ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

Εφαρμογη μεθοδου συλλογικης βαθιας μαθησης (Ensemble Deep Learning) για την εξορυξη δεδομενων υψηλης διαστατικοτητας

Μενυχτα Αικατερινη

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ
ΥΠΕΥΘΥΝΟΣ
ΑΡΙΣΤΕΙΔΗΣ Γ. ΒΡΑΧΑΤΗΣ

Λαμία, 2022

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

ΕΦΑΡΜΟΓΗ ΜΕΘΟΔΟΥ ΣΥΛΛΟΓΙΚΗΣ ΒΑΘΙΑΣ ΜΑΘΗΣΗΣ (ENSEMBLE DEEP LEARNING) ΓΙΑ ΤΗΝ ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΥΨΗΛΗΣ ΔΙΑΣΤΑΤΙΚΟΤΗΤΑΣ

Μενύχτα Αικατερίνη

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ
ΥΠΕΥΘΥΝΟΣ
ΑΡΙΣΤΕΙΔΗΣ Γ. ΒΡΑΧΑΤΗΣ

Λαμία, 2022

Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις [1], που προβλέπονται από της διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

1. *Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί* ***χωρίς να τα περικλείω σε εισαγωγικά*** *και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάσθηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.*

2. *Δέχομαι ότι η αυτολεξεί* ***παράθεση χωρίς εισαγωγικά****, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.*

3. *Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια*

4. *Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.*

Ημερομηνία: 17/03/2022

Η Δηλ.

Μενύχτα Αικατερίνη

(Υπογραφή)

(1) «Όποιος εν γνώσει του δηλώνει ψευδή γεγονότα ή αρνείται ή αποκρύπτει τα αληθινά με έγγραφη υπεύθυνη δήλωση του άρθρου 8 παρ. 4 Ν. 1599/1986 τιμωρείται με φυλάκιση τουλάχιστον τριών μηνών. Εάν ο υπαίτιος αυτών των πράξεων σκόπευε να προσπορίσει στον εαυτόν του ή σε άλλον περιουσιακό όφελος βλάπτοντας τρίτον ή σκόπευε να βλάψει άλλον, τιμωρείται με κάθειρξη μέχρι 10 ετών.

ΕΦΑΡΜΟΓΗ ΜΕΘΟΔΟΥ ΣΥΛΛΟΓΙΚΗΣ ΒΑΘΙΑΣ ΜΑΘΗΣΗΣ (ENSEMBLE DEEP LEARNING) ΓΙΑ ΤΗΝ ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΥΨΗΛΗΣ ΔΙΑΣΤΑΤΙΚΟΤΗΤΑΣ

Μενύχτα Αικατερίνη

Τριμελής Επιτροπή:

Βραχάτης Αριστείδης, Επίκουρος Καθηγητής (Επιβλέπων)

Πλαγιανάκος Βασίλειος, Καθηγητής

Τασουλής Σωτήριος, Επίκουρος Καθηγητής

Περίληψη

Τα τελευταία χρόνια είναι εμφανής η αύξηση του όγκου των δεδομένων, ενώ μεγάλες είναι οι προσπάθειες που γίνονται με σκοπό την καλύτερη διαχείριση και εκμετάλλευσή τους. Παράλληλα, αυξάνεται και ο όγκος των διαθέσιμων βιολογικών δεδομένων, καθώς μέσω αυτών είναι δυνατή η καλύτερη κατανόηση του κόσμου, η ανάπτυξη μεθόδων προσωποποιημένης θεραπείας κα διάγνωσης, η μελέτη της κυτταρικής ετερογένειας, ενώ δίνονται απαντήσεις σε πολλά σύγχρονα ερωτήματα. Ειδικότερα στον τομέα της αλληλούχισης RNA οργανισμών η σωστή ανάλυση των δεδομένων που προκύπτουν μέσα από εργαστηριακές διαδικασίες, μπορούν να βοηθήσουν την έρευνα και την αντιμετώπιση ασθενειών, την εύρεση των υπεύθυνων γονιδίων για διάφορες μεταλλάξεις ή παθήσεις και δυνητικά την εύρεση βιοδεικτών ή γονιδιακών εκφράσεων που σχετίζονται με τους διάφορους τύπους καρκίνου.

Για την καλύτερη και πιο αποτελεσματική διαχείριση και ερμηνεία της πληροφορίας, χρειάζεται να αναπτυχθούν μέθοδοι ανάλυσης με τη χρήση της τεχνητής νοημοσύνης και της βαθιάς μάθησης, εξειδικευμένων για βιολογικά δεδομένα. Μέσα σε αυτό το πλαίσιο, αφού γίνει μια αναφορά σε όρους που σχετίζονται με τα μεγάλα δεδομένα και τη μηχανική μάθηση, γίνεται μια εκτενής αναφορά στις πιο γνωστές, υπάρχουσες μεθόδους εξαγωγής χαρακτηριστικών και μείωσης διαστάσεων πολυδιάστατων (βιολογικών και μη) δεδομένων. Στην συνέχεια παρουσιάζεται η μέθοδος scVEC, η οποία υλοποιήθηκε για την παρούσα εργασία και είναι μια μέθοδος κατηγοριοποίησης βιολογικών δεδομένων από αλληλούχιση RNA μεμονωμένου κυττάρου (scRNA-seq data), που βασίζεται στη χρήση των Autoencoders και της συλλογικής (Ensemble) μάθησης. Στην αρχή η μέθοδος μειώνει τη διάσταση των δεδομένων εισόδου με τη χρήση Autoencoder, ειδικά διαμορφωμένου για βιολογικά δεδομένα αυτής της μορφής και με βάση τη διαθέσιμη βιβλιογραφία και στη συνέχεια στο νέο χώρο που προκύπτει εφαρμόζεται αλγόριθμος KNN, με σκοπό την κατηγοριοποίηση των κυττάρων με βάση τη γονιδιακή τους έκφραση. Η διαδικασία αυτή επαναλαμβάνεται για έναν καθορισμένο αριθμό επαναλήψεων και στο τέλος τα αποτελέσματα συγκεντρώνονται και η τελική πρόβλεψη επιλέγεται εφαρμόζοντας μια μέθοδο voting μέσω πλειοψηφίας (majority vote). Η μέθοδος αυτή φαίνεται να αποδίδει πολύ καλά σε τέτοιου είδους δεδομένα, τα οποία είναι πολυδιάστατα και αραιά, έχοντας μεγάλη ακρίβεια και ισχύ κατηγοριοποίησης. Τα αποτελέσματα και η ισχύς της μεθόδου επαληθεύεται και επιβεβαιώνεται από μια νέα παρόμοια μελέτη που δημοσιεύτηκε το Δεκέμβριο 2021, ονομάζεται scIAE και περιγράφεται στην βιβλιογραφική ανασκόπηση της εργασίας.

Abstract

During the last decades the increase in the volume of data is evident, while great efforts are made in order to better manage and exploit them. At the same time, the volume of available biological data is increasing, as through them it is possible to better understand the world, to develop methods of personalized treatment and diagnosis, to study cell heterogeneity, while many modern questions are answered. In the field of RNA sequencing the proper analysis of data obtained through laboratory procedures can help research and treat diseases, find the genes responsible for various mutations or diseases, and potentially find biomarkers or gene expressions related to the different types of cancer.

For better and more efficient management and interpretation of information, methods of analysis using artificial intelligence and deep learning, specialized in biological data, need to be developed. In this context, after referring to terms related to big data and machine learning, an extensive reference is made to the most well-known, existing methods of extracting features and dimensionality reduction of multidimensional (biological and non-biological) data. As detailed below the scVEC method is presented, which was implemented for the present thesis, which is a method of categorizing biological data from single cell RNA sequencing (scRNA-seq data), based on the use of Autoencoders and Ensemble learning. At first the method reduces the dimensions of the input data using Autoencoders, specially formulated for biological data of this format and based on the available literature and then in the new space that results, KNN algorithm is applied, in order to categorize the cells based on their genetic expression. This process is repeated for a specified number of repetitions and at the end the results are aggregated, and the final prediction is selected by applying a majority voting method. This method seems to perform very well on such data, which are multidimensional and sparse with lots of dropout events, having high accuracy and categorization power. The results and validity of the method are verified and confirmed by a new similar study published in December 2021, called scIAE and described in the literature review.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω αρχικά τον επιβλέποντα καθηγητή μου κ. Βραχάτη, για όλη την βοήθεια, την καθοδήγηση και τον χρόνο που διέθεσε για την ολοκλήρωση της παρούσας εργασίας μου. Επίσης θα ήθελα να ευχαριστήσω τους γονείς μου και την αδελφή μου για τη στήριξη και τη συνεχή ενθάρρυνση τους όλο αυτό το χρονικό διάστημα. Τέλος, ευχαριστώ όλους τους κοντινούς μου ανθρώπους και φίλους που η βοήθεια και η συνεισφορά τους, όσον αφορά την παρούσα εργασία αλλά και τη συνολική ψυχολογική στήριξη που μου προσέφεραν όλο αυτό το χρονικό διάστημα, ήταν και παραμένει πολύτιμη.

Περιεχόμενα

Κατάλογος Εικόνων

1. Εισαγωγή

1.1. Η εποχή της πληροφορίας και των «μεγάλων δεδομένων»

Τα τελευταία χρόνια ο όγκος της πληροφορίας αυξάνεται εκθετικά. Η πληροφορία μπορεί να προέρχεται από διάφορα είδη πηγών, ενώ δημόσιοι και ιδιωτικοί ακαδημαϊκοί και βιομηχανικοί τομείς συλλέγουν δεδομένα μεγάλου όγκου με σκοπό την βελτίωση των συμπερασμάτων που εξάγουν ή των υπηρεσιών και των αγαθών που παρέχουν. Τα δεδομένα αυτά που συλλέγονται καθημερινά είναι συνήθως ακατέργαστα και η επεξεργασία τους αποτελεί ακόμα και σήμερα μεγάλο ερευνητικό ενδιαφέρον. Ο τρόπος που τα δεδομένα αυτά συλλέγονται, κατηγοριοποιούνται και επεξεργάζονται είναι κρίσιμης σημασίας, καθώς μια λανθασμένη ερμηνεία τους μπορεί να οδηγήσει σε ερευνητικά λάθη ή μη αντικειμενικά συμπεράσματα που μπορούν με ευκολία να αποπροσανατολίσουν την έρευνα.

1.2. Δεδομένα μεγάλου όγκου

Με τον όρο "δεδομένα μεγάλου όγκου" συνήθως γίνεται λόγος για μεγάλη ποσότητα δεδομένων που δεν είναι δυνατή η διαχείρισή τους όταν χρησιμοποιούνται τα συνήθη λογισμικά ή διαδικτυακές πλατφόρμες (Dash *et al.*, 2019). Με βάση τον νόμο του Moore έχει παρατηρηθεί ότι ο αριθμός των τρανζίστορ σε ένα πυκνό ολοκληρωμένο κύκλωμα θα διπλασιάζεται κάθε χρόνο για τουλάχιστον μια δεκαετία από τότε, πράγμα που τελικά επαληθεύτηκε, καθώς πράγματι για αρκετά χρόνια ο αριθμός τους διπλασιαζόταν κάθε δύο χρόνια (L Xiu, 2019), (Lundstrom, 2003). Αυτό σημαίνει ότι ενώ η επεξεργαστική ισχύς αυξάνεται, δεν μπορεί πάντα να ακολουθήσει την πορεία αύξησης του όγκου των διαθέσιμων δεδομένων. Τα δεδομένα μεγάλου όγκου χρειάζονται περισσότερο διαθέσιμο χώρο αποθήκευσης, επεξεργαστική ισχύ και τη χρήση συγκεκριμένων και στοχευμένων μεθόδων ανάλυσης. Το θέμα των μεγάλων δεδομένων παρουσιάζεται πιο αναλυτικά σε επόμενο κεφάλαιο όπως και οι δυσκολίες και οι προκλήσεις που προκύπτουν στην προσπάθεια ανάλυσης και επεξεργασίας τους.

Όσον αφορά τα βιοϊατρικά δεδομένα υπάρχουν εξίσου πολλές πηγές που αυξάνουν τον όγκο τους, όπως ιατρικοί φάκελοι, φάκελοι ασθενών, αποτελέσματα εξετάσεων, δεδομένα από νοσοκομεία, από συσκευές που χρησιμοποιούνται στον κλάδο της υγείας κλπ. Τα δεδομένα συνήθως είναι ποσοτικά (quantitive), με χρήση μετρήσεων ή ποιοτικά (qualitive), χρησιμοποιώντας περιγραφές. Τα βιολογικά δεδομένα ανήκουν σε πολλούς και διαφορετικούς τύπους και μπορεί να περιλαμβάνουν αριθμητικά δεδομένα, αρχεία εικόνων, κειμένου, καταγεγραμμένα σήματα (Shortliffe and Barnett, 2014) κ.α. Υπάρχουν όμως και βιολογικά δεδομένα που εξάγονται από άλλου είδους δείγματα, όπως για παράδειγμα από κύτταρα. Μέσα από διάφορες βιολογικές διεργασίες από ένα κύτταρο μπορεί να

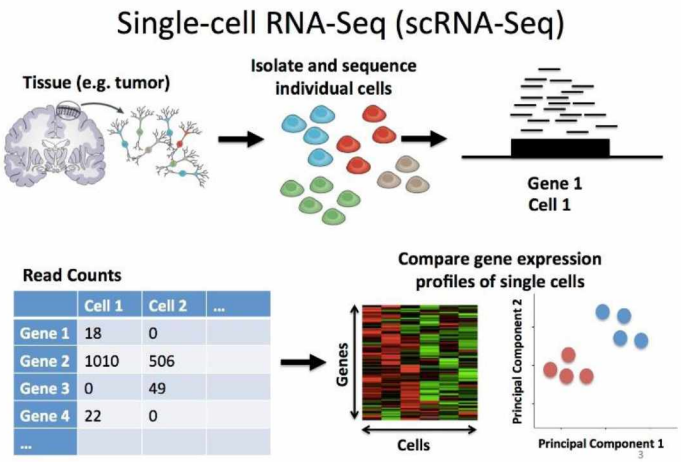
εξαχθεί πληθώρα δεδομένων που χρειάζονται περαιτέρω ανάλυση και ερμηνεία για να προκύψει κάποιο σημαντικό πόρισμα. Σημαντική είναι και η ανάγκη για ανάλυση και διάκριση των διαφόρων τύπων κυττάρων και ανάπτυξης μεθόδων μελέτης της κυτταρικής ετερογένειας. Τα γενετικά, βιοχημικά, μεταγραφικά και μεταβολικά δεδομένα ανήκουν στην κατηγορία των "ωμικών δεδομένων" (omics) (Vailati-Riboni, Palombo and Loor, 2017). Η ερμηνεία τους μπορεί να οδηγήσει σε σημαντικές αλλαγές στο σύστημα υγείας, την δημιουργία προσωποποιημένων μεθόδων θεραπείας ασθενών αλλά και την επίλυση προβλημάτων που απασχολούν την επιστημονική και ιατρική κοινότητα εδώ και χρόνια όπως η θεραπεία ή ανάπτυξη τρόπων αντιμετώπισης των διαφόρων τύπων καρκίνου. Για τους παραπάνω λόγους όσο περισσότερα δεδομένα συλλέγονται τόσο πιο εύκολη γίνεται η κατανόηση των βιολογικών διαδικασιών που μελετώνται κάθε φορά. Παράλληλα αναπτύσσεται και πληθώρα εφαρμογών για την εξαγωγή, ανάλυση και επεξεργασία όλων αυτών των δεδομένων που αφορά κυρίως τους τομείς της βιοπληροφορικής, κλινικής πληροφορικής, πληροφορικής απεικόνισης και πληροφορικής της δημόσιας υγείας.

1.3. Τεχνολογίες αλληλούχισης γονιδιωμάτων επόμενης γενιάς

Σημαντική είναι και η ανάπτυξη της τεχνολογίας NGS ή next generation sequencing (Arumugam, Uli and Annavi, 2019), (Next-Generation Sequencing (NGS) | Explore the Technology, Illumina, n.d.). Η NGS είναι μια τεχνολογία για μαζική και παράλληλη αλληλούχιση με υψηλή απόδοση, που χρησιμοποιείται για τον προσδιορισμό της σειράς των νουκλεοτιδίων σε ολόκληρα γονιδιώματα ή στοχευμένες περιοχές DNA ή RNA. Η τεχνολογία αυτή επιτρέπει στα εργαστήρια να εκτελούν μια ευρεία ποικιλία εφαρμογών και να μελετούν βιολογικά συστήματα με μεγαλύτερη λεπτομέρεια.

Ακόμη περισσότερο επαναστατική όμως θεωρείται η τεχνολογία scRNA-Seq (Single Cell RNA Sequencing) που παρέχει βαθύτερη ακόμα εικόνα για την πολυεπίπεδη πολυπλοκότητα διαφορετικών κυττάρων που ανήκουν στον ίδιο τύπο ιστού. Πιο συγκεκριμένα εξετάζει την πληροφορία που εξάγεται από την αλληλούχιση μεμονωμένων κυττάρων μέσω βελτιστοποιημένης NGS, παρέχοντας καλύτερης ποιότητας ανάλυση των κυτταρικών διαφορών και καλύτερη κατανόηση της λειτουργίας ενός και μόνο κυττάρου σε σχέση με το περιβάλλον του (scRNA-Seq, Illumina, n.d.). Η κατανόηση των βιολογικών διαδικασιών, που μελετώνται μέσω της παραπάνω τεχνολογίας, σε επίπεδο ενός και μόνο κυττάρου, βασίζεται στην μοναδικότητα των κυττάρων αυτών. Για αυτό και τα δεδομένα που προκύπτουν, έχουν μεγάλη ετερογένεια και παρέχουν σημαντικές πληροφορίες για την λειτουργία κάθε κυττάρου. Η πληροφορία αυτή μπορεί να προσφέρει μελλοντικά σημαντικά πορίσματα και έχει μεγάλες

δυνατότητες στην κατανόηση του ανθρώπινου γονιδιώματος, την εύρεση πολύπλοκων βιολογικών μηχανισμών, την αντιμετώπιση ασθενειών και την ανάπτυξη προσωποποιημένων μεθόδων θεραπείας. Ειδικά για τον τελευταίο τομέα, υπάρχουν ασθένειες και σύνδρομα που εκδηλώνονται με διαφορετικό τρόπο ή έκταση ανάμεσα στους ασθενείς και κάποιες φορές είναι δύσκολη η αναγνώριση και θεραπεία τους. Ακόμη τα συμπτώματα κάποιων ασθενειών μπορεί να είναι κοινά σε περισσότερες από μια ασθένειες, η κάθε μια από τις οποίες όμως χρήζει και διαφορετικής αντιμετώπισης. Για τους παραπάνω λόγους κρίνεται απαραίτητη η ανάπτυξη προσωποποιημένων μεθόδων διάγνωσης και θεραπείας ασθενειών. Τα δεδομένα που προέρχονται από την τεχνολογία scRNA-Seq είναι και τα δεδομένα που χρησιμοποιήθηκαν στο κύριο μέρος της εργασίας και για τα οποία θα γίνει εκτενέστερη περιγραφή σε αντίστοιχο κεφάλαιο.



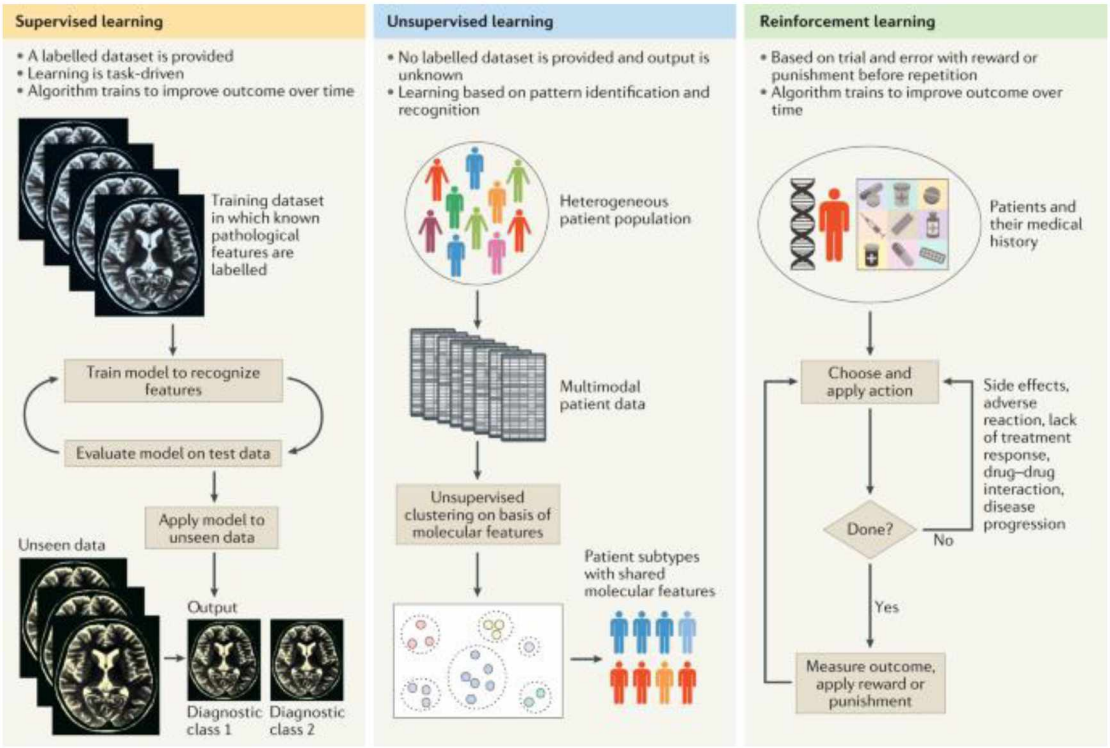
Εικόνα 1 Διαδικασία εργαστηριακής μεθόδου αλληλούχισης RNA μεμονωμένου κυττάρου

1.4. Επεξεργασία και εξόρυξη δεδομένων μεγάλου όγκου

Η επεξεργασία και ανάλυση των διαθέσιμων δεδομένων με σκοπό την εξαγωγή συμπερασμάτων χρήσιμων για τον άνθρωπο είναι ένας κλάδος της τεχνολογίας που τα τελευταία χρόνια έχει λάβει μεγάλες διαστάσεις. Ο κύριος σκοπός της ανάλυσης δεδομένων είναι να εξάγει ουσιαστικές πληροφορίες από τα δεδομένα, έτσι ώστε η γνώση που προκύπτει να μπορεί να χρησιμοποιηθεί για τη λήψη τεκμηριωμένων αποφάσεων. Οι επιστήμονες αναλύουν και ερμηνεύουν δεδομένα για να αναζητήσουν νόημα που μπορεί να χρησιμεύσει ως απόδειξη. Συχνά οι επιστήμονες προσπαθούν να προσδιορίσουν εάν οι μεταβλητές σχετίζονται μεταξύ τους και πόσο. Άλλες φορές χρησιμοποιούν τα

δεδομένα για να απαντήσουν σε ερωτήσεις που έχουν θέσει προκειμένου να λύσουν ένα πρόβλημα, όπως για παράδειγμα αν και κατά πόσο ένα φάρμακο είναι αποδοτικό.

