ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΪΑΤΡΙΚΗ

# ΣΥΓΚΡΙΣΗ ΑΠΟΔΟΣΗΣ ΤΑΞΙΝΟΜΗΤΩΝ ΣΕ ΔΕΔΟΜΕΝΑ RNA-SEQUENCING

## ΕΡΓΑΣΙΑ ΓΙΑ ΤΟ ΜΑΘΗΜΑ ΒΙΟΛΟΓΙΑ ΣΥΣΤΗΜΑΤΩΝ

ΕΚΠΩΝΗΘΗΚΕ ΑΠΟ ΤΟΝ
*[ΛΑΖΑΡΟ ΚΩΝΣΤΑΝΤΙΝΟ-ΠΑΝΑΓΙΩΤΗ, 1639]*

ΔΙΔΑΣΚΟΥΣΑ ΚΑΘΗΓΗΤΡΙΑ
*[ΚΟΝΤΟΥ ΠΑΝΑΓΙΩΤΑ]*

ΛΑΜΙΑ, 2021-2022

# Voting vs Classic Classifiers on RNA-sequencing data binary classification

Κωνσταντίνος-Παναγιώτης Λάζαρος

*Abstract*—**Η παρούσα εργασία είναι μια συγκριτική μελέτη μεταξύ απλών ταξινομητών και μιας μεθόδου ταξινόμησης βασισμένης στην σταθμισμένη ψηφοφορία. Πρόκειται για ένα πρόβλημα δυαδικής ταξινόμησης (binary classification) σε δεδομένα μεγάλου όγκου (RNA-sequencing) τα οποία προέρχονται από την δευτερογενή βάση δεδομένων Expression Atlas του Ευρωπαϊκού Ινστιτούτου Βιοπληροφορικής (EMBL-EBI) που αφορούν στον νέο κορωνοϊό SARS-CoV-2.**

## I. ΕΙΣΑΓΩΓΗ

Το RNA-Sequencing είναι μια τεχνική λειτουργικής γονιδιωματικής μέσω της οποίας εξετάζεται η ποσότητα και το είδος του RNA σε ένα δείγμα μέσω αλληλούχισης επόμενης γενιάς (Next Generation Sequencing). Αναλύει το μεταγράφωμα ένος οργανισμού, υποδεικνύοντας όχι μόνο ποιά γονίδια εκφράζονται και ποιά δεν εκφράζονται αλλά και σε τί βαθμό. Έχει πάρα πολλές εφαρμογές καθώς μας επιτρέπει να αναλύσουμε το μεραγράφωμα ένος οργανισμού, το συνολικό δηλαδή κυτταρικό περιεχόμενο mRNA, rRNA και tRNA. Η κατανόηση του μεταγραφώματος είναι ιδιαίτερα σημαντική για την εύρεση της σύνδεσης μεταξύ του γονιδιώματος και της πρωτεϊνικής έκφρασης και λειτουργίας. Το RNA-sequencing χρησιμοποιείται για δημιουργία προφίλ μεταγραφώματος, για τον εντοπισμό πολυμορφισμών ενός νουκλεοτιδίου (Single Nucleotide Polymorphism), κλπ.

Τα δεδομένα τα οποία προκύπτουν από RNA-sequencing μπορούν επίσης να χρησιμοποιηθούν και για προβλέψεις. Συγκεκριμένα, εάν έχουμε ένα σύνολο δεδομένων με τα γονίδια τα οποία εκφράζονται (καθώς και τον βαθμό στον οποίο εκφράζονται) για μια ομάδα ατόμων και εάν γνωρίζουμε την κατάσταση κάθε ατόμου (υγιές ή ασθενές) στην ομάδα, τότε το συγκεκριμένο σύνολο δεδομένων μπορεί να χρησιμοποιηθεί για την εκπαίδευση ενός ταξινομητή, έτσι ώστε να γίνεται πρόβλεψη για την κατάσταση στην οποία βρίσκεται ένα άτομο με βάση τα γονίδια τα οποία εκφράζονται στα κύτταρα του (και τον βαθμό έκφρασής τους).

