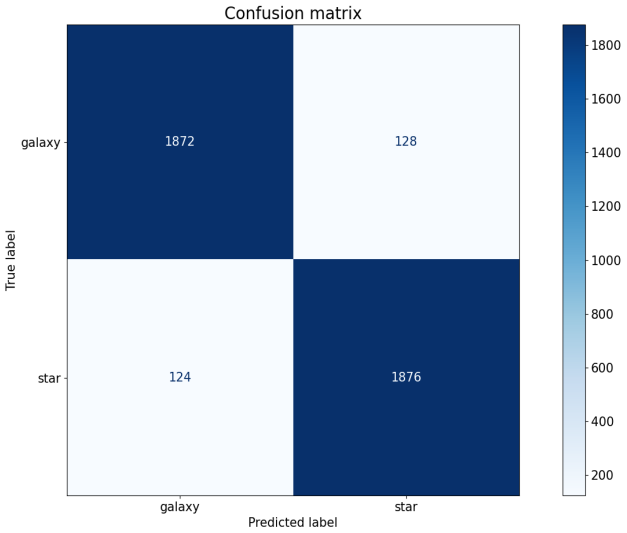
Για την αναζήτηση του μοντέλου και των παραμέτρων του έγινε χρήση της μεθόδου GridSearch, όπου συγκεκριμένα έγινε ανζήτηση του μοντέλου και των παραμέτρων του που επιτυγχάνει την υψηλότερη ακρίβεια στο σύνολο δεδομένων εκπαιδευσης. Κατά τη διαδικασία ελέχθηκαν τα μοντέλα με τους συνδιασμούς παραμέτρων που φαίνονται παρακάτω, με την επικύρωση να γίνεται απευθείας μέσω της μεθόδου με 5-fold cross validation, έτσι ώστε και πάλι όπως και προηγουμένως να χρησιμοποιείται το 80% του συνόλου για εκπαίδευση και το υπόλοιπο 20% για επικύρωση.
Τα μοντέλα και οι παράμετροι που εξετάστηκαν:

- *RF:*
  - *Estimators:* 100-400, step 50
  - *Criterion:* Gini, Entropy
- *SVM:*
  - *Kernel:* linear, polynomial, rbf
  - *C:* 10^0 – 10^4, 10 numbers

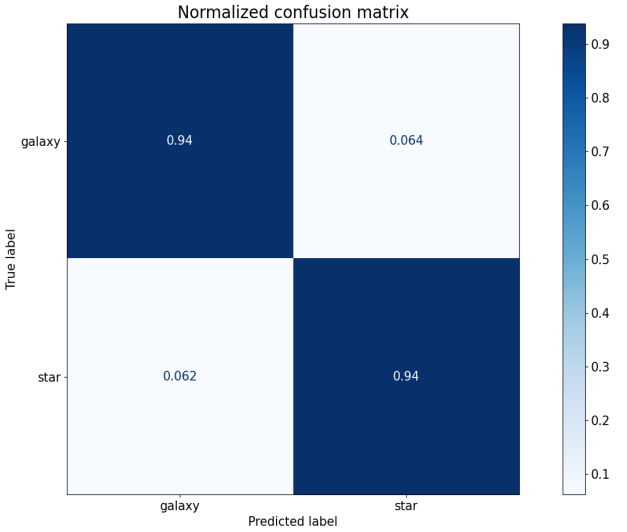
Από τη διαδικασία αυτή το καλύτερο μοντέλο για την επίλυση του προβλήματος μας αναδείχτηκε ο SVM με παραμέτρους C=7.7426, kernel: 'rbf', με μέση ακρίβεια *92.4%* και μέση απόκληση 0.01 σε όλα τα folds. Το εποτέλεσμα δεν εκπλήσει, δεδομένης της πληθώρας μηδενικών στο σύνολο δεδομένων και αφού η κατηγοριοποίηση γίνεται ανάμεσα σε δύο κλάσεις μόνο.