Η εξόρυξη πληροφορίας από βιολογικά ή μη δεδομένα μεγάλου όγκου, για τους λόγους που αναφέρθηκαν, είναι μια χρονοβόρα και δύσκολη διαδικασία. Η ποσότητα των δεδομένων αυξάνεται διαρκώς, ενώ τα διαθέσιμα μέσα δεν είναι πάντα ικανά να τα επεξεργαστούν και να τα αναλύσουν, είτε για λόγους επεξεργαστικούς είτε επειδή είναι χρονοβόρο. Για αυτόν τον λόγο έχουν γίνει προσπάθειες ανάπτυξης μεθόδων με την χρήση της τεχνητής νοημοσύνης και της βαθιάς μάθησης που κάνουν την διαδικασία ευκολότερη. Με την χρήση της τεχνητής νοημοσύνης είναι δυνατή η εύρεση μοτίβων, εξαγωγή και επιλογή κατάλληλων χαρακτηριστικών και σχηματισμός συνάψεων που ένας άνθρωπος αδυνατεί να ανακαλύψει και μάλιστα σε πολύ λιγότερο χρόνο. Ο λόγος είναι ότι ο άνθρωπος εστιάζει σε διαφορετικές παραμέτρους και χαρακτηριστικά των δεδομένων, ή επηρεάζεται από πρότερες γνώσεις για την εξαγωγή συμπερασμάτων. Αντίθετα η τεχνητή νοημοσύνη εστιάζει σε χαρακτηριστικά όπως οι αποστάσεις των δεδομένων στο χώρο και αριθμητικές πληροφορίες που εξάγει και απομνημονεύει για να ανακαλύψει πιθανές συσχετίσεις και μοτίβα στα διαθέσιμα δεδομένα ή άλλες φορές ξεκινά από τυχαίες ή αυθαίρετες καταστάσεις τις οποίες τροποποιεί βρίσκοντας, για παράδειγμα, κατάλληλες παραμέτρους, ώστε τελικά να καταλήξει σε ουσιαστικά συμπεράσματα. Βέβαια, ο τρόπος που η τεχνητή νοημοσύνη λειτουργεί και δημιουργεί συσχετίσεις δεν μας είναι ακόμη και σήμερα γνωστός, καθώς λειτουργεί σαν μαύρο κουτί.



Εικόνα 2 Κατηγορίες μεθόδων μηχανικής μάθησης και εφαρμογές σε βιοϊατρικά δεδομένα

1.5. Τεχνητή νοημοσύνη και μηχανική μάθηση

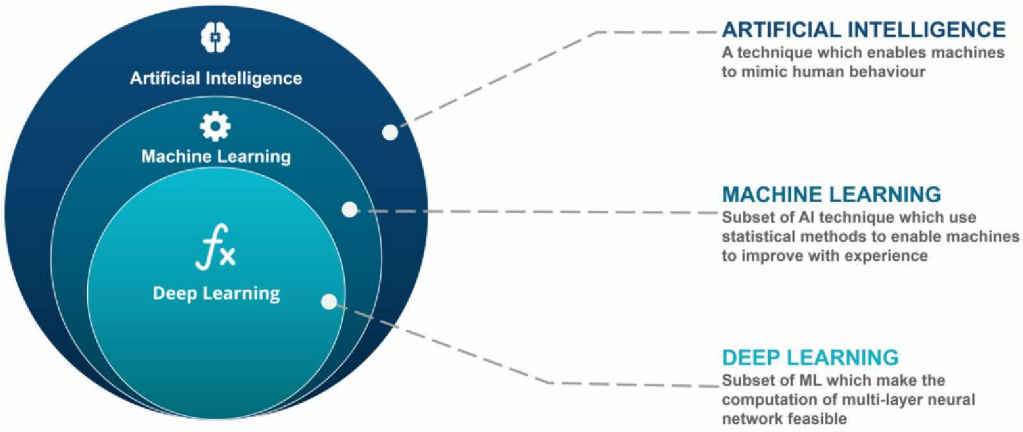
Η τεχνητή νοημοσύνη χρησιμοποιεί μηχανές και υπολογιστές με σκοπό να μιμηθεί τις ικανότητες επίλυσης προβλημάτων και λήψης αποφάσεων του ανθρώπινου μυαλού. Όμως όπως προαναφέρθηκε δεν λειτουργεί με τον ίδιο τρόπο. Υπάρχουν διαφορετικά συστήματα υπολογιστών με βάση τον ορθολογισμό και τη σκέψη έναντι της δράσης. Υπάρχουν συστήματα που σκέφτονται με τον ίδιο τρόπο με τους ανθρώπους και συστήματα που λειτουργούν και δρουν σαν άνθρωποι. Η τεχνητή νοημοσύνη είναι ένα πεδίο που χρησιμοποιείται για επίλυση προβλημάτων και συνδυάζει την επιστήμη των υπολογιστών με μεγάλα σύνολα δεδομένων με σκοπό να καταλήξει σε κάποια συμπεράσματα. Έχει πολλές πρακτικές εφαρμογές στην καθημερινή ζωή, όπως η αναγνώριση ομιλίας, η εξυπηρέτηση πελατών (πχ με την χρήση αυτοματοποιημένων μηνυμάτων με βάση τις απαντήσεις του πελάτη), η υπολογιστική όραση, (υπολογιστές και συστήματα εξάγουν πληροφορίες από ψηφιακές εικόνες), μηχανές που προτείνουν, με βάση διαθέσιμα δεδομένα και τάσεις, συγκεκριμένα προϊόντα σε πελάτες που κάνουν ηλεκτρονικές αγορές, αυτοματοποιημένες αγορές και πωλήσεις μετοχών κ.α..

Εικόνα 3 Εφαρμογές μεθόδων μηχανικής μάθησης

Υπάρχουν δύο τύποι τεχνητής νοημοσύνης. Η «αδύναμη» ή «στενή» τεχνητή νοημοσύνη (Narrow AI or Artificial Narrow Intelligence / ANI) εκπαιδεύεται και εστιάζεται στην εκτέλεση συγκεκριμένων εργασιών. Αυτή αποτελεί κυρίως το μεγαλύτερο μέρος της τεχνητής νοημοσύνης που χρησιμοποιείται σήμερα. Ο όρος «Narrow» πιθανότατα αποτελεί πιο ακριβή περιγραφή για αυτόν τον τύπο τεχνητής νοημοσύνης, καθώς είναι κάθε άλλο παρά αδύναμη. Επιτρέπει ορισμένες πολύ ισχυρές εφαρμογές, όπως το Siri της Apple, το Alexa της Amazon, το IBM Watson και αυτόνομα οχήματα [(What is Artificial Intelligence (AI)? | IBM, June 2020)]. Ο άλλος τύπος τεχνητής νοημοσύνης είναι η «ισχυρή» (Strong AI) και αποτελείται από την Τεχνητή Γενική Νοημοσύνη (AGI) και την Τεχνητή Σούπερ Νοημοσύνη (ASI). Η τεχνητή γενική νοημοσύνη (Artificial general intelligence / AGI) είναι μια θεωρητική μορφή τεχνητής νοημοσύνης όπου μια μηχανή θα έχει νοημοσύνη ίδια με αυτή των ανθρώπων, την ικανότητα να λύνει προβλήματα, να μαθαίνει και να σχεδιάζει το μέλλον. Η Τεχνητή Σούπερ Νοημοσύνη (Artificial Super Intelligence / ASI), γνωστή και ως υπερ ευφυΐα, θα ξεπερνούσε την ευφυΐα και την ικανότητα του ανθρώπινου εγκεφάλου. Παρόλο που η ισχυρή τεχνητή νοημοσύνη αποτελεί ένα καθόλα θεωρητικό πεδίο της τεχνητής νοημοσύνης χωρίς την ύπαρξη πρακτικών εφαρμογών, υπάρχει σαν πεδίο έρευνας και η ανάπτυξή του μελετάται από τους επιστήμονες.

Στην τεχνητή νοημοσύνη εντάσσεται η βαθιά μάθηση (deep learning) και η μηχανική μάθηση (machine learning) που αποτελούνται από αλγόριθμους τεχνητής νοημοσύνης και πιο συγκεκριμένα, η

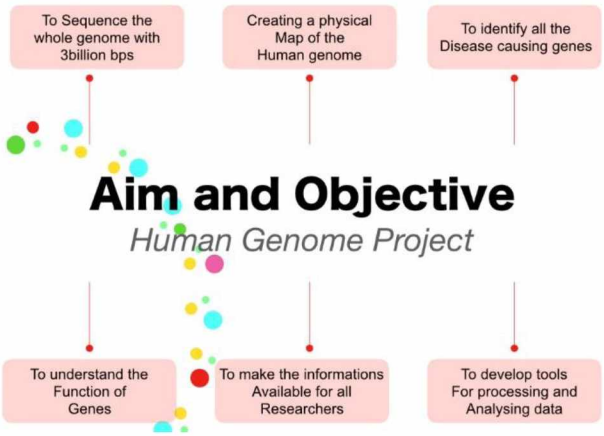
βαθιά μάθηση αποτελεί τμήμα της μηχανικής. Για αυτόν τον λόγο και η διάκριση ανάμεσά τους δεν είναι εύκολη υπόθεση και συχνά αυτές συγχέονται. Σκοπός και των δύο αποτελεί η ανάπτυξη μεθόδων και συστημάτων ταξινόμησης και πρόβλεψης έχοντας ως αφετηρία ένα αρχικό σύνολο δεδομένων. Η βαθιά μάθηση στην πραγματικότητα αποτελείται από νευρωνικά δίκτυα. Το "Deep" στη βαθιά μάθηση αναφέρεται σε ένα νευρωνικό δίκτυο που αποτελείται από περισσότερα από τουλάχιστον τρία επίπεδα τα οποία περιλαμβάνουν τις εισόδους, τα κρυφά επίπεδα και τις εξόδους. Ένα τέτοιο νευρωνικό δίκτυο μπορεί να θεωρηθεί αλγόριθμος βαθιάς μάθησης. Ο τρόπος με τον οποίο η βαθιά μάθηση και η μηχανική μάθηση διαφέρουν είναι στον τρόπο που μαθαίνει ο κάθε αλγόριθμος. Η μηχανική μάθηση χρειάζεται πιο δομημένα αρχικά δεδομένα προκειμένου το δίκτυο να μπορέσει να μάθει από αυτά και η αποτελεσματικότητά της εξαρτάται σε μεγάλο βαθμό από τον άνθρωπο. Αντίθετα στην βαθιά μάθηση ένα μεγάλο τμήμα της επεξεργασίας των εισαγόμενων δεδομένων αυτοματοποιείται και δεν είναι απαραίτητη η ανθρώπινη παρέμβαση. Συνεπώς είναι ευκολότερη η εισαγωγή μεγαλύτερων συνόλων δεδομένων. Βέβαια οι διακρίσεις όπως αναφέρθηκε είναι λεπτές καθώς ακόμα και στο πλαίσιο της βαθιάς μάθησης χρειάζεται κάποιου είδους προ επεξεργασίας των αρχικών δεδομένων προτού εισαχθούν στο νευρωνικό δίκτυο.



Εικόνα 4 Διάκριση ανάμεσα στις έννοιες μηχανική μάθηση, βαθιά μάθηση και τεχνητή νοημοσύνη

Κατά καιρούς έχουν αναπτυχθεί μέθοδοι που χρησιμοποιούν τεχνητή νοημοσύνη και μηχανική μάθηση με σκοπό την επεξεργασία και ανάλυση δεδομένων μεγάλου όγκου. Ειδικά όσον αφορά τον τομέα της βιοπληροφορικής, η τεχνητή νοημοσύνη έχει πολλές εφαρμογές και μπορεί να βοηθήσει στην λεπτομερέστερη και αποδοτικότερη ανάλυση των δεδομένων και την εύρεση μηχανισμών και σχέσεων που ο άνθρωπος αδυνατεί να βρει και να κατανοήσει. Για παράδειγμα η ανάλυση του ανθρώπινου γονιδιώματος ή the Human Genome Project (HGP), είναι μια από τις μεγαλύτερες προκλήσεις που

ξεκίνησε το 2008 και έχει σκοπό να αλληλουχίσει ολόκληρο το γονιδίωμα χιλιάδων ανθρώπων από ολόκληρο τον κόσμο (Laura Clarke, Susan Fairley et al., 2017). Το πρόγραμμα αυτό συνεχίζει να αναπτύσσεται ακόμη και σήμερα και αποτελεί ίσως το μεγαλύτερο σύνολο δεδομένων σε ολόκληρο τον κόσμο που περιλαμβάνει τόσες ανθρώπινες γενετικές παραλλαγές. Σκοπός του προγράμματος είναι να εφαρμόσει και να χρησιμοποιήσει όλη αυτήν την γνώση από τους διαθέσιμους γονότυπους, για να αναγνωρίσει πιθανούς φαινότυπους ασθενειών και να αναπτύξει προσωποποιημένα και πιο αποτελεσματικά φάρμακα.



Εικόνα 5 Στόχοι του προγράμματος χαρτογράφησης του ανθρώπινου γονιδιώματος

1.6. Ανάλυση βιολογικών δεδομένων με μεθόδους μηχανικής μάθησης

Όπως αναφέρθηκε τα βιολογικά δεδομένα είναι πολλών τύπων και προέρχονται από πολλές και διαφορετικές πηγές. Τα δεδομένα που προέρχονται από τεχνολογίες αλληλούχισης, όπως NGS και scRNA-Seq χρειάζονται πιο ειδικούς τρόπους ανάλυσης, λόγω των ιδιαίτερων χαρακτηριστικών τους και έχουν προβληματίσει τους επιστήμονες των δεδομένων για το αν και πόσο αποδοτικές είναι οι ήδη διαθέσιμες μέθοδοι με την χρήση της τεχνητής νοημοσύνης ή αν και πώς θα πρέπει να αναπτυχθούν τρόποι που λαμβάνουν υπόψη τους τα χαρακτηριστικά αυτά. Πάντως είναι σημαντικό να αναφερθεί ότι η επίλυση ενός τέτοιου ζητήματος απαιτεί την συνεργασία επιστημόνων τόσο από τον τομέα της επιστήμης των δεδομένων και της πληροφορικής, όσο και από τον τομέα της γενετικής και της βιολογίας. Έχουν αναπτυχθεί λοιπόν μέσα από αυτήν την συνεργασία κάποιοι τρόποι που αυξάνουν την ακρίβεια των ήδη υπαρχόντων μεθόδων λαμβάνοντας υπόψη τους τη μεγάλη ετερογένεια και το πλήθος των μηδενικών τιμών των δεδομένων. Χρειάζεται όμως παραπάνω προσπάθεια για να αναπτυχθούν ακόμη πιο αποδοτικοί και ειδικοί τρόποι για την επεξεργασία και ανάλυση των δεδομένων από scRNA-Seq και NGS που χρησιμοποιούν και εκμεταλλεύονται τα θετικά της

μηχανικής μάθησης για εφαρμογές όπως η ταξινόμηση, η ομαδοποίηση και η μείωση των διαστάσεών τους. Έτσι θα γίνει ακόμη πιο εύκολη η ανάλυση και η εύρεση μοτίβων και τάσεων σε αυτού του είδους τα δεδομένα.

Στα βιολογικά δεδομένα μεγάλου όγκου η μείωση των διαστάσεών τους αποτελεί ένα πρώτο στάδιο προ επεξεργασίας, προβάλλοντας τον υψηλής διάστασης χώρο με τις μετρήσεις της γονιδιακής έκφρασης σε έναν χώρο μικρότερης διάστασης. Ο βασικός σκοπός της διαδικασίας αυτής είναι η αναγνώριση μοτίβων ανάμεσα σε δείγματα που είναι δύσκολο να βρεθούν όταν αυτά βρίσκονται σε χώρο πολλών διαστάσεων. Όσον αφορά τα δεδομένα που προκύπτουν από την χρήση της τεχνολογίας scRNA-Seq, αυτά παρουσιάζουν κάποιες δυσκολίες οι οποίες τα καθιστούν δύσκολα στην επεξεργασία και ανάλυσή τους και η μείωση των διαστάσεών τους θεωρείται περίπλοκη. Στα δεδομένα αυτά είναι δυσκολότερη η διάκριση μοτίβων ή συνδέσεων και οι λόγοι που συμβαίνει αυτό αναλύονται αργότερα περεταίρω. Γενικοί μέθοδοι μείωσης διαστάσεων που χρησιμοποιούνται αποδοτικά για όλων των ειδών τα δεδομένα είναι η PCA (Principal Component Analysis) και η t-SNE (t – distributed Stochastic Neighbor Embedding) που έχουν αρκετά καλή ακρίβεια και στην ταξινόμηση των δειγμάτων από scRNA-Seq. Όμως έχουν αναπτυχθεί και πιο εξειδικευμένοι τρόποι μείωσης διαστάσεων που λαμβάνουν υπόψη τα χαρακτηριστικά των δεδομένων όπως η ZIFA (Zero Inflated Factor Analysis). Οι πιο γνωστές μέθοδοι μείωσης διαστάσεων μεγάλου όγκου δεδομένων περιγράφονται πιο αναλυτικά σε επόμενο κεφάλαιο.

1.7. Βασική ορολογία μηχανικής μάθησης

Εκτός από την μείωση της διάστασής τους, υπάρχουν και άλλου είδους μέθοδοι επεξεργασίας δεδομένων πολλών ή λίγων διαστάσεων. Η επιλογή της κατάλληλης μεθόδου μπορεί να εξαρτηθεί από την ποιότητα ή το είδος των διαθέσιμων δεδομένων. Αν για παράδειγμα διατίθενται δεδομένα με ετικέτα ή «labeled», δηλαδή δεδομένα που γνωρίζουμε αν ανήκουν σε κάποια συγκεκριμένη ομάδα ή κατηγορία (πχ καρκινικά ή μη κύτταρα) τότε συνήθως αυτά χρησιμοποιούνται για εκπαίδευση νευρωνικών δικτύων. Τα νευρωνικά δίκτυα, χρησιμοποιώντας μεθόδους μηχανικής ή βαθιάς μάθησης βρίσκουν μοτίβα στα δεδομένα και μπορούν, αν χρειαστεί, να χρησιμοποιήσουν αυτή τη γνώση που απέκτησαν για να κατηγοριοποιήσουν αργότερα δεδομένα χωρίς ετικέτα ή «unlabeled». Η διαδικασία αυτή ονομάζεται ταξινόμηση ή «classification». Πιο συγκεκριμένα στη μηχανική μάθηση, η ταξινόμηση αναφέρεται σε ένα πρόβλημα προγνωστικής μοντελοποίησης (predictive modeling) όπου προβλέπεται μια ετικέτα κλάσης για ένα συγκεκριμένο δείγμα από τα δεδομένα εισόδου.

Χαρακτηριστικός αλγόριθμος ταξινόμησης είναι ο KNN (K-Nearest Neighbor) που μπορεί να χρησιμοποιηθεί για την επίλυση προβλημάτων τόσο ταξινόμησης όσο και παλινδρόμησης (regression). Ο αλγόριθμος αυτός είναι εύκολο να εφαρμοστεί, αλλά επιβραδύνεται σημαντικά όσο αυξάνεται το μέγεθος των δεδομένων που χρησιμοποιούνται. Ο αλγόριθμος KNN στη φάση εκπαίδευσης αποθηκεύει απλώς το σύνολο των δεδομένων και όταν εισάγεται το σύνολο δεδομένων ελέγχου, τότε ταξινομεί κάθε δείγμα στην κατηγορία ή κλάση που τα στοιχεία της έχουν περισσότερα κοινά με αυτό. Ο KNN λειτουργεί βρίσκοντας τις αποστάσεις μεταξύ ενός συγκεκριμένου σημείου με όλα τα υπόλοιπα, επιλέγοντας τα K πιο κοντινά και υπολογίζει τον μέσο όρο των ετικετών (σε περίπτωση παλινδρόμησης) ή βρίσκει την πιο συχνή ετικέτα (στο περίπτωση ταξινόμησης) την οποία αποδίδει σε αυτό το στοιχείο.