Υπάρχουν πολλοί ταξινομητές που μπορούν να εκπαιδευτούν έτσι ώστε να κάνουν προβλέψεις σχετικά με την κατάσταση στην οποία βρίσκεται ένας οργανισμός. Το πρόβλημα είναι ότι ένας ταξινομητής από μόνος του μπορεί να μην είναι αρκετά "ισχυρός" όσον αφορά στις προβλέψεις που κάνει. Μια ενδιαφέρουσα τεχνική που επιλύει αυτό το πρόβλημα έως έναν βαθμό, είναι ο συνδυασμός των προβλέψεων από πολλούς διαφορετικούς ταξινομητές, με σκοπό να προκύψει μια τελική πρόβλεψη μέσω ψηφοφορίας.

Ένας ταξινομητής που βασίζεται στην ψηφοφορία, είναι ένας εκτιμητής ο οποίος εκπαιδεύει διάφορα απλά μοντέλα/ταξινομητές και δίνει σαν έξοδο μια πρόβλεψη που βασίζεται στον συνδυασμό των προβλέψεων κάθε απλού μοντέλου/ταξινομητή. Το κριτήριο συνδυασμού συνήθως έχει να κάνει με σταθμισμένη ψηφοφορία μεταξύ των αποτελεσμάτων κάθε ταξινομητή.

Υπάρχουν δύο τύποι κριτηρίων ψηφοφορίας:

1) Hard voting: Κάθε ταξινομητής "ψηφίζει" για μια κλάση και ως έξοδο παίρνουμε την κλάση που έχει την πλειοψηφία.
2) Soft voting: Κάθε ταξινομητής δίνει μια πιθανότητα για ένα σημείο δεδομένων να ανήκει σε μια από τις κλάσεις εξόδου. Οι προβλέψεις είναι σταθμισμένες ανάλογα με το βάρος που έχει ανατεθεί σε κάθε ταξινομητή και στην συνέχεια προστίθενται. Η κλάση εξόδου με το μεγαλύτερο άθροισμα σταθμισμένων πιθανοτήτων είναι αυτή που λαμβάνεται ως έξοδος.

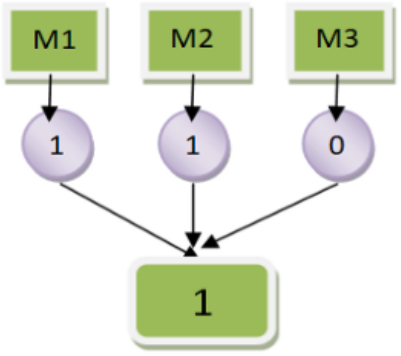


Fig. 1. Ψηφοφορία τύπου Hard Voting

Σκοπός της τεχνικής αυτής είναι να έχουμε καλύτερες προβλέψεις με μικρότερη πιθανότητα σφάλματος. Μέσω της παρούσας εργασίας συγκρίνεται η απόδοση ενός ταξινομητή που βασίζεται στην ψηφοφορία και των αντίστοιχων μεμονωμένων ταξινομητών που χρησιμοποιούνται για την σύνθεση του. Η σύγκριση γίνεται με χρήση ενός συνόλου δεδομένων που έχει προκύψει με την τεχνική του RNA-sequencing και αφορά σε ένα δείγμα case-control ατόμων όπου οι

ασθενείς πάσχουν από covid-19.

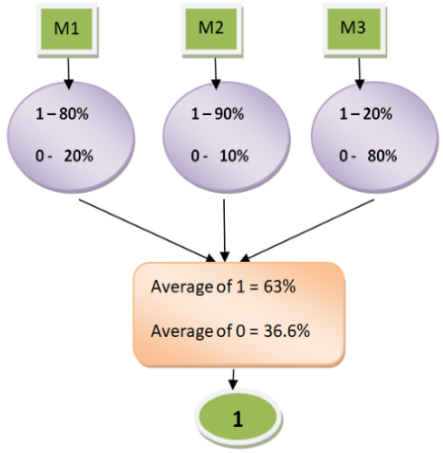


Fig. 2. Ψηφοφορία τύπου Soft Voting

## II. ΜΕΘΟΔΟΛΟΓΙΑ

Πρίν χτιστεί ο voting classifier, είναι απαραίτητο να γίνει προεπεξεργασία των δεδομένων. Στο σύνολο δεδομένων που χρησιμοποιείται, κάθε χαρακτηριστικό αντιπροσωπεύει ένα γονίδιο και η αριθμητική τιμή του, αποτελεί την ποσοτικοποίηση της έκφρασης του.