Αποτελέσματα

Εφαρμόζοντας το εκπαιδευμένο δίκτυο στις 2000 εικόνες του συνόλου δοκιμής παράχθηκαν τα ακόλουθα αποτελέσματα:



(α) Μη-Κανονικοποιημένος Πίνακας Σύγχισης

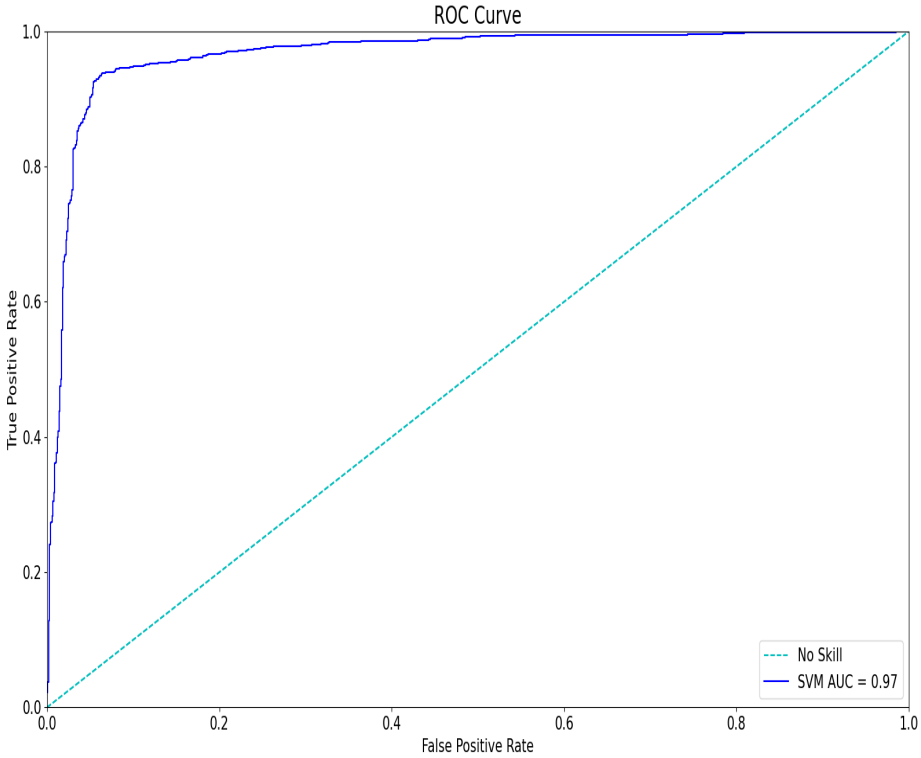


(β)Κανονικοποιημένος Πίνακας Σύγχισης

Σχήμα 2: Πίνακες Σύγχισης

Όπως φαίνεται και από τους πίνακες σύγχησης, τα αποτελέσματα είναι παρόμοια κι

εξαιρετικά και για τις δύο κλάσεις, αναλογιζόμενοι οτι οι πληροφορίες για την κατηγοριοποίηση των σωμάτων προέρχονται αποκλειστικά από φωτογραφίες, με ακρίβεια 93.7%.
Αυτό ενισχύεται και απο την καμπύλη ROC του μοντέλου, που μαρτυρά επίσης την πολύ καλή απόδοση, αφού ολόληρη η καμπύλη βρίσκερα αρκετά πιο πάνω απο την καμπύλη τυχαίας απόφασης και από κάτω της περικλέιεται επιφάνεια ίση με το 97% της συνολικής.