Αν τα δεδομένα που διαθέτουμε είναι unlabeled τότε η ανάλυσή τους απαιτεί άλλου είδους διαδικασία που ονομάζεται «clustering» ή ομαδοποίηση. Σε αυτήν την διαδικασία τον νευρωνικό δίκτυο δεν έχει πληροφορίες για το που ανήκει το κάθε διαθέσιμο δείγμα και προσπαθεί να εξάγει αυτήν την πληροφορία μέσα από εύρεση κοινών χαρακτηριστικών των δεδομένων τα οποία και τελικά ομαδοποιεί. Ειδικότερα η ομαδοποίηση χρησιμοποιεί μόνο τα δεδομένα εισόδου για εύρεση φυσικών ομάδων στο χώρο χαρακτηριστικών. Ένας χαρακτηριστικός αλγόριθμος ομαδοποίησης είναι ο αλγόριθμος k – means, που προσπαθεί να ομαδοποιήσει παρόμοια δείγματα ή σημεία με τη μορφή συστάδων. Ο αριθμός των ομάδων αντιπροσωπεύεται από έναν αριθμό K. Ο k-means υπολογίζει τα κέντρα των ομάδων και επαναλαμβάνει μέχρι να βρει το βέλτιστο κέντρο. Σε αυτόν τον αλγόριθμο, σκοπός είναι το άθροισμα της τετραγωνικής απόστασης μεταξύ των σημείων και του κέντρου να είναι ελάχιστο. Υπάρχουν πολλοί αλγόριθμοι ομαδοποίησης και ταξινόμησης που ο κάθε ένας έχει διαφορετικά θετικά και περιορισμούς.

Γενικά οι μέθοδοι που χρησιμοποιούν δεδομένα με ετικέτα ανήκουν στην κατηγορία των «Supervised Learning» μεθόδων ή μεθόδων επιβλεπόμενης μάθησης, ενώ οι μέθοδοι που χρησιμοποιούν δεδομένα χωρίς ετικέτα ανήκουν στην κατηγορία των «Unsupervised Learning» μεθόδων ή μεθόδων μη επιβλεπόμενης μάθησης. Υπάρχουν ακόμη και μέθοδοι που συνδυάζουν τις δύο παραπάνω κατηγορίες και ονομάζονται «Semi Supervised Learning» μέθοδοι. Οι Autoencoders που θα αναλυθούν περεταίρω σε επόμενο κεφάλαιο ανήκουν στην κατηγορία των unsupervised learning μεθόδων, αν και πρακτικά χρησιμοποιούν μεθόδους επιβλεπόμενης μάθησης για να εκπαιδευτούν, γεγονός που τους καθιστά «self – Supervised» (A Gentle Introduction to LSTM Autoencoders, Nov. 2018).

1.8. Χρήση μεθόδων συλλογικής μάθησης

Η ακρίβεια και η ταχύτητα της ταξινόμησης (classification) όπως αναφέρθηκε μειώνεται όσο αυξάνεται ο όγκος των δεδομένων, αλλά επηρεάζεται και αν ο αλγόριθμος εφαρμοστεί απευθείας στα αρχικά δεδομένα χωρίς κάποια μορφή προεπεξεργασίας τους. Η ακρίβεια του μπορεί ακόμη να επηρεαστεί από ένα μη ισορροπημένο σύνολο δεδομένων, τον υψηλό θόρυβο στα δεδομένα, ελλιπείς τιμές, περιττά δεδομένα κ.α.. Έχει παρατηρηθεί ότι ειδικά οι μέθοδοι ταξινόμησης είναι πολύ χρήσιμοι και έχουν πολλές εφαρμογές στον τομέα της βιοπληροφορικής. Όμως τα βιολογικά δεδομένα είναι συνήθως μεγάλου όγκου και αυτό δημιουργεί πρόβλημα στην χρήση μεθόδων επιβλεπόμενης μάθησης σε αυτά, καθώς η απόδοσή τους μειώνεται. Ακόμη και μετά από την εφαρμογή μεθόδων μείωσης διαστάσεων, ευρέως χρησιμοποιούμενες μέθοδοι επιβλεπόμενης μάθησης όπως ο απλοϊκός κατηγοριοποιητής bayes (naïve bayes), τα δέντρα αποφάσεων (decision trees), ο αλγόριθμος των K κοντινότερων γειτόνων (KNN) και οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines / SVM) είναι εξαιρετικά δύσκολο να εφαρμοστούν αποδοτικά σε δεδομένα πολλών διαστάσεων.

Για τους παραπάνω λόγους έχουν αναπτυχθεί μέθοδοι συλλογικής μάθησης ή «Ensemble Learning methods» οι οποίες συνδυάζουν πολλαπλά μοντέλα μηχανικής μάθησης με σκοπό την δημιουργία ισχυρότερων, που τελικά έχουν μεγαλύτερη ακρίβεια. Για παράδειγμα, οι Ensemble Classifiers έχουν μεγαλύτερη ακρίβεια στην ταξινόμηση δεδομένων μεγάλου όγκου σε σχέση με τους απλούς αλγόριθμους ταξινόμησης. Οι απλοί ταξινομητές δεν μπορούν να χειριστούν με την ίδια ευκολία τον θόρυβο και τα μη ισορροπημένα δεδομένα μεγάλου όγκου. Σε γενικές γραμμές υπάρχουν αρκετές προτεινόμενες μέθοδοι συλλογικής μάθησης που εφαρμόζονται ακόμα και σε βιολογικά δεδομένα μεγάλου όγκου, όπως για παράδειγμα σε δεδομένα από scRNA-Seq, άλλες περισσότερο και άλλες λιγότερο ακριβείς.

1.9. Περίληψη εργασίας

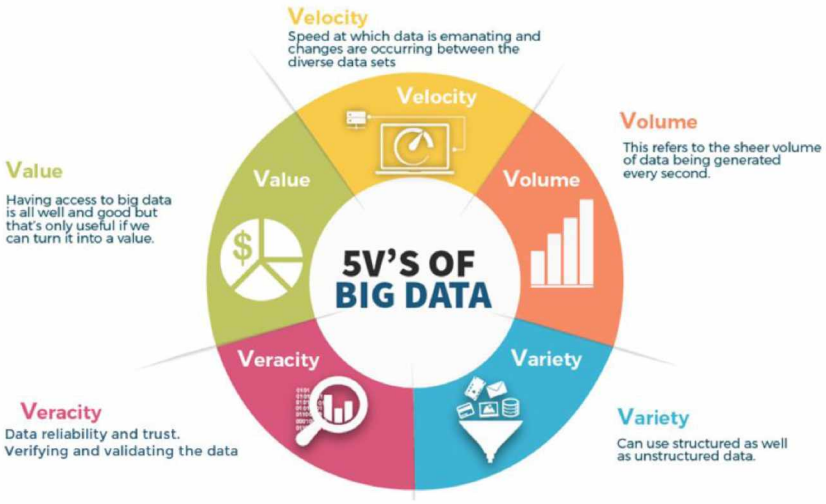
Στην παρούσα εργασία έχει γίνει χρήση Autoencoder, ενός νευρωνικού δικτύου κρυπτογράφησης – αποκρυπτογράφησης και συλλογικών ταξινομητών σε πολυδιάστατα δεδομένα από scRNA-Seq. Σκοπός είναι αρχικά η μείωση της διάστασής τους και η εξαγωγή ενός χώρου μειωμένης διάστασης από το αρχικό σύνολο δεδομένων με την χρήση ενός αριθμού (r) Autoencoder που μπορεί να προσδιοριστεί από το χρήστη. Πιο αναλυτικά, κάθε φορά που εισάγονται τα ίδια δεδομένα στο νευρωνικό αυτό δίκτυο, εκείνο βρίσκει διαφορετικά μοτίβα και συνδέσεις ανάμεσα στα δεδομένα, με αποτέλεσμα να προκύπτει κάθε φορά διαφορετικός χώρος μειωμένης διάστασης ακόμη και για το ίδιο

σύνολο αρχικών δεδομένων. Στην συνέχεια ο χώρος αυτός, που προκύπτει από το εσωτερικό τμήμα του Autoencoder, χωρίζεται σε δέκα επιμέρους υποχώρους μέσω μιας διαδικασίας 10-fold cross validation. Για κάθε έναν από αυτούς τους υποχώρους χρησιμοποιούνται όλοι οι υπόλοιποι με σκοπό την εκπαίδευση του κάθε ταξινομητή και η διαδικασία επαναλαμβάνεται δέκα φορές. Στο τέλος, τα αποτελέσματα των διαφορετικών ταξινομητών αποθηκεύονται και δημιουργείται ένας συλλογικός ταξινομητής. Η διαδικασία επαναλαμβάνεται r φορές με τη χρήση διαφορετικών Autoencoder και κατά συνέπεια διαφορετικών χώρων μειωμένης διάστασης για το ίδιο κάθε φορά σύνολο δεδομένων. Αφού αποθηκευτούν οι προβλέψεις των κλάσεων για κάθε έναν από τους r Autoencoders, τότε ως τελική κλάση κάθε δείγματος επιλέγεται εκείνη που προβλέφθηκε πιο συχνά (majority class) από τους συλλογικούς ταξινομητές (ensemble classifiers). Η διαδικασία αυτή αποτελεί μέθοδο συλλογικής μάθησης, αφού χρησιμοποιούνται βασικοί classifiers και νευρωνικά δίκτυα (VAE's) που τα αποτελέσματά τους στο τέλος της διαδικασίας συνδυάζονται για να δημιουργηθεί ένα ισχυρότερο μοντέλο. Περισσότερες λεπτομέρειες για το πρακτικό και θεωρητικό μέρος της εργασίας, αναφέρονται σε επόμενη παράγραφο.

Όσον αφορά την δομή της εργασίας, αρχικά γίνεται μια αναφορά στα δεδομένα πολλών διαστάσεων και τις τρεις διαστάσεις που αυτά εκτείνονται με βάση τον D. Laney αλλά και στα προβλήματα που προκύπτουν εξαιτίας της μεγάλης τους διαστατικότητας. Στην συνέχεια, διευκρινίζονται οι πιο σημαντικές ιδιαιτερότητες των πολυδιάστατων βιολογικών δεδομένων και οι λόγοι που αυτά χρίζουν διαφορετικής προσέγγισης προκειμένου να αναλυθούν και να επεξεργαστούν. Ακόμη, γίνεται μια σύντομη ανάλυση των γραμμικών και μη γραμμικών μεθόδων που χρησιμοποιούνται για την μείωση διάστασης και εξαγωγής χαρακτηριστικών των πολυδιάστατων δεδομένων, αλλά και των μεθόδων που μπορούν να χρησιμοποιηθούν για μείωση της διάστασης των βιολογικών δεδομένων και πιο συγκεκριμένα των δεδομένων scRNA-seq. Τέλος, παρουσιάζεται αναλυτικά το μοντέλο που αναπτύχθηκε στα πλαίσια της εργασίας, με τα αποτελέσματα που προέκυψαν αλλά και σχετικές προσεγγίσεις του θέματος από άλλες πηγές ή με την χρήση άλλων μεθόδων. Ο κώδικας που αναπτύχθηκε με κάποιες επεξηγήσεις βρίσκεται σε ειδικό παράρτημα στο τέλος της εργασίας.

2. Δεδομένα πολλών διαστάσεων

Ο Douglas Laney (Doug Laney, 2001) παρατήρησε ότι τα δεδομένα μεγάλου όγκου αυξάνονται σε τρεις διαφορετικές διαστάσεις, γνωστές και ως τρία V. Αυτές αποτελούν την ποσότητα ή όγκο των δεδομένων (volume), τον ρυθμό με τον οποίο τα δεδομένα συλλέγονται και διατίθενται για χρήση (velocity) και την ποικιλία ή τους τύπους (variety) των δεδομένων. Πολλοί προσπάθησαν να προσθέσουν περισσότερες διαστάσεις στον παραπάνω ορισμό, όμως ως περισσότερο αποδεκτή θεωρήθηκε η προσθήκη της διάστασης της «εγκυρότητας» ή «veracity» των δεδομένων που βοηθά στον διαχωρισμό των σημαντικών από τα μη σημαντικά δεδομένα, και στο τέλος, δημιουργεί μια βαθύτερη κατανόησή τους και του τρόπου με τον οποίο μπορούν να χρησιμοποιηθούν σε διάφορες εφαρμογές.



Εικόνα 6 Τα χαρακτηριστικά των μεγάλων δεδομένων, όπως διατυπώνονται σήμερα, με όρους που εμπλούτισαν τον ορισμό του Douglas Laney

Όμως τα δεδομένα πολλών διαστάσεων έχουν και άλλα χαρακτηριστικά που τα διαχωρίζουν από τα δεδομένα λίγων διαστάσεων. Ένα από αυτά είναι ο λόγος που τα δεδομένα συλλέγονται και ο χρόνος που αυτά παραμένουν αποθηκευμένα σε κάποιο σύστημα. Τα δεδομένα πολλών διαστάσεων ενώ συλλέγονται αρχικά με κάποιο στόχο, χρησιμοποιούνται συνήθως για διαφορετικούς και απρόβλεπτους από την αρχή λόγους. Για αυτόν τον λόγο, παραμένουν αποθηκευμένα για μεγαλύτερα χρονικά διαστήματα και μοιράζονται σε διαφορετικούς υπολογιστές ή ακόμη και διαφορετικές γεωγραφικές τοποθεσίες, είτε για λόγους ασφάλειας, είτε για λόγους χωρητικότητας. Επιπλέον, ενώ τα δεδομένα μικρού όγκου είναι συνήθως καλώς δομημένα και οργανωμένα, τα δεδομένα μεγάλου όγκου είναι συνήθως μη δομημένα και μπορεί να έχουν πολλές μορφές στα διαφορετικά αρχεία που

αποθηκεύονται, τα οποία ακολουθούν διαφορετικές αρχές και μπορεί να συνδέονται μέσω άλλων πόρων. Τέλος, ακόμη άλλη μια σημαντική διαφορά είναι ότι τα μεγάλα δεδομένα χρειάζονται περισσότερη προετοιμασία και έλεγχο για να χρησιμοποιηθούν, καθώς μπορεί να μην είναι ενημερωμένα, να χαθεί κάποιο σημαντικό μέρος που τα αχρηστεύει, να χρησιμοποιούνται σε αυτά διαφορετικές μετρικές ανάλογα με την γεωγραφική περιοχή που συλλέχθηκαν κ.α..

Για όλους τους παραπάνω λόγους είναι εμφανές ότι αυτού του είδους τα δεδομένα χρειάζονται άλλους τρόπους χειρισμού και εμφανίζουν διαφορετικές προκλήσεις και δυσκολίες σε σχέση με τα δεδομένα λίγων διαστάσεων.

2.1. Προκλήσεις και δυσκολίες

Ένα μεγάλο πρόβλημα των δεδομένων πολλών διαστάσεων που έχει ήδη αναφερθεί είναι ότι υπάρχουν δυσκολίες όσον αφορά την αποθήκευση, την μεταφορά, την αξιοπιστία, την προστασία και την ασφάλειά τους (θέματα προσωπικών δεδομένων, νόμος GDPR και προστασία από υποκλοπές). Ακόμη ένα μεγάλο πρόβλημα αποτελεί η οπτικοποίηση των δεδομένων μεγάλου όγκου. Μέσω της οπτικοποίησης γίνεται ευκολότερη η εύρεση μοτίβων και ομάδων (clusters) στα δεδομένα, όμως αν οι διαστάσεις τους είναι πολλές, η οπτικοποίηση δεν βοηθά στην ανάλυσή τους. Ακόμη είναι πολύ δύσκολο να βρεθούν σχέσεις ανάμεσα στις διαστάσεις των δεδομένων που δεν σχετίζονται με άλλες (λόγω του colinearity) και δίνουν κάποια μοναδική και χρήσιμη πληροφορία. Όμως έχουν γίνει και πιο συγκεκριμένες αναφορές σε κλασικά προβλήματα που προκύπτουν λόγω της ποσότητας αυτών των δεδομένων (Donoho, 2000).

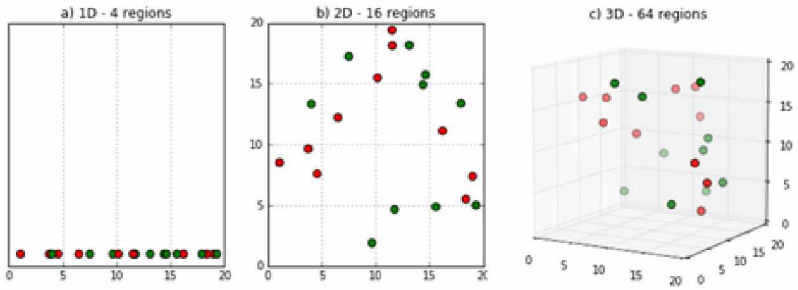
Στο παρελθόν, τα δεδομένα που χρησιμοποιούνταν, επεξεργάζονταν και αναλύονταν χαρακτηρίζονταν από μεγάλο αριθμό δειγμάτων, για καθένα από τα οποία αναλογούσε ένας μικρός αριθμός παρατηρήσεων. Αν, λοιπόν, θεωρήσουμε p την διάσταση του «άγνωστου» και n το πλήθος όσων είναι «γνωστά», τότε, με βάση την θεωρία, ισχύει το σενάριο «small p, large n». Αυτό το σενάριο αντανακλούσε και τους σύγχρονους περιορισμούς των υπολογιστών (Johnstone and Titterington, 2009). Αν λοιπόν προσπαθήσουμε ασυμπτωτικά να προσδιορίσουμε αυτό το σενάριο καταλήγουμε στο ότι όσο το $n \to \infty$, το p παραμένει μικρότερης τάξης από το n και στην πραγματικότητα (συνήθως) σταθερό. Το n χρησιμοποιείται για να περιγράψει το μέγεθος του δείγματος και το p για τις διαστάσεις. Μεταξύ των πιο γνωστών θεωρητικών συμπερασμάτων αυτού του τύπου είναι ο Νόμος των Μεγάλων Αριθμών (Laws of Large Numbers) και το Θεώρημα κεντρικού ορίου (the Central Limit Theorem).

Σύμφωνα με τον νόμο των Μεγάλων Αριθμών, ο δειγματικός μέσος ενός τυχαίου δείγματος μεγέθους n από έναν πληθυσμό μπορεί να προσδιοριστεί στατιστικά και η τιμή αυτή είναι ίση με τον μέσο όρο του πληθυσμού, όσο το $n \to \infty$. Για παράδειγμα, αν σε ένα ζάρι υπολογιστεί η μέση τιμή όλων των τιμών των πλευρών του προκύπτει ο αριθμός 3,5. Τότε σύμφωνα με τον παραπάνω νόμο, αν το ζάρι ριφθεί πολλές φορές, η μέση τιμή του δείγματος θα εξακολουθήσει να είναι το 3,5 και η διαπίστωση αυτή γίνεται ακόμη πιο ακριβής όσο αυξάνουμε τον αριθμό των ρίψεων.

Με βάση το Θεώρημα κεντρικού ορίου, η μέση κατανομή του δείγματος μιας τυχαίας μεταβλητής τείνει να γίνει κανονική, αν αθροιστούν οι ανεξάρτητες μεταβλητές που ανήκουν στο δείγμα και εάν το μέγεθος του δείγματος είναι αρκετά μεγάλο. Αυτό μπορεί να εξηγηθεί από το ότι ο μέσος όρος του δείγματος της κατανομής είναι ο πραγματικός μέσος όρος του πληθυσμού από τον οποίο ελήφθησαν τα δείγματα. Η διακύμανση της κατανομής του δείγματος, από την άλλη πλευρά, είναι η διακύμανση του πληθυσμού διαιρούμενη με το μέγεθος του δείγματος. Επομένως, όσο μεγαλύτερο είναι το μέγεθος του δείγματος της κατανομής, τόσο μικρότερη είναι η διακύμανση του μέσου όρου του.

Τέτοιου είδους θεωρήματα είναι πολύ χρήσιμα για την στατιστική, καθώς οδηγούν σε έγκυρα συμπεράσματα και έχουν αναρίθμητες πρακτικές εφαρμογές. Τα τελευταία χρόνια όμως ο ρυθμός συλλογής και αποθήκευσης δεδομένων έχει αυξηθεί, όπως και οι υπολογιστικές εγκαταστάσεις έχουν βελτιωθεί. Η εξέλιξη αυτή έχει οδηγήσει σε καταστάσεις που μέχρι πριν ήταν αδύνατο να αντιμετωπιστούν, όπως την συλλογή λίγων δειγμάτων τα οποία όμως έχουν πολλά χαρακτηριστικά ή διαστάσεις. Έτσι επικράτησε η περίπτωση «large p, small n» ή σε κάποιες άλλες περιπτώσεις «large p, large n». Στην πρώτη, φαίνεται το p να τείνει στο άπειρο γρηγορότερα από ότι το n, ενώ στην δεύτερη τα p και n τείνουν στο άπειρο με τον ίδιο ρυθμό.

Όταν λοιπόν για το σύνολο δεδομένων ισχύει ότι p >> n, δηλαδή ο χώρος των χαρακτηριστικών είναι πολύ μεγαλύτερος από τον αριθμό των δειγμάτων, τότε ισχύει και η «κατάρα της διαστατικότητας» ή «curse of dimensionality». Η έκφραση επινοήθηκε από τον Richard E. Bellman κατά την εξέταση προβλημάτων στη δυναμική βελτιστοποίηση (Bellman R.E., 1961).