Επομένως το σύνολο δεδομένων είναι ένας $m \times n$ πίνακας του οποίου οι γραμμές αποτελούν δείγματα/ανθρώπους που εξετάστηκαν και κάθε στήλη αντιπροσωπεύει ένα γονίδιο (το οποίο μπορεί να εκφράζεται, υπερεκφράζεται, υποεκφράζεται ή να μην εκφράζεται καθόλου). Θα πρέπει να γίνει αλλαγή στις τιμές των ετικετών εξόδου από κατηγορηματικές (covid-19, normal) σε αριθμητικές/δυαδικές τιμές (1,0) καθώς πρόκειται για πρόβλημα δυαδικής ταξινόμησης.

Στην συνέχεια θα πρέπει να γίνει μείωση των χαρακτηριστικών του συνόλου δεδομένων. Κάποια χαρακτηριστικά/γονίδια έχουν μηδενική τιμή για έναν μεγάλο αριθμό δειγμάτων/ανθρώπων στο σετ δεδομένων. Γίνεται λοιπόν έλεγχος και αν ένα γονίδιο έχει μηδενική τιμή για το 50% των δειγμάτων του συνόλου δεδομένων, τότε διαγράφεται. Έτσι, γίνεται μείωση χαρακτηριστικών, δηλαδή μείωση της διάστασης/πολυπλοκότητας του συνόλου δεδομένων. Τα δεδομένα επίσης κανονικοποιούνται έτσι ώστε οι τιμές για κάθε γονίδιο να βρίσκονται μεταξύ 0 και 1

Τέλος, το σύνολο δεδομένων θα πρέπει να χωριστεί σε training και test υποσύνολα. Το training set θα χρησιμοποιηθεί για την εκπαίδευση του voting classifier καθώς και των απλών ταξινομητών από τους οποίους αποτελείται, ενώ το test set θα χρησιμοποιηθεί στην συνέχεια για την αξιολόγηση των ταξινομητών όταν η εκπαίδευση έχει πλέον ολοκληρωθεί.

Εφόσον πλέον τα δεδομένα είναι επεξεργασμένα χτίζεται ο voting classifier ο οποίος θα αποτελείται από τρεις απλούς ταξινομητές οι οποίοι θα είναι οι εξής:

1) Random Forest Classifier
2) Support Vector Machine Classifier
3) Logistic Regression Classifier.

Για την παρούσα μελέτη χρησιμοποιείται η μέθοδος του soft voting. Σε κάθε ταξινομητή ανατίθενται βάρη. Στον random forest classifier έχει ανατεθεί βάρος 3 ενώ στον logistic regression classifier και στον Support vector machine έχει ανατεθεί το ίδιο βάρος το οποίο είναι ίσο με την μονάδα.

## III. ΑΠΟΤΕΛΕΣΜΑΤΑ

Εφόσον η εκπαίδευση έχει ολοκληρωθεί, γίνεται αξιολόγηση των ταξινομητών στο test set. Προκειμένου να αξιολογηθούν οι ταξινομητές χρησιμοποιούνται τρείς μετρικές οι οποίες είναι οι εξής:

1) Accuracy: μετρική που χρησιμοποιείται έτσι ώστε να δούμε πόσες προβλέψεις είναι αληθείς είτε είναι θετικές (στην περίπτωση μας: covid-19) είτε είναι αρνητικές (όχι covid-19, άρα υγιής). Το accuracy, δίνεται από τον τύπο:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

2) F1 score: Συνδυάζει το precision και το recall σε μια μετρική, υπολογίζοντας τον αρμονικό μέσο όρο μεταξύ τους.Δίνεται από τον γενικό τύπο Fbeta ο οποίος είναι ο εξής:

$$f1score = \frac{2 \times (precision \times recall)}{precision + recall}$$

Στο F1 score, το precision και το recall έχουν την ίδια αξία (ενώ στο f2 score, δίνεται διπλάσια σημασία στο recall σε σχέση με το precision).