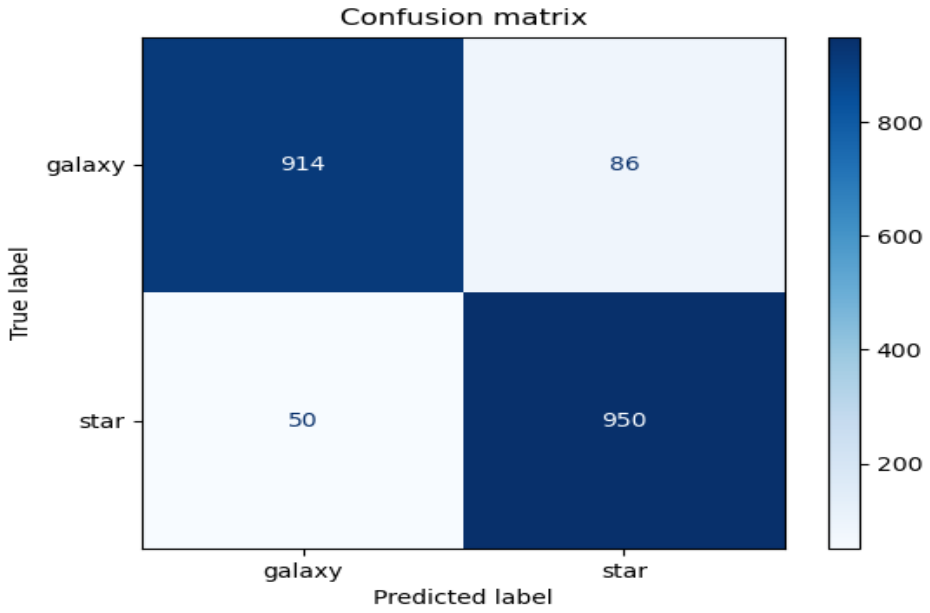


Σχήμα 3: ROC Καμπύλη

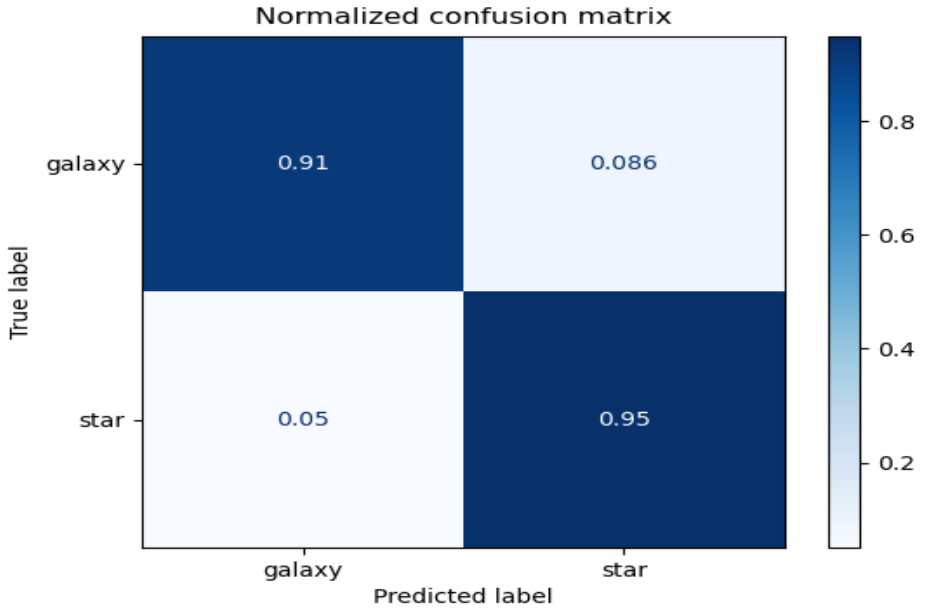
Λόγω του ότι τα παραδείγματα των κλάσεων είναι ισόποσα, οι παραπάνω μετρικές θεωρήθηκαν αρκετές για την εκτίμηση της ορθής λειτουργείας του μοντέλου.