Εικόνα 7 Η κατάρα της διαστατικότητας. Το σφάλμα αυξάνεται με την αύξηση του αριθμού των χαρακτηριστικών. Κάθε χαρακτηριστικό προσθέτει πληροφορίες και αν ληφθούν όλες υπόψη, μπορεί κάθε ένα δείγμα να διαχωριστεί τέλεια με τα υπόλοιπα. Ωστόσο, ένας άπειρος αριθμός χαρακτηριστικών απαιτεί έναν άπειρο αριθμό παραδειγμάτων εκπαίδευσης, εξαλείφοντας τη χρησιμότητα του νευρωνικού δικτύου στα δεδομένα του πραγματικού κόσμου.

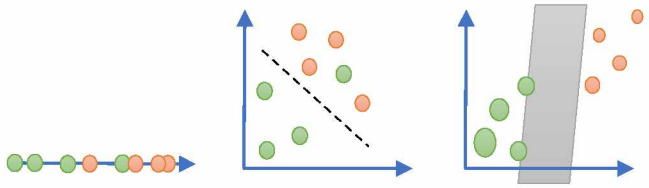
Έστω ότι διατίθενται 2 παρατηρήσεις, σε έναν ευκλείδειο χώρο με καρτεσιανές συντεταγμένες. Τότε για να υπολογίσουμε την ευκλείδεια απόσταση ανάμεσά τους χρησιμοποιούμε τον ακόλουθο και γνωστό τύπο:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Όπου n οι διαστάσεις του χώρου. Είναι εμφανές ότι για κάθε μια διάσταση που προστίθεται, η απόσταση των σημείων x,y αυξάνεται. Κάθε μια νέα διάσταση λοιπόν, αυξάνει τον χώρο των χαρακτηριστικών και αφού ο αριθμός των παρατηρήσεων παραμένει σταθερός, τότε ο χώρος αυτός γίνεται και πιο αραιός. Κατά συνέπεια η μέση απόσταση ανάμεσα στις παρατηρήσεις αυξάνεται, αφού ο χώρος μεγαλώνει και τα σημεία – παρατηρήσεις απομακρύνονται. Πιο συγκεκριμένα αν θεωρήσουμε ότι αυτές οι παρατηρήσεις ορίζουν μια ευθεία 45^o με τους καρτεσιανούς άξονες, τότε η απόσταση σε σχέση με τον αριθμό των διαστάσεων υπολογίζεται από την εξίσωση : $(x-y)\cdot\sqrt{n}$. Επίσης, αν θεωρήσουμε ένα καρτεσιανό σύστημα συντεταγμένων με μέγεθος 1/10 του μοναδιαίου κύβου, τότε για 10 διαστάσεις θα χρειαστούμε 10^10 σημεία, για κύβο με 20 διαστάσεις 10^20 κλπ.. Όσο ο αριθμός των διαστάσεων αυξάνεται τότε αντιλαμβανόμαστε ότι, με αυτόν τον ρυθμό, θα χρειαστεί να διαθέτουμε δεκάδες τρισεκατομμύρια σημεία – παρατηρήσεις.

Όσον αφορά την μηχανική και την βαθιά μάθηση, ειδικά για το κομμάτι της ομαδοποίησης ή της ταξινόμησης των δεδομένων και γενικότερα για όσους αλγόριθμους επεξεργάζονται και αναλύουν

δεδομένα με βάση τις αποστάσεις των σημείων, η μεγάλη διαστατικότητα και κατά συνέπεια ο αραιός χώρος των χαρακτηριστικών κάνει όλες τις παρατηρήσεις να φαίνεται ότι ισαπέχουν μεταξύ τους. Δηλαδή, αφού λόγω της «κατάρας της διαστατικότητας» δεν έχει διατηρηθεί η πληροφορία της μέσης απόστασης των δεδομένων, φαίνεται ότι όλα τα δεδομένα είναι το ίδιο κοντά ή μακριά σε σχέση με τα υπόλοιπα. Αυτό έχει ως αποτέλεσμα, στην προκειμένη περίπτωση της ομαδοποίησης, να μην μπορέσουν να δημιουργηθούν ομάδες με νόημα, αφού πρακτικά δεν υπάρχουν πληροφορίες χρήσιμες που θα εκμεταλλευτεί ο αλγόριθμος για να εξάγει κάποιο συμπέρασμα. Άλλες φορές, επειδή η πυκνότητα των παρατηρήσεων μειώνεται εκθετικά με την αύξηση των διαστάσεων, είναι πολύ πιθανό ο αλγόριθμος ή το μοντέλο που χρησιμοποιείται να μπορέσει να βρει μια γραμμή που να διαχωρίζει τέλεια τα δείγματα, οδηγώντας σε υπερπροσαρμογή ή «overfitting». Αυτό σημαίνει ότι το μοντέλο μπορεί να διαχωρίσει άριστα το σύνολο των παρατηρήσεων στις οποίες έχει εκπαιδευτεί, όμως παρουσιάζει δραματικά μικρή ακρίβεια στην πρόβλεψη νέων. Άρα η απόδοση των αλγόριθμων που χρησιμοποιούν αποστάσεις μειώνεται καθώς αυξάνονται οι διαστάσεις. Το πρόβλημα αυτό της μεγάλης διαστατικότητας, από την άλλη, κάποιες φορές ερμηνεύεται και ως «ευλογία», καθώς για τα δεδομένα που δεν είναι γραμμικά διαχωρίσιμα όταν έχουν λίγες διαστάσεις, η αύξηση των διαστάσεών τους μπορεί να τα μετατρέψει σε γραμμικά διαχωρίσιμα, όπου η εύρεση τοπικών ελαχίστων κατά τη διαδικασία της οπισθοδιάδωσης διευκολύνεται, και το πρόβλημα να διορθωθεί. Λύση στο παραπάνω πρόβλημα θεωρείται η μείωση των διαστάσεων των δεδομένων με διάφορες μεθόδους και τεχνικές που περιγράφονται αργότερα σε επόμενο κεφάλαιο.

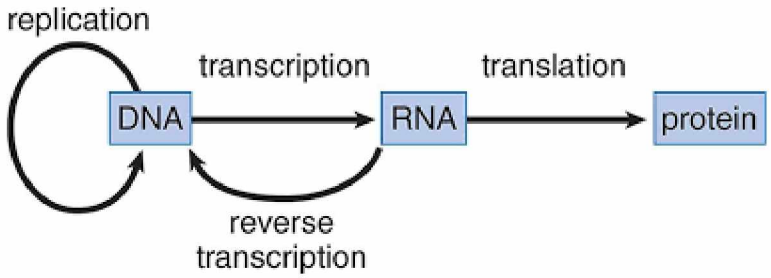


Εικόνα 8 Παράδειγμα βελτίωσης διαχωρισιμότητας των κλάσεων σε μεγαλύτερες διαστάσεις

2.2. Βιολογικά δεδομένα πολλών διαστάσεων

Όσον αφορά τον τομέα της βιολογίας, η επικρατέστερη μέθοδος είναι η αυτόματη, συστηματική συλλογή υπερ-πληροφορίας και λεπτομερειών για κάθε παρατηρούμενο δείγμα. Υπάρχουν διάφοροι τύποι βιολογικών δεδομένων που κάθε ένας από αυτούς δίνει διαφορετικές σημαντικές πληροφορίες που μπορεί να φανούν ιδιαίτερα χρήσιμες αν χρησιμοποιηθούν και ερμηνευθούν σωστά.

Αρχικά σημαντική πληροφορία μπορεί να συλλεχθεί μέσω της παρατήρησης της γονιδιακής έκφρασης των κυττάρων ή διαφορετικά της διαδικασίας που είναι υπεύθυνη για την μετάβαση της γενετικής πληροφορίας από γενιά σε γενιά, με διάφορες εργαστηριακές μεθόδους. Η διαδικασία της γονιδιακής έκφρασης είναι περίπλοκη βιολογικά και περιλαμβάνει την κωδικοποίηση της πληροφορίας που βρίσκεται στον πυρήνα ενός κυττάρου με τη μορφή DNA, την μεταγραφή της σε RNA και τέλος την μετάφρασή της σε πρωτεΐνες. Αν με κάποιον τρόπο γίνει δυνατός ο έλεγχος αυτής της διαδικασίας , τότε αυτό μπορεί να οδηγήσει σε σημαντικές ανακαλύψεις για τον τρόπο που η γενετική πληροφορία κωδικοποιείται και χρησιμοποιείται, ποια τμήματά της είναι σημαντικά και γιατί ή πώς μπορεί κάποιος να επέμβει και να διορθώσει κάποιο πιθανό σφάλμα, που συμβαίνει σε αυτήν την διαδικασία, προλαμβάνοντας κάποια ασθένεια ή ακόμη και τον καρκίνο.



Εικόνα 9 Κεντρικό δόγμα της μοριακής βιολογίας

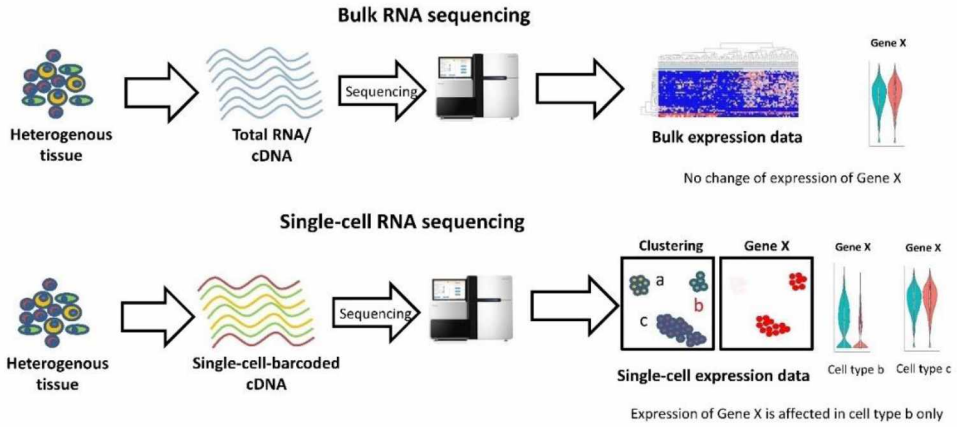
Υπάρχουν διάφοροι τύποι RNA και πρωτεϊνών που παράγονται κάθε φορά και η παραγωγή τους εξαρτάται από το πού ανήκει το συγκεκριμένο κύτταρο ή σε ποιόν ιστό, καθώς άλλα γονίδια εκφράζονται για παράδειγμα σε ένα μυϊκό κύτταρο και άλλα σε ένα νευρικό, με αποτέλεσμα να παράγονται διαφορετικού είδους RNA και πρωτεΐνες. Από τα διαφορετικά RNA που παράγονται ο τύπος που περιέχει τις πληροφορίες για την παραγωγή μιας πρωτεΐνης ονομάζεται αγγελιοφόρος RNA (mRNA), επειδή μεταφέρει την πληροφορία ή το μήνυμα από το DNA έξω από τον πυρήνα στο κυτταρόπλασμα του κυττάρου. Ανάλογα με την ποσότητα των mRNA που μπορούν να βρεθούν στο κυτταρόπλασμα ενός κυττάρου είναι δυνατή η κατανόηση της κύριας πληροφορίας που εκφράζεται στο συγκεκριμένο κύτταρο, αφού μπορούν να παραχθούν πολλά mRNA που κωδικοποιούν την ίδια πληροφορία στο ίδιο κύτταρο και κατά συνέπεια υποδηλώνουν την γονιδιακή έκφραση του συγκεκριμένου κυττάρου.

Η γονιδιακή έκφραση καθορίζεται κυρίως από το ίδιο το κύτταρο μέσω πληροφοριών που βρίσκονται κωδικοποιημένες στο DNA του πυρήνα του. Αν λοιπόν υπάρξει κάποια ανωμαλία στο κύτταρο, αυτή

θα επηρεάσει και την γονιδιακή έκφραση του κυττάρου αυτού. Αυτό μπορεί να προκληθεί από κάποια μετάλλαξη που πιθανώς έχει συμβεί κατά την διάρκεια της αντιγραφής και πολλαπλασιασμού των κυττάρων και έχει επηρεάσει το τμήμα του DNA που ευθύνεται για την γονιδιακή έκφραση. Αυτή η αλλαγή μπορεί να προκαλέσει την υπερέκφραση ενός συγκεκριμένου γονιδίου, την μειωμένη έκφρασή του, ή την μη έκφρασή του, ανάλογα με την σοβαρότητα της μετάλλαξης. Οποιαδήποτε αλλαγή στην γονιδιακή ρύθμιση μπορεί να προκαλέσει προβλήματα σε ολόκληρο τον οργανισμό, όπως για παράδειγμα δυσανεξίες ή προβλήματα με την άμυνα του οργανισμού ή να οδηγήσει ακόμη και σε καρκίνο.

Για να δημιουργηθεί μια μοριακή κατανόηση των κυττάρων και να εξαχθούν χρήσιμες πληροφορίες, τα κύτταρα μπορούν να αξιολογηθούν με διάφορους τρόπους, για παράδειγμα μέσω αναλύσεων αλληλουχιών γονιδιωματικού DNA, της δομής χρωματίνης, των αλληλουχιών αγγελιοφόρου RNA (mRNA), του RNA που δεν κωδικοποιεί κάποια πρωτεΐνη, την έκφραση των πρωτεϊνών, αν αυτές έχουν τροποποιηθεί κ.α.. Οι τεχνολογίες NGS και scRNA-Seq χρησιμοποιούν πληροφορίες σχετικές με την ποσότητα των mRNA μορίων ή αλλιώς το «μεταγράφωμα» στο δείγμα, με σκοπό να βρουν πιθανές αλλαγές στην γονιδιακή έκφραση και να συγκρίνουν την τυπική γονιδιακή έκφραση σε έναν οργανισμό ή κύτταρο, αντίστοιχα, σε σχέση με την έκφραση των γονιδίων σε ένα συγκεκριμένο άτομο ή ομάδα ατόμων. Αυτή η πληροφορία μπορεί να είναι πολύτιμη, καθώς μπορεί να αναγνωρίσει πιθανή προδιάθεση ενός ατόμου για μια ασθένεια, να προλάβει έγκαιρα την εμφάνισή της ή να οδηγήσει σε περισσότερο προσωποποιημένες και εξειδικευμένες μεθόδους θεραπείας.

Η τεχνολογία scRNA-Seq είναι μια γονιδιωματική προσέγγιση για την ανίχνευση και την ποσοτική ανάλυση μορίων mRNA σε ένα βιολογικό δείγμα και είναι χρήσιμη για τη μελέτη κυτταρικών αποκρίσεων. Η συγκεκριμένη τεχνολογία μπορεί να περιγράψει μόρια RNA μεμονωμένων κυττάρων με υψηλή ακρίβεια (Haque *et al.*, 2017). Τα δεδομένα που προκύπτουν από αυτήν την ανάλυση είναι συνήθως πολυδιάστατα και περιέχουν δείγματα ή κύτταρα για καθένα από τα οποία έχει μετρηθεί ο αριθμός των διαφορετικών μεταγραφωμάτων ή αλλιώς γονιδίων που εκφράζονται σε αυτό. Έτσι συλλέγονται πληροφορίες για την γονιδιακή έκφραση σε κάθε κύτταρο του δείγματος, οι οποίες στην συνέχεια αναλύονται και επεξεργάζονται. Τέτοιου είδους δεδομένα χρησιμοποιήθηκαν και στην παρούσα εργασία με σκοπό την μείωση της διάστασής του και την εφαρμογή σε αυτά μεθόδων μηχανικής μάθησης.



Εικόνα 10 Διαφορά NGS με scRNA-seq

2.3. Δυσκολίες και ιδιαίτερα χαρακτηριστικά των βιολογικών δεδομένων μεγάλου όγκου

Τα δεδομένα που προκύπτουν από την scRNA-Seq είναι πολυδιάστατα και περιέχουν τις μετρήσεις των μεταγραφωμάτων που βρέθηκαν σε κάθε διαθέσιμο δείγμα ή κύτταρο. Η ανάλυση τέτοιου είδους δεδομένων παρουσιάζει όμως κάποιες δυσκολίες. Αυτά είναι συνήθως αραιά, με πολλές μηδενικές ή σχεδόν μηδενικές τιμές. Αυτό συμβαίνει αρχικά λόγω της μικρής ποσότητας RNA που καταγράφεται από κάθε κύτταρο, για διάφορους λόγους, όπως για παράδειγμα το γεγονός ότι ο κύκλος ζωής των κυττάρων δεν είναι συγχρονισμένος, οπότε κάθε κύτταρο είναι και σε διαφορετικό στάδιο ανάπτυξης και συνεπώς κύτταρα που ανήκουν στον ίδιο ιστό εκφράζουν διαφορετικά ή άλλης ποσότητας ίδια γονίδια. Ο δεύτερος λόγος οφείλεται σε τεχνικούς παράγοντες που έχουν σχέση με την αποτελεσματικότητα της καταγραφής του RNA, όπως dropout γεγονότα που οδηγούν σε ελλιπείς πληροφορίες.

Η μεγάλη περιεκτικότητα των scRNA-Seq δειγμάτων σε μηδενικές ή σχεδόν μηδενικές τιμές, καθιστά τα δεδομένα αυτά «zero-inflated». «Zero-inflated» ή μηδενικά διογκωμένα ονομάζονται τα δεδομένα εάν το ποσό των παρατηρούμενων μηδενικών είναι μεγαλύτερο από το ποσό των προβλεπόμενων μηδενικών. Σε αυτή την περίπτωση, το μοντέλο δεν μπορεί να υπολογίσει σωστά τα μηδενικά, γεγονός που υποδηλώνει μηδενικό πληθωρισμό στα δεδομένα. Για αυτόν τον λόγο η μείωση των διαστάσεων αυτών των δεδομένων γίνεται περίπλοκη, καθώς είναι δυσκολότερη η διάκριση μοτίβων ή συνδέσεων ανάμεσα στα διαφορετικά δείγματα αφού το μοντέλο δεν μπορεί να υπολογίσει σωστά και να βγάλει λογικά συμπεράσματα από την ύπαρξη όλων αυτών των μηδενικών τιμών.

Αυτά τα μηδενικά δεδομένα που καθιστούν το σύνολο δεδομένων «Zero-Inflated» ονομάζονται «drop-outs». Τα «dropout events» προκαλούνται συνήθως από λάθη που προκύπτουν κατά την διάρκεια της αντίστροφης μεταγραφής των RNA στην διαδικασία της αλληλούχισης. Τα λάθη αυτά οδηγούν σε εσφαλμένη εύρεση της έκφρασης ενός γονιδίου για τα μεμονωμένα κύτταρα, χωρίς αυτή η διαφορά να μπορεί να ερμηνευθεί, με βάση την έκφραση του ίδιου γονιδίου σε άλλα αντίστοιχα κύτταρα του ίδιου κυτταρικού τύπου. Πιο συγκεκριμένα, παρουσιάζονται στα δεδομένα μηδενικές ή σχεδόν μηδενικές τιμές κατά την μέτρηση των γονιδίων που εκφράζονται σε ένα κύτταρο, που δεν μπορούν να ερμηνευτούν με βάση τα διαθέσιμα δείγματα. Έτσι τα διάφορα μοντέλα που προσπαθούν να ερμηνεύσουν τα δεδομένα αυτά παρουσιάζουν μικρότερη απόδοση, εκτός και αν ληφθεί υπόψη η ιδιαιτερότητα αυτή.

Ο αριθμός των μετρήσεων UMI, που αντιπροσωπεύουν τον απόλυτο αριθμό των παρατηρούμενων μεταγραφωμάτων ανά κύτταρο αντί για την χρήση των αναγνώσεων ή reads, που περιλαμβάνουν την συναγόμενη αλληλουχία ζευγών βάσεων (ή πιθανότητες ζεύγους βάσεων) που βρίσκονται σε αυτό, μπορεί να χρησιμοποιηθεί κάποιες φορές για να διορθώσει προσωρινά και σε ορισμένες περιπτώσεις το παραπάνω πρόβλημα. Ας θεωρήσουμε ένα μεμονωμένο κύτταρο i που περιέχει t_i μετάγραφα mRNA. Έστω n_i ο συνολικός αριθμός των UMI για το ίδιο κύτταρο. Όταν το κύτταρο υποβάλλεται σε επεξεργασία με ένα πρωτόκολλο scRNA-Seq, λύεται και στη συνέχεια ένα μέρος των μεταγραφωμάτων του μαγνητίζονται από κάποια σφαιρίδια (beads). Στην συνέχεια γίνεται μια σειρά από πολύπλοκες βιοχημικές αντιδράσεις, που περιλαμβάνουν διάφορα στάδια μέχρι την τελική δημιουργία των μετρήσεων UMI. Σε καθένα από αυτά τα στάδια, κάποιο κλάσμα των μορίων από το προηγούμενο στάδιο χάνεται. Συγκεκριμένα, η αντίστροφη μεταγραφάση που χρησιμοποιείται για την εξαγωγή DNA από RNA, είναι ένα αναποτελεσματικό και επιρρεπές σε σφάλματα ένζυμο. Επομένως, ο αριθμός των μετρήσεων UMI που αντιπροσωπεύουν το κύτταρο είναι πολύ μικρότερος από τον αριθμό των μεταγραφωμάτων του αρχικού ($n_i \ll t_i$). Επιπλέον, τα μόρια που επιλέγονται και γίνονται με επιτυχία UMI είναι μια τυχαία διαδικασία. Ισχύει ότι γονίδια με μεγάλη σχετική αφθονία μεταγραφωμάτων στο αρχικό κύτταρο είναι πιο πιθανό να έχουν μη μηδενικούς αριθμούς UMI, αλλά γονίδια με μικρή σχετική αφθονία μπορεί να παρατηρηθούν με μετρήσεις UMI ακριβών μηδενικών (exact zeros). Τελικά οι μετρήσεις UMI που έχουν υπολογιστεί μέσω της διαδικασίας που περιεγράφηκε νωρίτερα, είναι ένα πολυωνυμικό δείγμα των πραγματικών βιολογικών μετρήσεων, που περιέχουν πληροφορίες που σχετίζονται μόνο με τα πρότυπα έκφρασης στο κύτταρο, το οποίο είναι θεμιτό, καθώς αυτή είναι η βασική πληροφορία που χρειαζόμαστε για την ανάλυση αυτού του είδους των δεδομένων. Τέτοιου είδους μετρήσεις χρησιμοποιούνται και σε μεθόδους μείωσης διαστάσεων όπως η GLMPCA που περιγράφεται παρακάτω, σε επόμενο κεφάλαιο (Townes *et al.*, 2020).