3) ROC-AUC: Το AUC είναι το εμβαδόν κάτω από την καμπύλη ROC (Area Under the Curve). Η καμπύλη ROC δείχνει την αντίθεση που υπάρχει μεατξύ του ρυθμού αληθώς θετικών προβλέψεων (true positive rate) και του ρυθμού ψευδώς θετικών προβλέψεων (false positive rate). Προφανώς, όσο πιο μεγάλο είναι το true positive rate και όσο πιο μικρό είναι το false positive rate, τόσο το καλύτερο για τον ταξινομητή (η καμπύλη θα βρίσκεται πάνω και προς τα αριστερά). Δίνεται από τον τύπο:

$$\int_0^1 TPR(FPR^{-1}(x))\,dx$$

Από τα αποτελέσματα της αξιολόγησης, γίνεται προφανές ότι ο ταξινομητής που βασίζεται στην ψηφοφορία (voting classifier) έχει καλύτερη απόδοση όσον αφορά την ακρίβεια (accuracy) και το f1 score σε σχέση με τους απλούς ταξινομητές. Το ROC-AUC

score στην περίπτωση του voting classifier είναι καλύτερο σε σχέση με τον Random forest και τον logistic regression classifier και λίγο χειρότερο από αυτό του SVM classifier. Ίσως αυτή η διαφορά να μειωθεί με αλλαγή των βαρών που ανατέθηκαν στο soft voting.

| Model metrics on RNA-seq dataset | | | |
|---|---|---|---|
| Model | Accuracy | F1 score | ROC-AUC |
| RFC | 0.871 | 0.917 | 0.772 |
| LRC | 0.865 | 0.910 | 0.808 |
| SVM | 0.878 | 0.917 | 0.836 |
| VC | 0.893 | 0.930 | 0.824 |

Παρατίθεται και διάγραμμα με την πιθανότητα ταξινόμησης του πρώτου δείγματος στις δύο κλάσεις για όλους τους ταξινομητές.

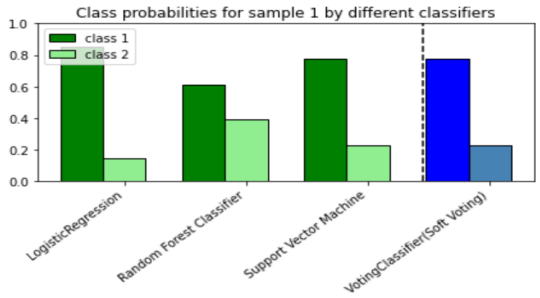


Fig. 3. DoubleUnet KID-dataset evaluation sample images

## IV. ΕΠΙΛΟΓΟΣ

Η παρούσα εργασία αποτελεί μια συγκριτική μελέτη της απόδοσης μερικών απλών ταξινομητών έναντι ενός ταξινομητή που βασίζεται στην ψηφοφορία (ο voting classifier χρησιμοποιεί τους τρεις απλούς ταξινομητές προκειμένου να διεξαχθεί η ψηφοφορία).

Τα δεδομένα τα οποία χρησιμοποιήθηκαν προέκυψαν με την μέθοδο του RNA-sequencing και αφορούν στην έκφαση γονιδίων σε ένα case-control group για covid-19.

Σε μελλοντική μελέτη θα μπορούσε να γίνει σύγκριση της απόδοσης ενός ταξινομητή που βασίζεται στην ψηφοφορία και ενός πιο σύνθετου ταξινομητή/νευρωνικού δικτύου.

## References


[1] Citation: Zararsız G, Goksuluk D, Korkmaz S, Eldem V, Zararsiz GE, Duru IP, et al. (2017) A comprehensive simulation study on classification of RNA-Seq data. PLoS ONE 12(8): e0182507. https://doi.org/10.1371/journal.pone.0182507

[2] Bauer, E., Kohavi, R. An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. Machine Learning 36, 105–139 (1999). https://doi.org/10.1023/A:1007515423169

[3] Cortes, C., Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.

[4] Ho, T. K. (1995). Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition (Vol. 1, pp. 278–282).

[5] Cox, D. R. (1958). The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological), 20(2), 215–232.