Αναζήτηση Συμαντικότερου Χαρακτηριστικού

Όσον αφορά το σημαντικότερο χαρακτηριστικό εκ των δύο που χρησιμοποιήθηκαν, αυτό τελικά ήταν το χρώμα του αντικειμένου, όπως και είχε εικασθεί εξαρχής. Για να προκύψει αυτό το συμπέρασμα, έγινε διαχωρισμός του συνόλου των χαρακτηριστικών στα επιμέρους δύο και για μια αρχική ένδειξη, ο ταξινομητής εφαρμόστηκε στα νέα δύο σύνολα δεδομένων. Η ακρίβεια με χρήση των μορφολογικών χαρακτηριστικών των ουράνιων σωμάτων έδωσε ακρίβεια ίση με 81.3% στο σύνολο δοκιμής, ενώ με χρήση μόνο του ιστογράμματος χρώματος η ακρίβεια που επετεύχθη ήταν ίση με *92.7%* (μόλις *1%* χαμηλότερη ακρίβεια από το αρχικό).
Για να εξακριβωθεί η ορθότητα του παραπάνω αποτελέσματος η μέθοδος του GridSearch με τις ίδιες παραμέτρους, εφαρμόστηκε και στα δύο νέα (χωριστά) σύνολα δεδομένων (ο διαχωρισμός έγινε τόσο στο σύνολο εκπαίδευσης όσο και στο σύνολο δοκιμής). Και με αυτή τη μέθοδο τα χρωματικά χαρακτηριστικά φάνηκαν να υπερτερούν έναντι των μορφολογικών, για τον απλούστατο λόγο ότι η αναζήτηση και επαίδευση των μοντέλων με χρήση μορφολογικών χαρακτηριστικών διήρκεσε πλέον των δύο ημερών. Αντιθέτως, η μέση ακρίβεια που επετεύχθη στο σύνολο εκπαίδευσης ήταν *90%* με τυπική απόκλιση 0.01, με το μοντέλο που πέτυχε τη μεγαλύτερη ακρίβεια να είναι ένας SVM με παραμέτρους C=2.7826, kernel: 'rbf'.
Με εφαρμογή του παραπάνω μοντέλου στο σετ δοκιμής προέκυψαν οι παρακάτω πίνακες και καμπύλη ROC:

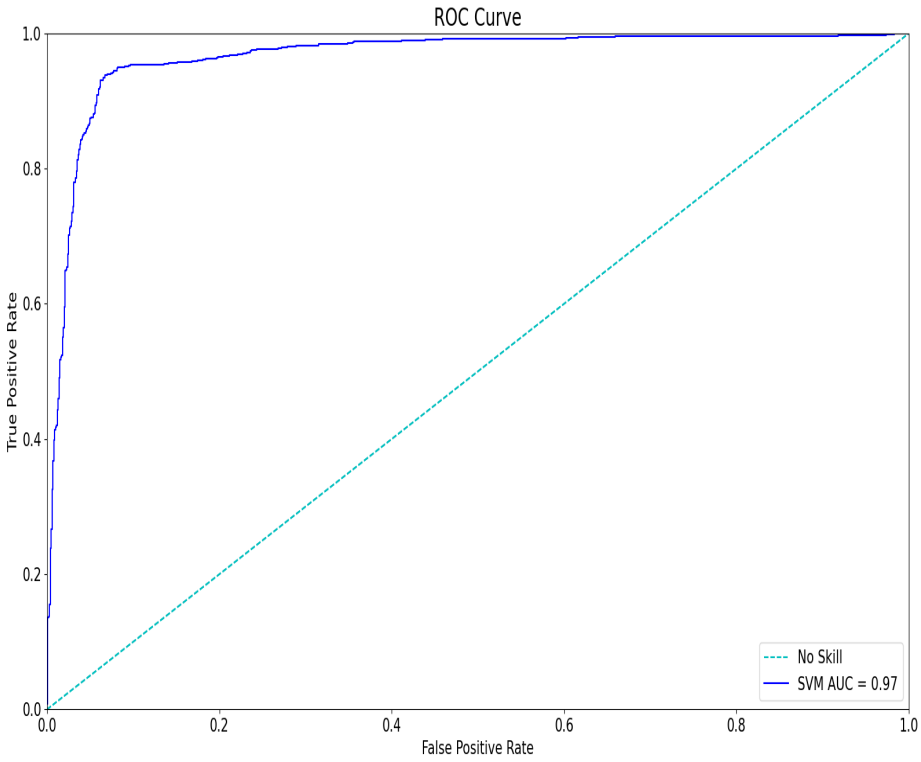


(α) Μη-Κανονικοποιημένος Πίνακας Σύγχισης



(β)Κανονικοποιημένος Πίνακας Σύγχισης

Σχήμα 4: Πίνακες Σύγχισης για Χαρακτηριστικό Χρώματος



Σχήμα 5: ROC Καμπύλη για Χαρακτηριστικό Χρώματος

Από τα παραπάνω φαίνεται πως και αυτό το μοντέλο μπρεί να λειτουργήσει αρκούντως καλά, στα ίδια επίπεδα με το βασικό. Ωστόσο, εδώ παρατηρείται μια μικρή διαφορά. Το μοντέλο τείνει να ταξινομεί ακριβέστερα τα αντικείμενα της κλάσης των αστέρων με ακρίβεια *95%* έναντι του *91%* της κλάσης των γαλαξιών. Αυτό βέβαια είναι και αναμενόμενο, δεδομένης της υψυλότερης έντασης ακρινοβολίας των αστέρων (πιο έντονο χρώμα), ειδικότερα στις φωτογραφίες που χρησιμοποιήθηκαν.

Συμπεράσματα

Καταλήγοντας, από τα αποτελέσματα της εν λόγω εργασίας παρατειρείται πως η ταξινόμηση ουράνιων σωμάτων με βάση μόνο χαρακτηριστικά που προκύπτουν από φωτογραφίες τους φέρνει ικανοποιητικά αποτελέσματα, δεδομένης και της χαμηλής υπολογιστικής ενέργειας που απαιτεί η διαδικασία σε σχέση με άλλες ακριβέστερες, επιπλέον χαρακτηριστικά ακτινοβολία σε μη-οπτικά μήκη κύματος, απόσταση από τη Γη κ.ά.
Ωστόσο, πιθανώς να μπορούν να προκύψουν ακόμα καλύτερα αποτελέσματα και με τα παραπάνω μοντέλα , αν γίνει περαιτέρω προεργασία των εικόνων, πχ με την εφαρμογή HOGs, τα οποία θα μειώσουν τις διαστάσεις κάθε εικόνας και θα συμπυκνώσουν τις πληροφορίες που περιέχει. Κάτι τέτοιο είναι πιθανό να μελετηθεί εκτενέστερα σε κάποια επόμενη εργασία.
Τέλος, αναλογιζόμενοι τα διαφορετικά χαρακτηριστικά που φέρουν τα διάφορα ουράνια σώματα (πχ πλανήτες, αστέρες, νεφελώματα κ.ά) είναι εύκολο το παραπάνω πρόβλημα να προσαρμοστεί ώστε το μοντέλο να έχει τη δυνατότητα να αναγνωρίσει και να κατηγοριοποιήσει αυτά τα σώματα σε περισσότερες των δύο κλάσσεων. Ωστόσο, εικάζεται πως το βέλτιστο μοντέλο πιθανώς να μην είναι και πάλι ένας SVM, αφού τα RF τείνουν να αποδίδουν καλυτερα από τους τελευταίους σε προβλήματα πολλαπλών κλάσσεων.

Βιβλιογραφία

[1] L. Shamir, Automatic morphological classification of galaxy images, Monthly Notices of the Royal Astronomical Society 399 (3) (2009) 1367–1372. arXiv:https://academic.oup.com/mnras/article-pdf/399/3/1367/18693417/mnras0399-1367.pdf, doi:10.1111/j.1365-2966.2009.15366.x.
URL https://doi.org/10.1111/j.1365-2966.2009.15366.x

[2] S. Sharma, R. Sharma, Classification of astronomical objects using various machine learning techniques, in: V. Jain, G. Chaudhary, M. C. Taplamacioglu, M. S. Agarwal (Eds.), Advances in Data Sciences, Security and Applications, Springer Singapore, Singapore, 2020, pp. 275–283.

[3] N. M. Ball, R. J. Brunner, Data mining and machine learning in astronomy, International Journal of Modern Physics D 19 (07) (2010) 1049–1106.

[4] D. Baron, Machine learning in astronomy: A practical overview, arXiv preprint arXiv:1904.07248 (2019).