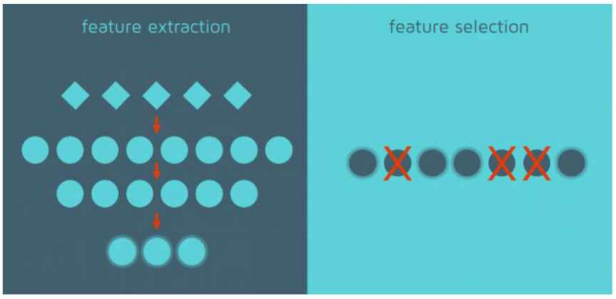
Ένα ακόμα πρόβλημα που έχουν τα δεδομένα αυτά λόγω του όγκου τους είναι το πρόβλημα «μεγάλο p μικρό n» που αναφέρθηκε και νωρίτερα. Ένα κοινό παράδειγμα ενός προβλήματος p >> n στα βιολογικά δεδομένα από μικροσυστοιχίες γονιδίων η και στα δεδομένα από scRNA-Seq, είναι ότι συνήθως υπάρχουν χιλιάδες γονίδια (p) και μόνο δεκάδες δείγματα (n). Το πρόβλημα σε αυτήν την περίπτωση δημιουργείται όταν αυτές χρησιμοποιούνται για διάγνωση/ταξινόμηση ή πρόβλεψη, αφού είναι αδύνατη η εύρεση των απαραίτητων σχέσεων από τα δεδομένα και κατά συνέπεια η χρήση τους με αποτελεσματικό τρόπο (van de Geer and van Houwelingen, 2004). Τέτοιου είδους προβλήματα συνήθως οδηγούν σε υπερπροσαρμογή του συνόλου δεδομένων εκπαίδευσης όταν χρησιμοποιούνται μοντέλα μηχανικής μάθησης (Hastie and Tibshirani, 2003).

Το παραπάνω πρόβλημα μπορεί ενδεχομένως να λυθεί με την χρήση μεθόδων «ποινών» (penalty values) που οδηγούν σε αραιές αναπαραστάσεις. Όμως, παρά το γεγονός ότι υπάρχει ανάγκη για εφαρμογή κατάλληλης cross-validation μεθόδου, για τον σκοπό της πρόβλεψης, θα πρέπει να αποφεύγεται οποιαδήποτε (βιολογική) ερμηνεία του συνόλου των επεξηγηματικών μεταβλητών που επιλέγονται με αυτόν τον τρόπο και των συντελεστών παλινδρόμησής τους (van de Geer and van Houwelingen, 2004). Άρα και στην συγκεκριμένη περίπτωση η χρήση μεθόδων μείωσης διαστάσεων αυτού του είδους των δεδομένων με σκοπό την εξαγωγή πληροφορίας με βιολογική σημασία κρίνεται απαραίτητη.

Ακόμη, όσον αφορά την εφαρμογή κλασικών μεθόδων σχολιασμού των διαφόρων τύπων κυττάρων, στα οποία είναι απαραίτητη η αποτελεσματική ομαδοποίηση των βιολογικών δεδομένων, η αναγνώριση των διαφορετικών τύπων κυττάρων εξαρτάται από τη μέθοδο που χρησιμοποιείται και την επιλογή των βέλτιστων παραμέτρων, που συνήθως βρίσκονται πειραματικά (Yin et al., Jan 2022). Επίσης δεν υπάρχει μια κοινή ορολογία για τα διάφορα είδη κυττάρων, με αποτέλεσμα κάθε κατηγορία να χαρακτηρίζεται διαφορετικά με βάση το εργαστήριο στο οποίο διεξήχθη το πείραμα, γεγονός που οδηγεί σε πρακτικές δυσκολίες και περεταίρω προβλήματα στην αξιοποίηση των συμπερασμάτων κάθε έρευνας. Για τον σκοπό αυτό θεωρείται αναγκαία η κατασκευή ενός «ευρετηρίου» όπου κάθε τύπος κυττάρου θα έχει συγκεκριμένη ονομασία και κατηγορία στην οποία ανήκει. Τα προβλήματα αυτά ίσως μπορέσουν να λυθούν χρησιμοποιώντας μεθόδους κατηγοριοποίησης βασισμένους στην τεχνητή νοημοσύνη και μεθόδους εξαγωγής χαρακτηριστικών.

3. Αλγόριθμοι μείωσης διαστάσεων και εξαγωγής χαρακτηριστικών δεδομένων μεγάλου όγκου

Η μείωση της διάστασης των δεδομένων έχει πολλές πρακτικές εφαρμογές και χρησιμοποιείται ιδιαίτερα τα τελευταία χρόνια λόγω της εκθετικής αύξησης των δεδομένων. Κάποια από τα πλεονεκτήματά της περιλαμβάνουν την μείωση του απαραίτητου χρόνου εκπαίδευσης και την αύξηση της απόδοσης των μεθόδων μηχανικής μάθησης, την ανάγκη λιγότερης υπολογιστικής ισχύος, την επίλυση του προβλήματος της υπερ-προσαρμογής, που αναφέρθηκε νωρίτερα, την ευκολότερη και περισσότερο κατανοητή οπτικοποίηση των δεδομένων κ.α.. Υπάρχει πληθώρα τρόπων που χρησιμοποιούνται για την μείωση της διάστασης των δεδομένων μεγάλου όγκου και οι πιο γνωστές από αυτές αναλύονται παρακάτω. Υπάρχουν δύο είδη τεχνικών που χρησιμοποιούνται για τη μείωση της διάστασης των πολυδιάστατων δεδομένων, κάθε μια όμως έχει διαφορετική προσέγγιση. Ο πρώτος τρόπος διατηρεί μόνο τα πιο σημαντικά χαρακτηριστικά στο σύνολο δεδομένων και αφαιρεί τα περιττά χαρακτηριστικά, χωρίς να εφαρμόσει κάποιο είδος μετασχηματισμού στο αρχικό σύνολο των χαρακτηριστικών (feature selection). Χαρακτηριστικά παραδείγματα αυτών των μεθόδων είναι η Backward elimination, Forward selection και τα Random forests, με τα οποία όμως δεν θα ασχοληθούμε ιδιαίτερα. Ο δεύτερος τρόπος λειτουργεί βρίσκοντας έναν συνδυασμό νέων χαρακτηριστικών, εφαρμόζοντας κατάλληλους μετασχηματισμούς στα αρχικά δεδομένα (feature extraction). Δηλαδή το νέο σύνολο χαρακτηριστικών περιέχει διαφορετικές τιμές σε σχέση με τις αρχικές. Αυτή η μέθοδος μπορεί να διαχωριστεί περεταίρω σε Γραμμικές μεθόδους και σε Μη γραμμικές μεθόδους. Η ανάλυση πρωτευουσών συνιστωσών (PCA), η ανάλυση παραγόντων (FA) και η γραμμική διακριτική ανάλυση (LDA) είναι παραδείγματα γραμμικών μεθόδων μείωσης διαστάσεων. Η PCA πυρήνα, η t-SNE, η πολυδιάστατη κλιμάκωση (MDS) και η ισομετρική χαρτογράφηση (Isomap) είναι παραδείγματα μη γραμμικών μεθόδων μείωσης διαστάσεων. Οι πιο βασικές και συχνά χρησιμοποιούμενες μέθοδοι γραμμικών και μη γραμμικών μεθόδων μείωσης διαστάσεων αναλύονται παρακάτω.

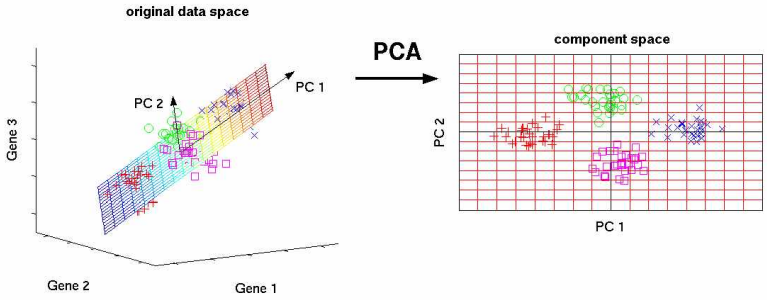


Εικόνα 11 Σύγκριση μεθόδων μείωσης διαστάσεων των μεγάλων δεδομένων

3.1. Γραμμικές Μέθοδοι μείωσης διαστάσεων και feature extraction

3.1.1. Principal Component Analysis (PCA)

Η PCA είναι ένας αλγόριθμος μηχανικής μάθησης (un-supervised) που χρησιμοποιείται για μείωση διαστάσεων. Περιληπτικά ο αλγόριθμος αυτός μετατρέπει ένα σύνολο συσχετιζόμενων μεταβλητών (p) σε ένα σύνολο με μικρότερο αριθμό μη συσχετιζόμενων μεταβλητών k (k<p) που ονομάζονται κύριες συνιστώσες, ενώ διατηρεί όσο το δυνατόν μεγαλύτερο μέρος της διακύμανσης (variance) του αρχικού συνόλου δεδομένων. Πιο συγκεκριμένα η PCA προσπαθεί να υπολογίσει τους άξονες στους οποίους παρατηρείται η μέγιστη διασπορά των δεδομένων. Δηλαδή βρίσκεται ο άξονας που παρέχει τις περισσότερες πληροφορίες για τα δεδομένα, που είναι αυτός με τις μεγαλύτερες διασπορές ανάμεσά τους. Σκοπός είναι να διατηρηθεί όσο δυνατόν μεγαλύτερο μέρος της διακύμανσης και πληροφορίας των αρχικών δεδομένων και χρησιμοποιείται όταν χρειάζεται να υπάρχει η γνώση για τα δείγματα που διαφέρουν περισσότερο (variation).



Εικόνα 12 Εφαρμογή PCA σε δεδομένα τριών διαστάσεων

Για να κατανοήσουμε τον τρόπο με τον οποίο λειτουργεί η PCA χρειάζεται να κάνουμε μια σύντομη αναφορά στους όρους συνδιακύμανση (covariance) και πίνακας συνδιακύμανσης (covariance matrix) (Statistical and Mathematical Concepts behind PCA | by Rukshan Pramoditha | Data Science 365 |

Medium, 2020). Η συνδιακύμανση είναι η τιμή που αντικατοπτρίζεται από το πόσο δύο τυχαίες μεταβλητές συσχετίζονται μεταξύ τους, δηλαδή είναι δείκτης της συμπεριφοράς μιας μεταβλητής σε σχέση με την συμπεριφορά μια άλλης. Μπορεί να υπάρξει θετική, αρνητική ή μηδενική συνδιακύμανση ανάλογα με την συσχέτιση των δύο μεταβλητών. Ο πίνακας συνδιακύμανσης είναι ένας πίνακας, συμμετρικός και θετικά ημιορισμένος, με μη αρνητικές ιδιοτιμές, που περιλαμβάνει τις πιθανές τιμές συνδιακύμανσης που μπορούν να υπολογιστούν για όλα τα χαρακτηριστικά ενός συνόλου δεδομένων.

Η PCA μπορεί να περιγραφεί προσεγγιστικά από την ακόλουθη εξίσωση:

$$Ax = \lambda x$$

Όπου A ένας πίνακας n×n, λ η ιδιοτιμή του A και χ το ιδιοδιάνυσμα που σχετίζεται με την ιδιοτιμή λ και για το οποίο ισχύει η παραπάνω σχέση. Αν θέλουμε να περιγράψουμε επακριβώς τον αλγόριθμο μπορούμε να χρησιμοποιήσουμε την ακόλουθη σχέση :

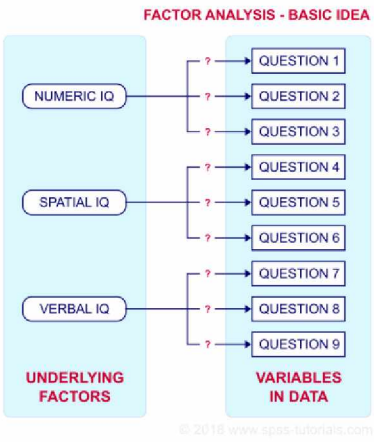
$$A = U\Sigma U^{-1}$$

Στην σχέση αυτή, ο A είναι ένας πίνακας διάστασης n×n, όπου για την PCA είναι ο πίνακας συνδιακύμανσης που περιγράψαμε νωρίτερα. Το Σ αντιπροσωπεύει όλες τις ιδιοτιμές σε μορφή ενός διαγώνιου πίνακα. Αυτές δείχνουν την ποσότητα της μεταβλητότητας μέσα στο σύνολο των δεδομένων που υποδεικνύεται από την συγκεκριμένη ιδιοτιμή. Δηλαδή κάθε ιδιοτιμή υποδηλώνει πόσο χρήσιμος ή πόσες πληροφορίες δίνει το κάθε ιδιοδιάνυσμα στο αρχικό σύνολο δεδομένων. Όσο μεγαλύτερη η ιδιοτιμή, τόσο μεγαλύτερη η προσφορά του αντίστοιχου ιδιοδιανύσματος σε χρήσιμη πληροφορία. Τέλος το U αντιπροσωπεύει όλα τα ιδιοδιανύσματα σε μορφή ενός τετραγωνικού πίνακα n×n.

Από τις ιδιοτιμές που υπολογίστηκαν παραπάνω μόνο ένα ποσοστό, αυτών με την μεγαλύτερη τιμή και κατά συνέπεια το μεγαλύτερο ποσοστό πληροφορίας και μεταβλητότητας, είναι απαραίτητο και χρησιμοποιείται. Η διαδικασία μείωσης διαστάσεων είναι πρακτικά ο πολλαπλασιασμός ανάμεσα στον πίνακα των επιλεγμένων ιδιοτιμών και των δεδομένων που πρόκειται να μετασχηματιστούν. Έτσι τα δεδομένα υφίστανται ένα μετασχηματισμό που διατηρεί όσο το δυνατό μεγαλύτερο μέρος της πληροφορίας, μειώνοντας τις διαστάσεις του αρχικού συνόλου δεδομένων.

3.1.2. Factor analysis (FA)

Ο κύριος στόχος της Ανάλυσης Παραγόντων είναι η εύρεση «κρυμμένων» μεταβλητών (latent) που δεν υπολογίζονται άμεσα σε μια μεμονωμένη μεταβλητή αλλά προκύπτουν από άλλες μεταβλητές του συνόλου δεδομένων. Αυτές οι μεταβλητές ονομάζονται παράγοντες και δίνουν ουσιαστικές πληροφορίες για τα δεδομένα καθώς μπορούν να ερμηνευτούν έχοντας κάποια φυσική σημασία. Αν για παράδειγμα βρεθούν δύο «κρυμμένες» μεταβλητές τότε το μοντέλο ονομάζεται μοντέλο ανάλυσης δύο παραγόντων. Συνήθως σε κάθε παράγοντα ανατίθενται τουλάχιστον τρεις μεταβλητές. Η βασική υπόθεση της FA είναι ότι υπάρχουν τέτοιου είδους μεταβλητές στα διαθέσιμα δεδομένα. Για το σκοπό αυτό χρειάζεται υπολογιστεί ο πίνακας συνδιακύμανσης, οι ιδιοτιμές και να βρεθεί ο αριθμός των παραγόντων (ιδιοτιμών) που εξηγούν όσο το μεγαλύτερο δυνατό ποσοστό της διακύμανσης των δεδομένων(Complete Guide to Factor Analysis (Updated 2022) - Qualtrics, 2022). Στην συνέχεια, ελέγχεται ποιες είναι οι μεταβλητές από τα αρχικά δεδομένα που χρησιμοποιούνται για να καθορίσουν τον κάθε παράγοντα και δίνεται μια λογική ερμηνεία στον καθένα από αυτούς. Έτσι μπορούν πλέον να χρησιμοποιηθούν οι τιμές των νέων παραγόντων για περεταίρω ανάλυση των αρχικών δεδομένων, αφού πλέον έχει διατηρηθεί ένα μεγάλο μέρος της πληροφορίας αποθηκευμένο σε αυτούς και η διάσταση των αρχικών δεδομένων μειώθηκε.

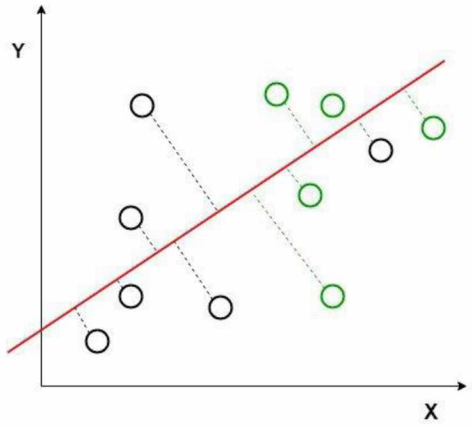


Εικόνα 13 Παράδειγμα ανάλυσης παραγόντων

3.1.3. Linear Discriminant Analysis (LDA)

Η LDA ή γραμμική ανάλυση διακριτότητας είναι μια μέθοδος επιβλεπόμενης μάθησης, που χρησιμοποιείται συνήθως για ταξινόμηση πολλαπλών κατηγοριών, προσπαθώντας να μεγιστοποιήσει την απόσταση ανάμεσα στις διαφορετικές κλάσεις ή ως τεχνική μείωσης διαστάσεων. (Using Linear Discriminant Analysis (LDA) for data Explore: Step by Step. | Blog, Jul 2017). Χρησιμοποιεί την

γνώση από τις ετικέτες των δεδομένων με σκοπό να διακρίνει τα δεδομένα εκπαίδευσης. Η τεχνική LDA βρίσκει έναν γραμμικό συνδυασμό χαρακτηριστικών εισόδου που βελτιστοποιεί τη διαχωριστικότητα (separability) των κλάσεων, σε αντίθεση με την PCA που επιχειρεί να βρει ένα σύνολο ασύνδετων στοιχείων μέγιστης διακύμανσης στο σύνολο δεδομένων. Πιο συγκεκριμένα η LDA προβάλλει όλα τα δεδομένα εισόδου σε μια και μόνο ευθεία. Η ποιότητα της μεθόδου, εξαρτάται από την κλίση της ευθείας και κατά συνέπεια από την ποιότητα του διαχωρισμού των δεδομένων που ανήκουν στην ίδια ομάδα και την απόσταση των σημείων που ανήκουν στην ίδια ομάδα. Αν τα δεδομένα στην ευθεία δεν διαχωρίζονται ξεκάθαρα, τότε ο αλγόριθμος αναζητά άλλη ευθεία, ώστε να εξασφαλιστεί ο βέλτιστος διαχωρισμός ανάμεσα στα δεδομένα.



Εικόνα 14 Εύρεση βέλτιστης ευθείας για προβολή των σημείων, λαμβάνοντας υπόψη και τις δύο διαστάσεις, ώστε να επιτευχθεί η μέγιστη διαχωρισιμότητα των κλάσεων και η ελάχιστη απόσταση των σημείων που ανήκουν στην ίδια.

Έστω ένα σύνολο δεδομένων X = [x1, x2,..., xN], xi∈Rd και έστω N1 τα δεδομένα που ανήκουν στο υποσύνολο D1 και έχουν ετικέτα λ1 και N2 δεδομένα που ανήκουν στο υποσύνολο D2 και έχουν ετικέτα λ2. Σχηματίζοντας έναν γραμμικό συνδυασμό των στοιχείων του (X), προκύπτει το παρακάτω βαθμωτό γινόμενο:

$$y = w^t x$$

Το w καθορίζει την κατεύθυνση της ευθείας και είναι δείκτης του καλού ή όχι διαχωρισμού των δεδομένων πάνω σε αυτήν. Το μέτρο του |w| κλιμακώνει το y. Σκοπός του αλγόριθμου είναι η εύρεση της βέλτιστης κατεύθυνσης του w που υπολογίζεται από το μέσο των δειγμάτων:

$$m_i = \frac{1}{N_i} \sum_{x \epsilon D_i} x$$

Η μέση τιμή όλων των δειγμάτων είναι η προβολή του m_i:

$$\tilde{m}_i = \frac{1}{N_i} \sum_{y \epsilon Y_i} y = \frac{1}{N_i} \sum_{x \epsilon D_i} w^t x = w^t m_i$$

Έτσι υπολογίζεται η διαφορά ανάμεσα στις προβαλλόμενες μέσες τιμές ως:

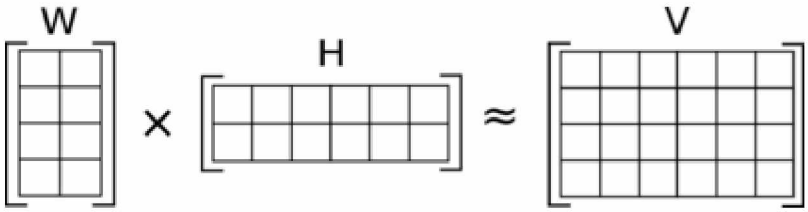
$$|\tilde{m}_1 - \tilde{m}_2| = |w^t(m_1 - m_2)|$$

Αν λοιπόν τα δεδομένα δεν είναι ξεκάθαρα διαχωρίσιμα, αρκεί να αυξηθεί η διαφορά των προβαλλόμενων μέσω τιμών και κατά συνέπεια η απόσταση ανάμεσα στις ομάδες. Με αυτόν τον τρόπο επιτυγχάνεται η μείωση των διαστάσεων των δεδομένων αντικαθιστώντας τις παρατηρήσεις με γραμμικούς συνδυασμούς που συγκεντρώνουν την μεγαλύτερη ποσότητα πληροφορίας των δεδομένων αυτών.

3.1.4. Non-negative Matrix Factorization (NMF)

Η παραγοντοποίηση σε μη αρνητικές μήτρες (NMF) είναι μια τεχνική μη επιβλεπόμενης μάθησης μείωσης διαστάσεων καθώς εξάγει αυτόματα αραιά και ουσιαστικά χαρακτηριστικά από ένα σύνολο μη αρνητικών διανυσμάτων δεδομένων, πολλών διαστάσεων, τα οποία αποσυνθέτει (ή παραγοντοποιεί), ώστε τελικά να προκύψουν αναπαραστάσεις χαμηλότερης διάστασης. Αυτά τα διανύσματα χαμηλότερης διάστασης είναι μη αρνητικά, και κατά συνέπεια οι συντελεστές τους είναι μη αρνητικοί. Η συγκεκριμένη μέθοδος χρησιμοποιείται όταν σε πολυδιάστατα δεδομένα υπάρχουν χαρακτηριστικά που είναι διφορούμενα ή έχουν ασθενή προβλεψιμότητα. Συνδυάζοντας χαρακτηριστικά, η NMF μπορεί να παράγει μοτίβα με νόημα. Κάθε νέο διάνυσμα που δημιουργείται είναι ένας γραμμικός συνδυασμός του αρχικού συνόλου χαρακτηριστικών και έχει ένα σύνολο συντελεστών, που αποτελούν μέτρο του βάρους κάθε επιμέρους χαρακτηριστικού του. Η NMF χρησιμοποιεί τεχνικές της πολυμεταβλητής ανάλυσης και της γραμμικής άλγεβρας και χρησιμοποιώντας έναν αρχικό πίνακα (A), η NMF δημιουργεί δύο πίνακες (W και H). Ο πίνακας W

περιλαμβάνει την βάση της NMF και ο Η τους συντελεστές (βάρη) για αυτή. Ο πολλαπλασιασμός αυτών των δύο πινάκων οδηγεί προσεγγιστικά στην εύρεση του αρχικού πίνακα δεδομένων.



Εικόνα 15 Απεικόνιση παραγοντοποίησης μη αρνητικού πίνακα (NMF).

Ο αλγόριθμος τροποποιεί επαναληπτικά τις τιμές των W και Η έτσι ώστε το γινόμενό τους να πλησιάζει το Α. Η τεχνική διατηρεί μεγάλο μέρος της πληροφορίας των αρχικών δεδομένων και εγγυάται ότι η βάση και τα βάρη δεν είναι αρνητικά. Αυτή είναι η σημαντικότερη ιδιότητα της μεθόδου, καθώς τα χαρακτηριστικά στον πίνακα W (της βάσης) επιτρέπουν μόνο τη χρήση πρόσθεσης και όχι της αφαίρεσης. Έτσι επιτυγχάνεται η ιδιότητα της «μερικής μάθησης» ή part learning της NMF (Zhenhua Li, Xiang Li et al., 2012). Ο αλγόριθμος τερματίζει όταν το σφάλμα προσέγγισης συγκλίνει ή επιτυγχάνεται ένας καθορισμένος αριθμός επαναλήψεων.

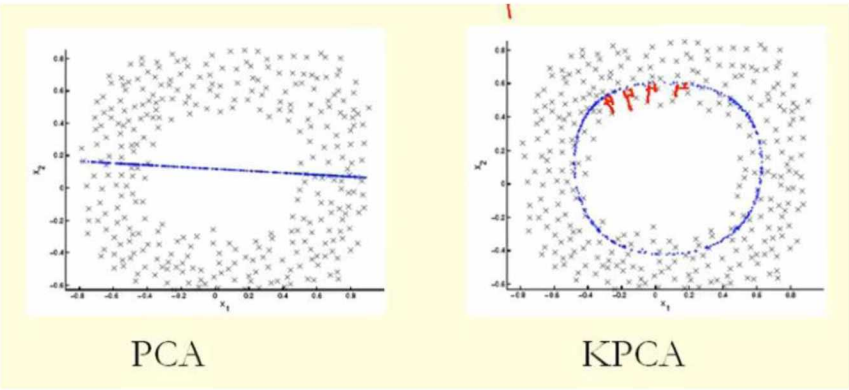
Ο αλγόριθμος NMF πρέπει να αρχικοποιηθεί πρίν αρχίσει να τροποποιεί τις τιμές των W και Η και η κατάλληλη αρχικοποίηση μπορεί να είναι κρίσιμη για την απόκτηση ουσιαστικών αποτελεσμάτων. Συνήθως χρησιμοποιείται μια τυχαία αρχικοποίηση των τιμών των δύο πινάκων με βάση μια ομοιόμορφη κατανομή. Αυτή η προσέγγιση λειτουργεί καλά στις περισσότερες περιπτώσεις (Topic Modeling Articles with NMF. Extracting topics is a good… | by Rob Salgado | Towards Data Science, Apr 2020), (17 Non-Negative Matrix Factorization, Oracle, no date)17 .

3.2. Μη Γραμμικές Μέθοδοι μείωσης διαστάσεων και feature extraction

3.2.1. Kernel PCA

Η Kernel PCA ή PCA πυρήνα είναι μια μη γραμμική τεχνική μείωσης διαστάσεων που χρησιμοποιεί πυρήνες. Μπορεί επίσης να θεωρηθεί ως η μη γραμμική μορφή της απλής PCA και χρησιμοποιείται σε μη γραμμικά σύνολα δεδομένων όπου η απλή PCA δεν μπορεί να χρησιμοποιηθεί αποτελεσματικά. Στη συγκεκριμένη μέθοδο τα δεδομένα εισόδου μετασχηματίζονται αρχικά από μια συνάρτηση πυρήνα και προβάλλονται προσωρινά σε έναν νέο χώρο χαρακτηριστικών υψηλότερης διάστασης, στον οποίο όλες οι κλάσεις γίνονται γραμμικά διαχωρίσιμες. Ο νέος αυτός χώρος είναι ένα πίνακας του οποίου το

κάθε στοιχείο έχει μετασχηματιστεί από μια συνάρτηση $k_{i,j} = k(x_i, x_j)$. Η συνάρτηση αυτή είναι μια συνάρτηση Kernel (Shawe-Taylor and Cristianini, 2004), η οποία μπορεί να είναι οποιαδήποτε συνάρτηση δημιουργεί ένα θετικά ημιορισμένο Kernel πίνακα K. Άρα τα δεδομένα υφίστανται μετασχηματισμούς σε χώρο μεγαλύτερης διάστασης. Στη συνέχεια ο πίνακας αυτός κεντροποιείται αφαιρώντας τον μέσο όρο από όλα τα στοιχεία του πίνακα K. Έτσι τα δεδομένα του πίνακα αυτού έχουν μηδενικό μέσο. Στο σημείο αυτό είναι πλέον εύκολο να εφαρμοστεί ο κλασικός αλγόριθμος της PCA, υπολογίζοντας τα ιδιοδιανύσματα και τις ιδιοτιμές του πίνακα K, όπως έχει ήδη αναλυθεί νωρίτερα. Τα δεδομένα, λοιπόν, σε δεύτερο επίπεδο μετασχηματίζονται τελικά σε χώρο μικρότερης διάστασης, όπου μπορούν να χρησιμοποιηθούν για περεταίρω αναλύσεις.



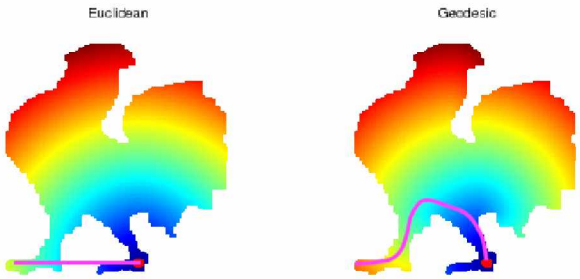
Εικόνα 16 Εφαρμογή PCA πυρήνα για μη γραμμικά διαχωρίσιμα δεδομένα

3.2.2. Multidimensional Scaling (MDS)

Η μέθοδος MDS ή πολυδιάστατης κλιμάκωσης (Torgerson, 1958), έχει σκοπό τη διατήρηση των αποστάσεων μεταξύ των σημείων στα αρχικά δεδομένα, ενώ μειώνει τη διάστασή τους δημιουργώντας έναν χώρο μειωμένης διάστασης στον οποίο χαρτογραφούνται οι αποστάσεις αυτές. Μπορεί να περιγραφεί και ως μια οπτική αναπαράσταση αποστάσεων ή ανομοιοτήτων μεταξύ συνόλων αντικειμένων. Ο αλγόριθμος MDS τοποθετεί κάθε αντικείμενο σε ένα χώρο N διαστάσεων, έτσι ώστε οι αποστάσεις μεταξύ των αντικειμένων να διατηρούνται όσο το δυνατόν καλύτερα. Τα σημεία που μοιάζουν περισσότερο (ή έχουν μικρότερες αποστάσεις) είναι πιο κοντά μεταξύ τους στο γράφημα σε σχέση με τα σημεία που είναι λιγότερο όμοια (ή έχουν μεγαλύτερες αποστάσεις). Υπάρχουν δύο τύποι αλγορίθμων MDS, ο μετρικός και ο μη μετρικός. Ο σκοπός της πολυδιάστατης κλιμάκωσης είναι να χαρτογραφήσει τη σχετική θέση των αντικειμένων χρησιμοποιώντας δεδομένα που δείχνουν πώς διαφέρουν τα σημεία. Η MDS είναι μια καλή τεχνική για την διατήρηση τόσο των καθολικών (global) όσο και των τοπικών (local) δομών των δεδομένων πολλών διαστάσεων.

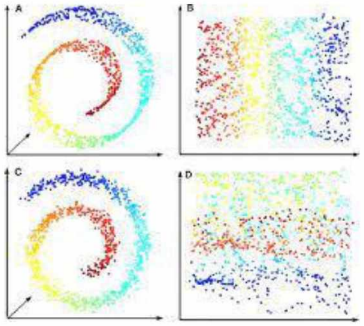
3.2.3. Isomap

Η μέθοδος Isometric mapping ή Isomap εκτελεί μη γραμμική μείωση διαστάσεων μέσω ισομετρικής χαρτογράφησης που διατηρεί τις εσωτερικές αποστάσεις ανάμεσα στα κοντινά δεδομένα του χώρου χρησιμοποιώντας όμως την έννοια της γεωδεσιακής απόστασης των δεδομένων. Η γεωδαισιακή αποτελεί την ελάχιστου μήκους απόσταση ανάμεσα σε δύο σημεία, λαμβάνοντας όμως υπόψη το σχήμα του χώρου, σε αντίθεση για παράδειγμα με την ευκλείδεια που υπολογίζεται μέσω του μήκους μιας ευθείας γραμμής που συνδέει τα δύο σημεία.



Εικόνα 17 Γεωδαισιακή απόσταση ενός ζεύγους σημείων

Η μέθοδος συνδέει κάθε σημείο x_i (i = 1, 2, ..., n), υπολογίζοντας την καμπύλη ή γεωδαισιακή απόσταση, με τους k πλησιέστερους γείτονες x_{ij} (j = 1, 2, ..., k) κατασκευάζοντας ένα γράφημα γετνίασης G και μειώνει τη διάσταση υπολογίζοντας την συντομότερη διαδρομή μέσω άλλων αλγόριθμων (πχ. Dijkstra). Οι γεωδαισιακές αποστάσεις μεταξύ όλων των σημείων δεδομένων στο X υπολογίζονται, σχηματίζοντας έναν πίνακα με ζεύγη γεωδαισιακών αποστάσεων. Οι μειωμένων διαστάσεων αναπαραστάσεις y_i των σημείων δεδομένων xi στο χώρο των λίγων διαστάσεων Y υπολογίζονται κλιμακώνοντας τον πίνακα με τα ζεύγη γεωδαισιακών αποστάσεων. Σκοπός της Isomap είναι η διατήρηση των γεωδαισιακών αποστάσεων των δεδομένων στον χώρο μικρότερης διάστασης και επιτυγχάνει καλύτερο διαχωρισμό ανάμεσα στις κλάσεις σε σχέση με άλλες μεθόδους μείωσης διαστάσεων.



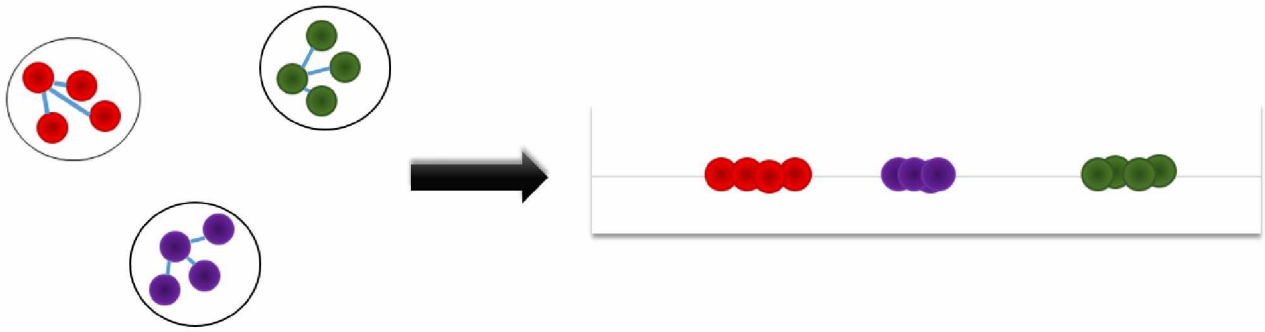
Εικόνα 18 Απεικόνιση της Isomap σε ένα ελβετικό σύνολο δεδομένων ρολού ή swiss roll dataset

3.2.4. t-distributed Stochastic Neighbor Embedding (t-SNE)

Η μέθοδος t-Κατανεμημένη Στοχαστική Ενσωμάτωση γειτόνων ή t-SNE χρησιμοποιείται κυρίως για οπτικοποίηση πολυδιάστατων δεδομένων προσδίδοντας σε κάθε σημείο μια θέση σε ένα χάρτη δύο ή τριών διαστάσεων. Αυτό είναι ιδιαίτερα σημαντικό για τα πολυδιάστατα δεδομένα που βρίσκονται σε πολλές διαφορετικές, αλλά σχετικά λίγων διαστάσεων θέσεις στον χώρο, όπως εικόνες αντικειμένων που ανήκουν σε πολλές κλάσεις και φαίνονται από διαφορετικές οπτικές γωνίες.

Η μέθοδος t-SNE προσπαθεί να διατηρήσει την εσωτερική δομή των δεδομένων, προκειμένου να μην χαθούν τυχόν χωρικές πληροφορίες (*t-SNE: Behind the Math. Being one of the most talked about… | by Sushanth Sreenivasa | Towards Data Science*, no date; *t-SNE clearly explained. An intuitive explanation of t-SNE… | by Kemal Erdem (burnpiro) | Towards Data Science*, no date). Πιο συγκεκριμένα, προσπαθεί να διατηρήσει μαζί τα κοντινά σημεία του χώρου πολλών διαστάσεων στο χώρο μικρότερης διάστασης. Παρόλα αυτά η απόσταση μεταξύ των μη γειτονικών σημείων δεν διατηρείται μετά τους μετασχηματισμούς. Η t-SNE υπολογίζει την ομοιότητα των σημείων, με βάση τις αποστάσεις μεταξύ τους και τα ομαδοποιεί με βάση την ομοιότητά τους, διατηρώντας την εσωτερική δομή τους και σχηματίζοντας ομάδες σε χώρο λιγότερων διαστάσεων. Αυτό γίνεται μέσω της μέτρησης των αποστάσεων όλων των σημείων του αρχικού χώρου και κατασκευής ενός πίνακα με τις τιμές ομοιότητας των σημείων. Κάθε σημείο, για να μπορέσει ο αλγόριθμος να λειτουργήσει σωστά, ορίζεται ότι έχει μηδενική ομοιότητα με τον εαυτό του. Κάθε φορά, αφού τοποθετήσει τυχαία τα σημεία στον νέο χώρο μικρότερης διάστασης, ο αλγόριθμος υπολογίζει τις αποστάσεις αυτού από όλα τα υπόλοιπα του χώρου και κινεί σταδιακά κάθε ένα σημείο, ώστε ο πίνακας αποστάσεων ή ομοιότητας του χώρου μικρότερης διάστασης να μοιάζει όσο το δυνατό περισσότερο στον αρχικό πίνακα ομοιοτήτων. Με αυτόν τον τρόπο επιτυγχάνεται ο διαχωρισμός των σημείων που δεν ανήκουν

στην ίδια κλάση και η διατήρηση των παρόμοιων σημείων του αρχικού χώρου κοντά και στο χώρο μειωμένης διάστασης.



Εικόνα 19 Εφαρμογή της t-SNE στις δύο διαστάσεις

Η απόσταση μεταξύ όλων των σημείων μετριέται και απεικονίζεται κατά μήκος της κατανομής T Student επειδή έχει «βαρύτερες ουρές». Η κατανομή T είναι σε θέση να διαφοροποιήσει περισσότερο τα σημεία που βρίσκονται πιο μακριά σε σχέση με τα σημεία που ανήκουν στην ίδια ομάδα λόγω των «ουρών» της. Έτσι εξασφαλίζεται η διαφοροποίηση των τιμών που είναι πιο μακριά από τη μέση τιμή και κατά συνέπεια διαφέρουν περισσότερο. Όσο περισσότερο η κατανομή αυτή πλησιάζει την κανονική τόσο πιο κοντινά και άρα όμοια είναι τα σημεία του δείγματος.

3.2.5. Diffusion Map

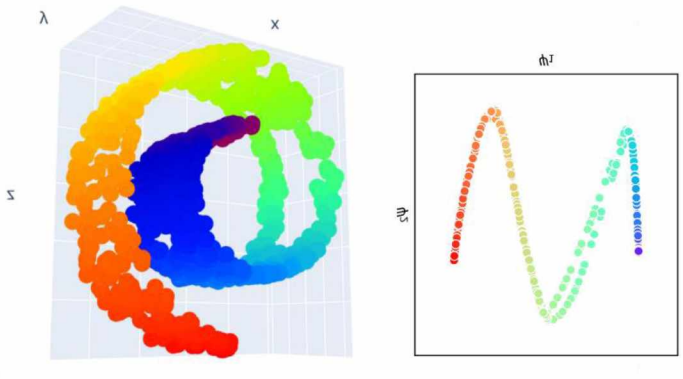
Η τεχνική του diffusion map ή χάρτη διάχυσης (Porte and Herbst, 2008), επικεντρώνεται στην ανακάλυψη του manifold ή τοπολογικού χώρου, από τον οποίο έχουν προέλθει τα δεδομένα. Πιο συγκεκριμένα, η συνδεσιμότητα (connectivity) του συνόλου δεδομένων, που υπολογίζεται με τη χρήση ενός μέτρου τοπικής ομοιότητας, συνήθως ευκλείδειας απόστασης, χρησιμοποιείται για τη δημιουργία ενός γράφου.

Ο χρόνος ή ο αριθμός των «jumps» από τον έναν κόμβο του γραφήματος στον άλλο, είναι δείκτης ομοιότητας των σημείων και αποκαλύπτει την γεωμετρία του χώρου. Κάθε τυχαίος «περίπατος» στο γράφο οδηγεί και σε διαφορετική σειρά πρόσβασης στα δεδομένα. Έτσι η πιθανότητα διασύνδεσης ενός σημείου ή κόμβου με έναν άλλο καθορίζεται από μια σχέση της μορφής: connectivity(x, y) = p(x, y). Υπολογίζοντας για κάθε διαφορετικό περίπατο τις πιθανότητες σύνδεσης ενός κόμβου x με έναν άλλο y, κατασκευάζεται ένας πίνακας P ή «diffusion matrix», που περιλαμβάνει τις πιθανότητες πρόσβασης από έναν κόμβο σε έναν άλλο. Ανάλογα με τον αριθμό των «jumps» που θεωρούμε ότι

βρίσκουν κοντινούς κόμβους, υψώνουμε τον πίνακα P στην αντίστοιχη δύναμη. Δηλαδή για δύο άλματα ο πίνακας περιλαμβάνει όλες οι διαδρομές από το σημείο x στο σημείο y. Ομοίως, ο $P^t_{x,y}$ αθροίζει όλες τις διαδρομές μήκους t από το σημείο x στο σημείο y.

Καθώς υπολογίζονται οι πιθανότητες P_t , όσο το t αυξάνεται, παρατηρείται το σύνολο δεδομένων σε διαφορετικές κλίμακες. Αυτή ονομάζεται διαδικασία διάχυσης. Ανακαλύπτονται τόσο νέοι τοπολογικοί χώροι και τοπικά χαρακτηριστικά των δεδομένων, όσο και κλάσεις που δεν είναι συνδεδεμένες μεταξύ τους. Όσο αυξάνεται το t (δηλαδή καθώς η διαδικασία διάχυσης προχωράει), η πιθανότητα επιλογής μιας διαδρομής κατά μήκος της υποκείμενης γεωμετρικής δομής του συνόλου δεδομένων αυξάνεται. Αυτό συμβαίνει επειδή, κατά μήκος αυτής, τα σημεία είναι πυκνά και ως εκ τούτου στενά συνδεδεμένα (η συνδεσιμότητα είναι συνάρτηση της Ευκλείδειας απόστασης μεταξύ δύο σημείων). Τα μονοπάτια σχηματίζονται κατά μήκος σύντομων, μεγάλης πιθανότητας αλμάτων. Μονοπάτια που δεν ακολουθούν αυτή τη δομή περιλαμβάνουν ένα ή περισσότερα μεγάλα, άλματα χαμηλής πιθανότητας, τα οποία μειώνουν τη συνολική πιθανότητα επιλογής της συγκεκριμένης διαδρομής. Είναι ακόμη σημαντικό να αναφερθεί ότι η πιθανότητα επιλογής μιας διαδρομής δεν είναι συμμετρική, δηλαδή p(x, y) ≠ p(y, x), καθώς σε κάθε βήμα παρουσιάζονται νέες επιλογές αλμάτων που μπορεί να επιλεχθούν. Η απόσταση διάχυσης είναι μικρή εάν υπάρχουν πολλά μονοπάτια υψηλής πιθανότητας μήκους t μεταξύ δύο σημείων.

Όσο η διαδικασία διάχυσης προχωράει, αποκαλύπτεται και η γεωμετρική δομή των δεδομένων κυρίως μέσω του υπολογισμού των αποστάσεων διάχυσης. Ο χάρτης διάχυσης διατηρεί την εγγενή γεωμετρία ενός συνόλου δεδομένων και δεδομένου ότι η χαρτογράφηση μετρά τις αποστάσεις σε μια δομή χαμηλότερης διάστασης, αναμένεται να βρεθεί ότι απαιτούνται λιγότερες συντεταγμένες για την αναπαράσταση των σημείων αυτών στο νέο χώρο. Η εύρεση των διαστάσεων που διατηρούν τις γεωγραφικές πληροφορίες των δεδομένων γίνεται με τη χρήση των ιδιοτιμών των ιδιοδιανυσμάτων του πίνακα P. Η μείωση της διάστασης των δεδομένων επιτυγχάνεται παραβλέποντας ορισμένες διαστάσεις στο χώρο διάχυσης και διατηρώντας τα ιδιοδιανύσματα του πίνακα P, με τις m μεγαλύτερες ιδιοτιμές που αντιστοιχούν σε αυτά και κατά συνέπεια τις αντίστοιχες διαστάσεις. Τα ιδιοδιανύσματα με τις μεγαλύτερες ιδιοτιμές εκτείνονται σε έναν χώρο μικρότερης διαστατικότητας όπου τα αρχικά δεδομένα μπορούν να αναπαρασταθούν αποτελεσματικά.



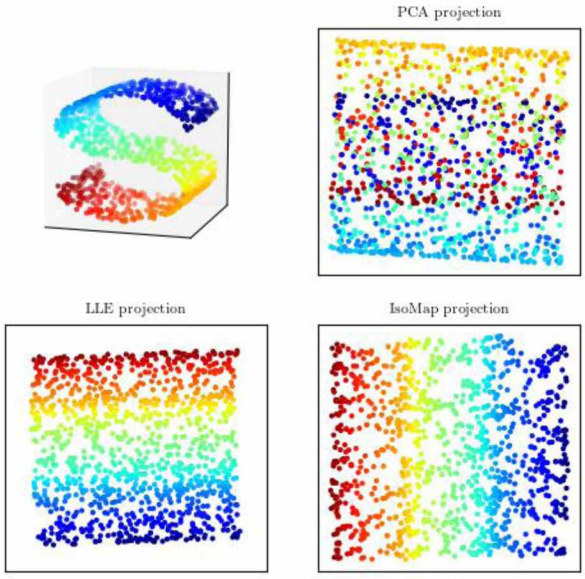
Εικόνα 20 Εφαρμογή του χάρτη διάχυσης για μείωση διάστασης του ελβετικού συνόλου δεδομένων ρολού

3.2.6. locally linear embedding (LLE)

Η LLE (Saul and Roweis, 2000) είναι μια μέθοδος μείωσης διαστάσεων, που σκοπός της είναι να μειώσει την διάσταση των αρχικών δεδομένων ώστε να διατηρηθούν τα γεωμετρικά χαρακτηριστικά τους και ειδικότερα οι αποστάσεις ανάμεσα στους κοντινότερους γείτονες (locally). Εκτελεί σε αυτούς μια σειρά από επαναλαμβανόμενες PCA, τα αποτελέσματα των οποίων συγκρίνονται και συνενώνονται, ώστε να καλυφθεί ολόκληρος ο του αρχικός χώρος (globally), για την εύρεση του καλύτερου μη γραμμικού χώρου χαμηλότερης διάστασης (non-linear embedding). Πιο συγκεκριμένα εκμεταλλεύεται την τοπική γεωμετρία τους και στην συνέχεια ενώνει τα τμήματα που προκύπτουν με στόχο τη διατήρηση της γεωμετρίας ολόκληρου του χώρου, σε έναν χώρο λιγότερων διαστάσεων, σε αντίθεση με παρόμοιες μεθόδους, όπως η PCA που λαμβάνουν εξ' αρχής υπόψη τους ολόκληρη την γεωμετρία του χώρου. Βασική προϋπόθεση για την εφαρμογή αυτής της μεθόδου είναι ο αρχικός χώρος να είναι συνεχής και τα δεδομένα να μην έχουν πολύ θόρυβο.

Αρχικά για κάθε σημείο του χώρου επιλέγονται οι k κοντινότεροι γείτονες και δημιουργούνται γραμμικά βάρη (linear weights), για να μπορέσει να κατασκευαστεί ξανά το συγκεκριμένο σημείο με βάση τους κοντινότερους γείτονες που ανατέθηκαν σε αυτό. Η διαδικασία εύρεσης αντιπροσωπευτικών βαρών ολοκληρώνεται όταν ελαχιστοποιηθεί μια συνάρτηση κόστους. Αυτός είναι και ο λόγος που η συνέχεια στον αρχικό χώρο δεδομένων είναι επιθυμητή, καθώς αν οι γείτονες βρίσκονται πολύ μακριά από το συγκεκριμένο σημείο δεν θα μπορέσουν να υπολογιστούν αντιπροσωπευτικά βάρη. Το ίδιο συμβαίνει και αν στα δεδομένα υπάρχει πολύς θόρυβος. Η επιλογή κατάλληλου αριθμού γειτόνων μπορεί να επηρεάσει εξίσου σημαντικά την απόδοση της μεθόδου, καθώς μια πολύ μεγάλη ή πολύ μικρή τιμή γειτόνων αποτυγχάνει να διατηρήσει τις απαραίτητες γεωγραφικές πληροφορίες των δεδομένων.

Στη συνέχεια, δημιουργείται ένας πίνακας βαρών όλων των σημείων που υπολογίστηκαν νωρίτερα. Με αυτόν τον τρόπο δεν είναι απαραίτητη η διατήρηση όλων των διαστάσεων των σημείων του χώρου αλλά μόνο ο πίνακας Y των βαρών. Τέλος γίνεται η χαρτογράφηση του νέου αυτού χώρου σε έναν χώρο λιγότερων διαστάσεων, μέσω της επιλογής των m μεγαλύτερων ιδιοτιμών που προκύπτουν από τα ιδιοδιανύσματα του πίνακα Y. Σε σύγκριση με άλλες μεθόδους μη επιβλεπόμενης μάθησης όπως η t-SNE, η Isomap και η LLE χρησιμοποιείται κυρίως για να εξάγει υπάρχουσες πληροφορίες από έναν συνεχή, χαμηλών διαστάσεων χώρο. Από την άλλη πλευρά, η t-SNE εστιάζει στην δομή των δεδομένων τοπικά και προσπαθεί να «εξάγει» ομαδοποιημένες τοπικές ομάδες αντί να «ξεδιπλώσει» (unfold) τις πληροφορίες στον υπάρχοντα συνεχή χώρο.



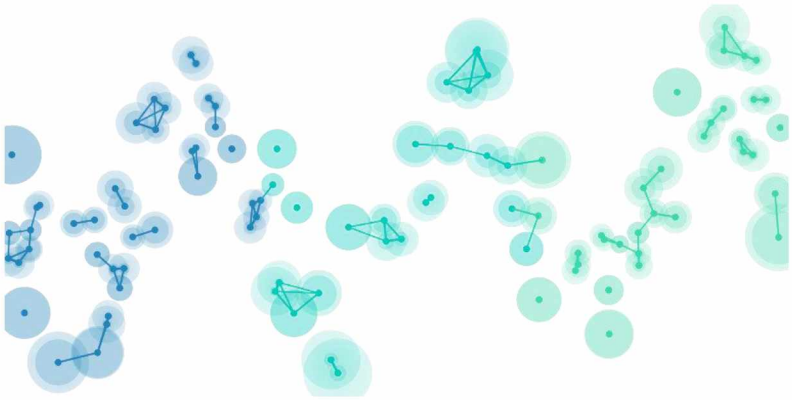
Εικόνα 21 Σύγκριση της LLE με άλλες μεθόδους μείωσης διαστάσεων. Σε σχέση με την PCA επιτυγχάνει καλύτερο διαχωρισμό των κλάσεων και συγκριτικά με την Isomap εξαλείφει την ανάγκη εκτίμησης των αποστάσεων ανά ζεύγη μεταξύ ευρέως διαχωρισμένων σημείων και ανακτά τη συνολική μη γραμμική δομή από τοπικές πληροφορίες των δεδομένων.

3.2.7. Uniform Manifold Approximation and Projection (UMAP)

Η μέθοδος UMAP (McInnes *et al.*, 2018) είναι μέθοδος μη επιβλεπόμενης μάθησης που, σε αντίθεση με την t-SNE, έχει τη δυνατότητα να διατηρήσει εκτός από την τοπική και την καθολική δομή των δεδομένων. Δηλαδή δίνει και την δυνατότητα διατήρησης της πληροφορίας της απόστασης ανάμεσα στις διαφορετικές ομάδες και όχι μόνο τις αποστάσεις των δεδομένων μέσα σε αυτές. Η UMAP

στηρίζεται σε μαθηματικά θεωρήματα και όλα τα βήματα που εφαρμόζει είναι μαθηματικά αποδεδειγμένα.

Η διαδικασία περιλαμβάνει αρχικά την κατασκευή ενός γραφήματος μεγάλης διάστασης και στην συνέχεια την προβολή του σε χώρο μικρότερης διάστασης για την δημιουργία ενός νέου. Σκοπός του αρχικού γραφήματος είναι να βρεθεί προσεγγιστικά το σχήμα ή η τοπολογία των αρχικών δεδομένων. Κάθε ένα από αυτά τα δεδομένα ονομάζονται 0-simplex και σύμφωνα με το θεώρημα «Nerve» η τοπολογία αυτή ή σχήμα των δεδομένων μπορεί να βρεθεί αν συνδεθούν αυτά τα 0-simplexes με τα γειτονικά τους δημιουργώντας 1, 2 ή περισσότερων διαστάσεων simplexes. Για να βρει τα γειτονικά σημεία, η UMAP δημιουργεί έναν κύκλο με συγκεκριμένη ακτίνα (radius) γύρω από κάθε ένα σημείο και συνδέει αυτά που εφάπτονται μεταξύ τους. Όμως με αυτόν τον τρόπο μπορεί να προκύψουν τμήματα γράφου που δεν είναι συνδεδεμένα μεταξύ τους, λόγω αραιών περιοχών στον χώρο των δεδομένων ή και σημεία που γύρω τους εφάπτονται παραπάνω από μια ακτίνες, στις πυκνότερες περιοχές. Το δεύτερο πρόβλημα σε συνδυασμό με την κατάρα της διαστατικότητας για τα πολυδιάστατα δεδομένα, κάνει τα σημεία να απέχουν σχεδόν το ίδιο από τα γειτονικά τους και ως αποτέλεσμα ο αλγόριθμος αδυνατεί να εξάγει κάποια ουσιαστική τοπολογική πληροφορία για την κατασκευή συνεχούς γράφου.



Εικόνα 22 Δημιουργία συνάψεων δεδομένων με βάση τις αποστάσεις ανάμεσά τους. Σε δεδομένα που έχουν μεγάλη ακτίνα σχηματίζονται συνάψεις με μικρότερο βάρος.

Το παραπάνω πρόβλημα λύνει η επιλογή μιας μεταβλητού μέτρου ακτίνας, μεγάλης για αραιές περιοχές και μικρής για πυκνές, αντί για την χρήση ακτίνας σταθερού μέτρου, χωρίς αυτό να επηρεάζει την πληροφορία που θα εξαχθεί, γεγονός που έχει αποδειχτεί και μαθηματικά μέσω της γεωμετρίας Riemannian. Σημαντική είναι η επιλογή κατάλληλου, με βάση τα αρχικά δεδομένα, αριθμού γειτόνων

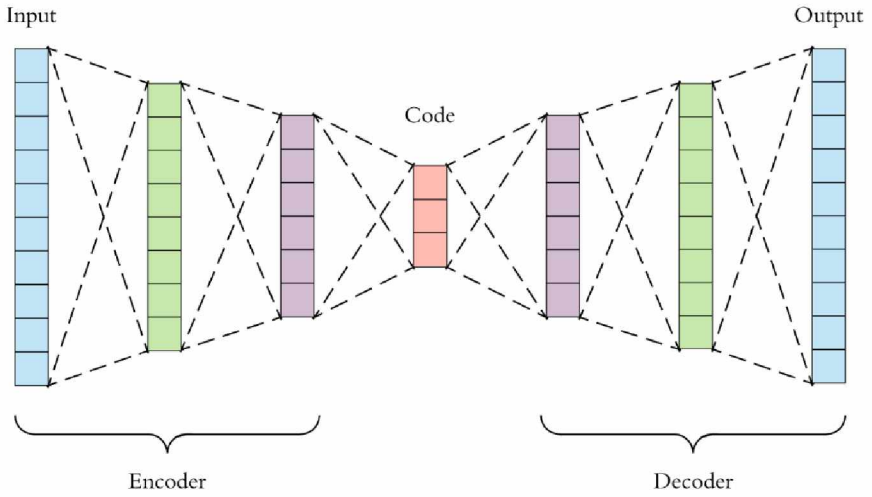
k, αφού όσο μικρότερο είναι το k τόσο πιο πολύ διατηρείται η δομή των δεδομένων σε τοπικό επίπεδο, ενώ όσο μεγαλώνει διατηρείται η δομή τους σε μεγαλύτερη κλίμακα. Άρα πρέπει να επιλεχθεί κατάλληλη τιμή για να διατηρηθεί μια ισορροπία ανάμεσα στην διατήρηση της τοπικής (local) ή ολικής (global) δομής των δεδομένων. Όσο μεγαλύτερη είναι η απόσταση των συνδεδεμένων σημείων, τόσο μικρότερο βάρος δίνεται σε αυτές τις διασυνδέσεις και τόσο πιο μικρή είναι η πιθανότητά τους. Έτσι δημιουργείται ο τελικός γράφος, στον οποίο διατηρούνται οι τοπολογικές πληροφορίες των δεδομένων σε τοπικό και καθολικό επίπεδο, που, στην συνέχεια, προβάλλεται σε χώρο λιγότερων διαστάσεων μέσω ενός αλγορίθμου διάταξης ενός κατευθυνόμενου γραφήματος. Σε αυτόν τον χώρο ο νέος γράφος που δημιουργείται επιλέγει να διατηρήσει μόνο τις πιο στενές και πιθανές διασυνδέσεις ανάμεσα στα σημεία δημιουργώντας και χωρίζοντας σε ομάδες τα αρχικά δεδομένα. Συμπερασματικά η UMAP είναι γρηγορότερη και διατηρεί μεγάλη ποσότητα τοπολογικής πληροφορίας (global και local) σε σχέση με άλλες παρόμοιες μεθόδους όπως η t-SNE.

3.2.8. Autoencoders

Μια σχετικά νέα μέθοδος μείωσης διαστάσεων που προτάθηκε ε το 1986 από την ομάδα του G. E. Hinton και επανήλθε σχετικά πρόσφατα ως τεχνική βαθιάς μάθησης (Hinton and Salakhutdinov, 2006), είναι οι Autoencoders ή νευρωνικά δίκτυα κρυπτογράφησης αποκρυπτογράφησης . Οι Autoencoders είναι νευρωνικά δίκτυα που χρησιμοποιούνται για την εύρεση κωδικοποιημένων αναπαραστάσεων, μικρότερης διάστασης, πολυδιάστατων δεδομένων και σκοπός τους είναι η ανακατασκευή των δεδομένων αυτών από την κωδικοποιημένη, μικρότερης διάστασης, αναπαράστασή τους χρησιμοποιώντας οπισθοδιάδωση ή backpropagation.

Θεωρείται ότι οι Autoencoders ανήκουν στις τεχνικές μη επιβλεπόμενης μάθησης, όμως μπορούν να χαρακτηριστούν και ως self-supervised, καθώς οι ετικέτες των δεδομένων παράγονται από τα ίδια τα δεδομένα (A Gentle Introduction to LSTM Autoencoders, Nov. 2018). Η δομή τους είναι παρόμοια με εκείνη του Multilayer Perceptron, όμως αντί να υπολογίζει μια έξοδο y από την είσοδο χ, οι Autoencoders ανακατασκευάζουν την είσοδο x στην έξοδο. Πιο συγκεκριμένα, οι Autoencoders αποτελούνται από τρία επιμέρους επίπεδα που είναι ο κωδικοποιητής (Encoder) , το κρυφό επίπεδο με την συμπιεσμένη είσοδο (Code / Bottleneck / latent space) και ο αποκωδικοποιητής (Decoder). Το επίπεδο του κωδικοποιητή συμπιέζει τα δεδομένα εισόδου σε μια μικρότερη αναπαράσταση (μικρότερης διάστασης), όπου τα παρόμοια δεδομένα βρίσκονται πιο κοντά στον χώρο. Ο χώρος αυτός ονομάζεται λανθάνων ή latent space. Το επίπεδο του αποκωδικοποιητή παίρνει την κωδικοποιημένη

πληροφορία του μεσαίου επιπέδου ως είσοδο και προσπαθεί να ανακατασκευάσει τα αρχικά δεδομένα εισόδου, με την μέγιστη δυνατή ακρίβεια, στην αρχική τους διάσταση.



Εικόνα 23 Λεπτομερής περιγραφή της εσωτερικής δομής ενός Autoencoder

Η λειτουργία του encoder περιγράφεται σε από την σχέση $y = s(Wx + b)$, όπου s είναι κάποια μη γραμμική συνάρτηση ενεργοποίησης όπως η σιγμοειδής (Sigmoid), η Rectified linear (ReLU) ή Ανορθωτής, η συνάρτηση ενεργοποίησης υπερβολικής εφαπτομένης (tanh) και άλλες. Η λειτουργία του decoder μπορεί να περιγραφεί μαθηματικά ως $x_r = s(W'x + b')$, όπου τα W′ και b' δεν είναι απαραίτητα οι αντιστροφές των προηγούμενων πινάκων.

Η ποιότητα της ανακατασκευής των αρχικών δεδομένων από τον decoder, καθορίζεται από μια συνάρτηση απώλειας ή loss function. Όσο πιο μικρή η τιμή της συνάρτησης αυτής, τόσο πιο πολύ όμοια είναι τα δεδομένα της εξόδου με τα αρχικά δεδομένα και κατά συνέπεια, η συμπιεσμένη αναπαράσταση των δεδομένων στον λανθάνοντα χώρο είναι ακριβέστερη. Στην παρούσα εργασία χρησιμοποιήθηκε ως συνάρτηση απώλειας του μοντέλου Autoencoder που κατασκευάστηκε, η MSE ή μέσο τετραγωνικό σφάλμα. Αυτή υπολογίζει τα τετράγωνα των αποστάσεων ανάμεσα στην πρόβλεψη που έκανε το νευρωνικό δίκτυο και τιμή της εισόδου που έπρεπε να προβλεφθεί. Μαθηματικά περιγράφεται ως $MSE = \frac{1}{n}\sum(x - x_r)^2$, όπου x_r η πρόβλεψη και x η αρχική είσοδος.

Υπάρχουν πολλές πρακτικές εφαρμογές των Autoencoder οι οποίες τον καθιστούν ιδιαίτερα χρήσιμο εργαλείο σε πολλές περιπτώσεις. Εκτός από την μείωση της διάστασης πολυδιάστατων δεδομένων, χρησιμοποιείται για ανίχνευση ανωμαλιών και την αφαίρεση θορύβου από εικόνες. Ακόμη είναι

διαθέσιμες διάφορες μορφές και τύποι Autoencoder, που ο καθένας χρησιμοποιείται και για διαφορετικούς λόγους όπως ο απλός ή «Vanilla Autoencoder», ο Convolutional Autoencoder, που χρησιμοποιείται κυρίως για εικόνες, ο Variational Autoencoder, όταν χρειάζεται μεγαλύτερη ακρίβεια και έλεγχος της συμπιεσμένης αναπαράστασης των δεδομένων του λανθάνοντα χώρου, ο Sparse Autoencoder και ο Denoising Autoencoder που ανήκουν στην κατηγορία των Regularized Autoencoders κ.α.. Στην εργασία αυτή θα γίνει εκτενέστερη αναφορά στον Variational Autoencoder, αφού αυτός ο τύπος επιλέχθηκε να εφαρμοστεί ως τεχνική μείωσης διαστάσεων για τα δεδομένα από scRNA-Seq, συνδυαστικά και με άλλες μεθόδους, με σκοπό τη βέλτιστη ανάλυση και κατανόησή τους. Είναι σημαντικό να αναφερθεί ότι υπάρχουν ήδη περιπτώσεις που οι Autoencoders χρησιμοποιούνται με σκοπό την εξαγωγή πληροφορίας από βιολογικά δεδομένα τύπου scRNA-seq (Yue Deng, Feng Bao et al., 2019; Eraslan et al., 2019).

3.3. Αλγόριθμοι μείωσης διαστάσεων και feature extraction δεδομένων από scRNA-Seq

Λόγω της ποιότητας των δεδομένων από scRNA-Seq και των ιδιαίτερων χαρακτηριστικών τους είναι δύσκολη, όπως έχει ήδη αναφερθεί, η ερμηνεία και η εύρεση μοτίβων σε ακατέργαστα (raw) δεδομένα αυτού του τύπου. Για αυτό το λόγο θεωρείται κρίσιμης σημασίας η εύρεση ποιοτικών και αποδοτικών μεθόδων μείωσης της διάστασης αυτών των δεδομένων, για να εξαχθούν πληροφορίες με βιολογική σημασία, που η ερμηνεία τους θα οδηγήσει σε συμπεράσματα που μπορούν να χρησιμεύσουν στην βιολογική και ιατρική έρευνα ή την ανάπτυξη εξατομικευμένων μεθόδων θεραπείας. Κάποιες από τις υπάρχουσες ήδη μεθόδους μπορούν να εφαρμοστούν με επιτυχία και σε αυτού του είδους τα δεδομένα, ενώ έχουν γίνει και προσπάθειες δημιουργίας περισσότερων ειδικών μεθόδων που λαμβάνουν υπόψη τους τα ειδικά χαρακτηριστικά των δεδομένων από scRNA-Seq, όπως παρουσιάζονται παρακάτω.

3.3.1. Generalized Linear Model Principal Component Analysis (GLM-PCA)

Η GLM-PCA είναι μια μέθοδος που αναπτύχθηκε πρόσφατα (B. Du, X. Kong, X. Feng, 2020),(Landgraf and Lee, 2020) για περιπτώσεις που τα δεδομένα δεν ακολουθούν την κανονική κατανομή, όπως στις περιπτώσεις δεδομένων από scRNA-seq, καθώς μπορεί να αναιρέσει το φαινόμενο «batch effect», όπου τα δεδομένα επηρεάζονται από μη βιολογικούς παράγοντες κατά τη διεξαγωγή ενός πειράματος μοριακής βιολογίας. Σε αυτήν την περίπτωση οι υπάρχουσες τεχνικές μπορεί να καταλήξουν σε ψευδείς πληροφορίες σχετικά με την μεταβλητότητα των δεδομένων (variability) και να βρουν διαφοροποιήσεις που δεν υπάρχουν πραγματικά σε αυτού του είδους τα δεδομένα κατά την μείωση της διάστασής τους. Αντίθετα η τεχνική αυτή μπορεί να χρησιμοποιηθεί σε

δεδομένα χωρίς να έχει προηγηθεί κάποιου είδους επεξεργασία ή κανονικοποίηση, ενώ συγκεκριμένα για τα scRNA-seq, αν αντί για τον αριθμό των αναγνώσεων (reads) χρησιμοποιηθούν οι μετρήσεις UMI (UMI counts), που αντιπροσωπεύουν τον απόλυτο αριθμό των παρατηρούμενων μεταγράφων ανά κύτταρο, τότε θα λυθεί το πρόβλημα των zero-inflated δεδομένων. Το πρόβλημα αυτό, όπως έχει ήδη αναφερθεί, είναι ένα από τα πιο βασικά προβλήματα των δεδομένων από τεχνική αλληλούχισης μεμονωμένου κυττάρου, καθώς υπάρχουν πολλές μηδενικές τιμές που δεν μπορούν να χρησιμοποιηθούν σωστά για να εξαχθεί κάποια σημαντική πληροφορία, καθώς το ποσό των παρατηρούμενων είναι μεγαλύτερο από το ποσό των προβλεπόμενων μηδενικών.

Το συγκεκριμένο μοντέλο εμπνεύστηκε από τα γενικευμένα γραμμικά μοντέλα και την οικογένεια των εκθετικών κατανομών πιθανοτήτων, όπως τις κατανομές Bernoulli, Gaussian και τα μοντέλα Markov και μπορεί να εφαρμοστεί σε οποιαδήποτε τέτοια κατανομή. Αν τα διαθέσιμα δεδομένα είναι μετρήσεις, μπορεί να χρησιμοποιηθεί Poisson ή αρνητική διωνυμική πιθανότητα. Έστω Y η μεταβλητή αποτελέσματος. Μια θεμελιώδης πτυχή των GLM είναι ότι το μοντέλο θορύβου θεωρείται ότι ακολουθεί μια εκθετική κατανομή πιθανότητας. Το παραπάνω μπορεί μαθηματικά να αναπαρασταθεί ως εξής:

$$\log f_Y(y;\theta) = c(y) + y\theta - \kappa(\theta)$$

Όπου θ φυσική παράμετρος. Η φυσική παράμετρος είναι στην πραγματικότητα συνάρτηση του μέσου όρου $\theta = \theta(\mu)$. Η μέση τιμή είναι $\kappa'(\theta) = \mu$ και η διακύμανση είναι $\kappa''(\theta)$. Σκοπός της GLMPCA είναι να βρει το καλύτερο rank ή βαθμός k, που δηλώνει τον μέγιστο αριθμό των γραμμικά ανεξάρτητων στηλών ή διαστάσεων, για να προβάλλει τον πίνακα με τις φυσικές παραμέτρους από το κορεσμένο μοντέλο $\tilde{\Theta}$ (δηλαδή το μοντέλο που ταιριάζει απόλυτα στα δεδομένα επειδή έχει τόσες εκτιμώμενες παραμέτρους, όσες και οι τιμές που πρέπει να προσαρμοστούν). Σκοπός της διαδικασίας είναι η ελαχιστοποίηση της κατάλληλης απόκλισης (deviance) από τα δεδομένα και η μεγιστοποίηση της αντικειμενικής συνάρτησης (Hunter and Lange, 2004).

3.3.2. Zero-Inflated Factor Analysis (ZIFA)

Η μέθοδος ZIFA ή Zero-Inflated Factor Analysis (Pierson and Yau, 2015) χρησιμοποιείται για μείωση διαστάσεων δεδομένων από scRNA-seq που είναι μηδενικά διογκωμένα ή zero-inflated. Οι ήδη υπάρχουσες μέθοδοι συνήθως δεν λαμβάνουν σοβαρά υπόψη τους τον μεγάλο αριθμό των μηδενικών

στα δεδομένα και τα ιδιαίτερα χαρακτηριστικά τους , γεγονός που επηρεάζει σημαντικά την απόδοση των μεθόδων μείωσης διαστάσεων. Πιο συγκεκριμένα η μέθοδος ZIFA βασίζεται στη μέθοδο ανάλυσης παραγόντων ή Factor Analysis, στο οποίο προσθέτει και ένα επιπλέον επίπεδο διαμόρφωσης μηδενικού πληθωρισμού. Η μέθοδος ZIFA προϋποθέτει οι διαχωρίσιμες καταστάσεις ή υποτύποι των κυττάρων να υπάρχουν αρχικά ως σημεία σε έναν λανθάνοντα (μη παρατηρητέο) χώρο χαμηλής διάστασης ή latent space. Αυτά στη συνέχεια προβάλλονται σε έναν λανθάνοντα χώρο γονιδιακής έκφρασης υψηλών διαστάσεων μέσω ενός γραμμικού μετασχηματισμού και της προσθήκης θορύβου μέτρησης κανονικής κατανομής. Κάθε μέτρηση έχει τότε κάποια πιθανότητα να μηδενιστεί μέσω του μοντέλου dropout ή μοντέλου μηδενισμού που διαμορφώνει τη λανθάνουσα κατανομή των τιμών έκφρασης. Αυτό επιτρέπει την ενσωμάτωση και χρήση της πληροφορίας που προέρχεται και από τα παρατηρούμενα μη διογκωμένα δεδομένα γονιδιακής έκφρασης ενός κυττάρου. Η παράμετρος κλιμάκωσης στο μοντέλο dropout μπορεί χρησιμεύσει στην εύρεση και ερμηνεία των μηδενικών τιμών των δεδομένων και την ανακάλυψη μοτίβων.

Έστω N ο αριθμός των δειγμάτων, D ο αριθμός γονιδίων και K ο επιθυμητός αριθμός τελικών διαστάσεων. Τα δεδομένα (N×D) είναι σε μορφή πίνακα Y= [y 11,…,y ND], όπου για y_{ij},το i δηλώνει το δείγμα ή κύτταρο και το j το αντίστοιχο γονίδιο. Τα δεδομένα θεωρείται ότι παράγονται από την προβολή των σημείων σε έναν χώρο μικρότερης διάστασης (N×K) Z= [z 1,…,z N] όπου K≪D. Ισχύουν τα παρακάτω:

$$z_i \sim Normal(0, I), x_i | z_i \sim Normal(Az_i + \mu, W)$$

$$h_{i,j} | x_{i,j} \sim Bernoulli\,(p_0), \quad y_{i,j} = \begin{cases} x_{i,j}, h_{i,j} = 0 \\ 0, h_{i,j} = 1 \end{cases}$$

Όπου I ο ταυτοτικός πίνακας μεγέθους KxK, A ο πίνακας μεγέθους DxK με τους παράγοντες από την ανάλυση παραγόντων, H ο πίνακας μεγέθους DxN που χρησιμοποιείται ως μάσκα, W=diag(σ_1^2,…,σ_D^2) πίνακας μεγέθους DxD και μ ένας πίνακας μέσων όρων μεγέθους Dx1. Θέτουμε την πιθανότητα μηδενισμού της συνάρτησης dropout του λανθάνοντα χώρου ως $p_0 = \exp(-\lambda x_{i,j}^2)$, όπου λ είναι μια σταθερά εκθετικής μείωσης του μοντέλου μηδενικού πληθωρισμού.

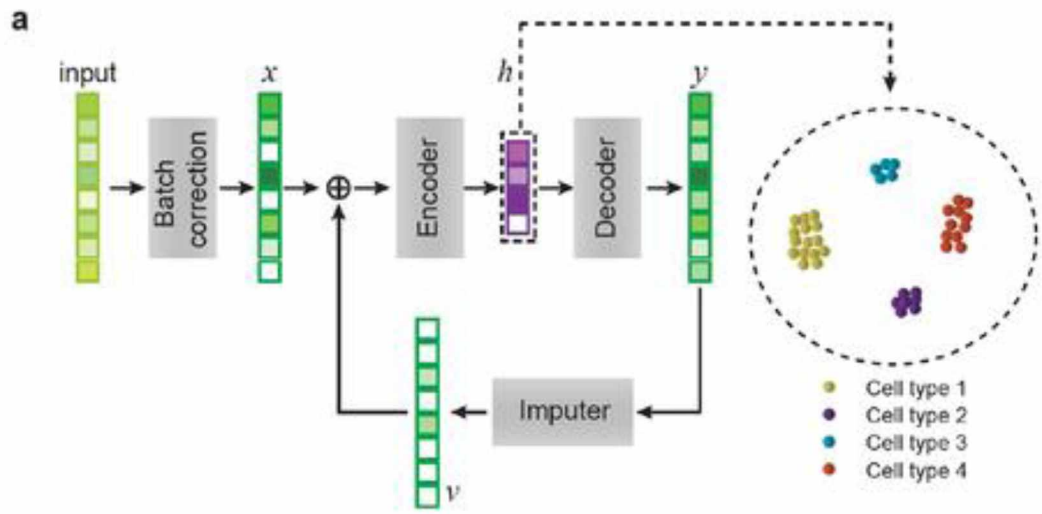
Αρχικά, με δεδομένο έναν πίνακα με τις εκφράσεις των γονιδίων Υ, η μέθοδος αρχικοποιεί τις τιμές μ, σ^2, λ και τον πίνακα Α, για τις οποίες προσπαθεί να υπολογίσει τις βέλτιστες τιμές, ώστε να μεγιστοποιηθεί η πιθανότητα $p(Y|\theta)$ όπου Θ = [Α,σ^2,μ,λ]. Σε κάθε βήμα υπολογίζονται νέες τιμές για τις παραμέτρους του Θ και νέα πιθανότητα p με βάση αυτές. Ο αλγόριθμος ολοκληρώνεται όταν η πιθανότητα p λάβει την αναμενόμενη τιμή που μεγιστοποιεί την λογαριθμική συνάρτηση πιθανοφάνειας (complete log likelihood).

Προκειμένου να μειωθεί ο χρόνος εκτέλεσης, υπάρχει η δυνατότητα διαχωρισμού των γονιδίων κάθε κυττάρου σε μη επικαλυπτόμενες ομάδες και η εκτέλεση του αλγόριθμου, χωρίς αυτό να επηρεάζει την αποτελεσματικότητά του.

3.3.3. scScope

Η μέθοδος scScope (Y. Deng, F. Bao et al., 2019) είναι μια μέθοδος βαθιάς μάθησης που χρησιμοποιεί αναδρομικά νευρωνικά δίκτυα (recurrent neural networks), για να αποφασίσει από πού προέρχονται οι μηδενικές ή σχεδόν μηδενικές αναγνώσεις στα δεδομένα από scRNA-seq και αν από αυτές μπορεί να εξαχθεί κάποια σημαντική πληροφορία βιολογικής σημασίας ή αν οφείλεται σε κάποιο τεχνικό λάθος ή βιολογική διαδικασία και δεν θα έπρεπε να ληφθεί υπόψη. Η διαδικασία επιτυγχάνεται μέσα από μια σειρά Τ βημάτων. Για Τ=1 η μέθοδος μετατρέπεται σε έναν απλό Autoencoder. Για την προεπιλεγμένη τιμή Τ=2 έχει βρεθεί ότι η scScope έχει την βέλτιστη ισορροπία ανάμεσα στην απόδοση και την ταχύτητα επεξεργασίας των δεδομένων. Η μέθοδος αυτή μπορεί με επιτυχία να αναιρέσει το φαινόμενο «batch effect», που συμβαίνει όταν μη βιολογικοί παράγοντες σε ένα πείραμα προκαλούν αλλαγές στα δεδομένα που εξάγονται από αυτό, την εκμάθηση κυτταρικών χαρακτηριστικών, τον καταλογισμό των «dropout events», δηλαδή των περιπτώσεων που σε ένα κύτταρο γίνεται έκφραση ενός γονιδίου σε μικρό ποσοστό ενώ δεν εκφράζεται καθόλου σε άλλα κύτταρα αλλά και την παράλληλη εκπαίδευση του μοντέλου.

Η scScope αποτελείται από τέσσερα επίπεδα, σε καθένα από τα οποία οι παράμετροι υφίστανται βελτιστοποιήσεις. Τα δεδομένα εισόδου δεν πρέπει να περιλαμβάνουν αρνητικές τιμές. Ακόμα η μέθοδος δίνει τη δυνατότητα διόρθωσης του λεγόμενου «batch effect» το οποίο, όπως αναφέρθηκε νωρίτερα, μπορεί να επηρεάσει σημαντικά και ουσιαστικά τα αποτελέσματα που εξάγονται από το μοντέλο.



Εικόνα 24 Δομή του scScope

Το επίπεδο διόρθωσης του παραπάνω φαινομένου fB() περιγράφεται και μαθηματικά:

$$x_c = f_B(\tilde{x}_c) = r(\tilde{x}_c - Bu_c)$$

Όπου $\tilde{x}_c$ είναι το προφίλ του κυττάρου εισόδου, το δυαδικό πειραματικό διάνυσμα δείκτη επιδράσεων κάθε «παρτίδας» ή batch $u_c \in \mathbb{R}^k$ (η μηδενική καταχώριση υποδηλώνει την παρτίδα των $\tilde{x}_c$), όπου K ο αριθμός της παρτίδας και $B \in \mathbb{R}^{N x K}$ ο πίνακας διόρθωσης της παρτίδας που εξάγεται από το δίκτυο. Το r δηλώνει την συνάρτηση ενεργοποίησης RELU.

Στην συνέχεια υπάρχει το επίπεδο του κωδικοποιητή (Encoder), που μειώνει και συμπιέζει τις διαστάσεις των αρχικών δεδομένων $x_c \in \mathbb{R}^N$, στον πίνακα μειωμένης διάστασης $h_c \in \mathbb{R}^M$, όπου M<N. Μαθηματικά μπορεί να περιγραφεί και ως:

$$h_c = f_E(x_c) = r(W_E x_c + b_E)$$

Όπου οι παράμετροι $W_E \in \mathbb{R}^{MxN}$ και $b_E \in \mathbb{R}^M$ μαθαίνονται και βελτιστοποιούνται κατά την διάρκεια του αλγόριθμου.

Το επίπεδο του αποκωδικοποιητή (Decoder) αποσυμπιέζει τα δεδομένα από τον λανθάνοντα χώρο μειωμένης διάστασης $h_c \in \mathbb{R}^M$ σε έναν χώρο $y_c \in \mathbb{R}^N$, ίδιας διάστασης με τα αρχικά δεδομένα και η λειτουργία του δίνεται από τον τύπο:

$$y_c = f_D(h_c) = r(W_D h_c + b_D)$$

Με παραμέτρους εκμάθησης τα $W_E \in \mathbb{R}^{NxM}$ και $b_E \in \mathbb{R}^N$.

Τέλος υπάρχει το επίπεδο απόδοσης, τεκμαρτό επίπεδο ή imputer, το οποίο είναι επίπεδο αυτοδιόρθωσης και αποδίδει τιμές ή υπολογίζει τις καταχωρήσεις που λείπουν. Με σκοπό τη μείωση των παραμέτρων που πρέπει να υπολογιστούν, το επίπεδο του αποκωδικοποιητή εκφράστηκε τελικά ως εξής:

$$u_c = r\left(W_U y_c + b_U\right) \in \mathbb{R}^p$$

Όπου το p έχει οριστεί ως 64. Η απόδοση τιμών ορίζεται μαθηματικά ως:

$$u_c = Pz_c[r(W_V u_c + b_V)] = r(W_D h_c + b_D)$$

Το Z_c περιλαμβάνει τα μηδενικά εκτιμημένα γονίδια του κυττάρου c. Όλες οι τιμές που δεν ανήκουν σε αυτό εκτιμώνται ως μηδενικές από τον χειριστή Pz_c. Αφού προκύψει το διάνυσμα v_c, υπολογίζεται το διορθωμένο προφίλ έκφρασης ενός κυττάρου $\hat{x}_c = x_c + v_c$. Αυτό το διορθωμένο προφίλ αποστέλλεται ξανά στον encoder προκειμένου να γίνει εκ νέου εκμάθηση μια νέας λανθάνουσας αναπαράστασης μειωμένης διάστασης. Όλες οι παραπάνω παράμετροι υπολογίζονται και διορθώνονται με σκοπό την ελαχιστοποίηση της συνάρτησης απώλειας Pz_c που βελτιστοποιεί τις παραμέτρους χρησιμοποιώντας μόνο τις μη μηδενικές καταχωρήσεις. Μέσω της παραπάνω διαδικασίας επιτυγχάνεται η μείωση των διαστάσεων των δεδομένων. Η μέθοδος εστιάζει στις μη μηδενικές τιμές και αφού η διαδικασία ολοκληρωθεί είναι πλέον ευκολότερη η εύρεση μοτίβων και η ανακάλυψη υποπληθυσμών κυττάρων που ήταν δυσκολότερο λόγω των «dropout events» να βρεθούν σωστά.

3.3.4. Άλλες χρησιμοποιούμενες μέθοδοι μείωσης διαστάσεων δεδομένων scRNA-seq

Φυσικά, οι παραπάνω δεν είναι οι μόνες μέθοδοι που έχουν ανακαλυφθεί για μείωση της διάστασης των δεδομένων από scRNA-seq, είναι όμως μέθοδοι που θα μπορούσαν πιθανόν να συνδυαστούν με την μέθοδο που ακολουθεί και χρησιμοποιήθηκε στην παρούσα εργασία, με σκοπό την βελτιστοποίησή της. Κάθε μέθοδος από τις παραπάνω έχει και διαφορετική προσέγγιση, όμως οι δύο τελευταίες εστιάζουν στο πρόβλημα που δημιουργείται από την μεγάλη ύπαρξη μηδενικών ή σχεδόν μηδενικών

τιμών λόγω των γεγονότων dropout και προσπαθούν να δώσουν μια ερμηνεία και να χρησιμοποιήσουν αυτού του είδους τα δεδομένα. Τα δεδομένα από αλληλούχιση μεμονωμένου κυττάρου έχουν μεγάλο ποσοστό θορύβου και ιδιαίτερα χαρακτηριστικά και για αυτόν τον λόγο, όπως έχει ήδη αναφερθεί, δεν είναι πάντα αποτελεσματική η χρήση των μεθόδων που περιεγράφηκαν στο προηγούμενο κεφάλαιο. Από τις μεθόδους εκείνες οι πιο αποδοτικές και χρησιμοποιούμενες είναι κυρίως η μέθοδος PCA και η t-SNE. Μέθοδοι ειδικά ανεπτυγμένες για βιολογικά δεδομένα είναι ακόμα η probabilistic Count Matrix Factorization (pCMF (Durif *et al.*, 2019)), ισχυρή για οπτικοποίηση και ομαδοποίηση δεδομένων από scRNA-seq, η Deep Count Autoencoder network (DCA) (Eraslan, Simon et al., 2019), η ZINB-WaVE (Risso *et al.*, 2018)κ.α..

4. Βιβλιογραφική Ανασκόπηση

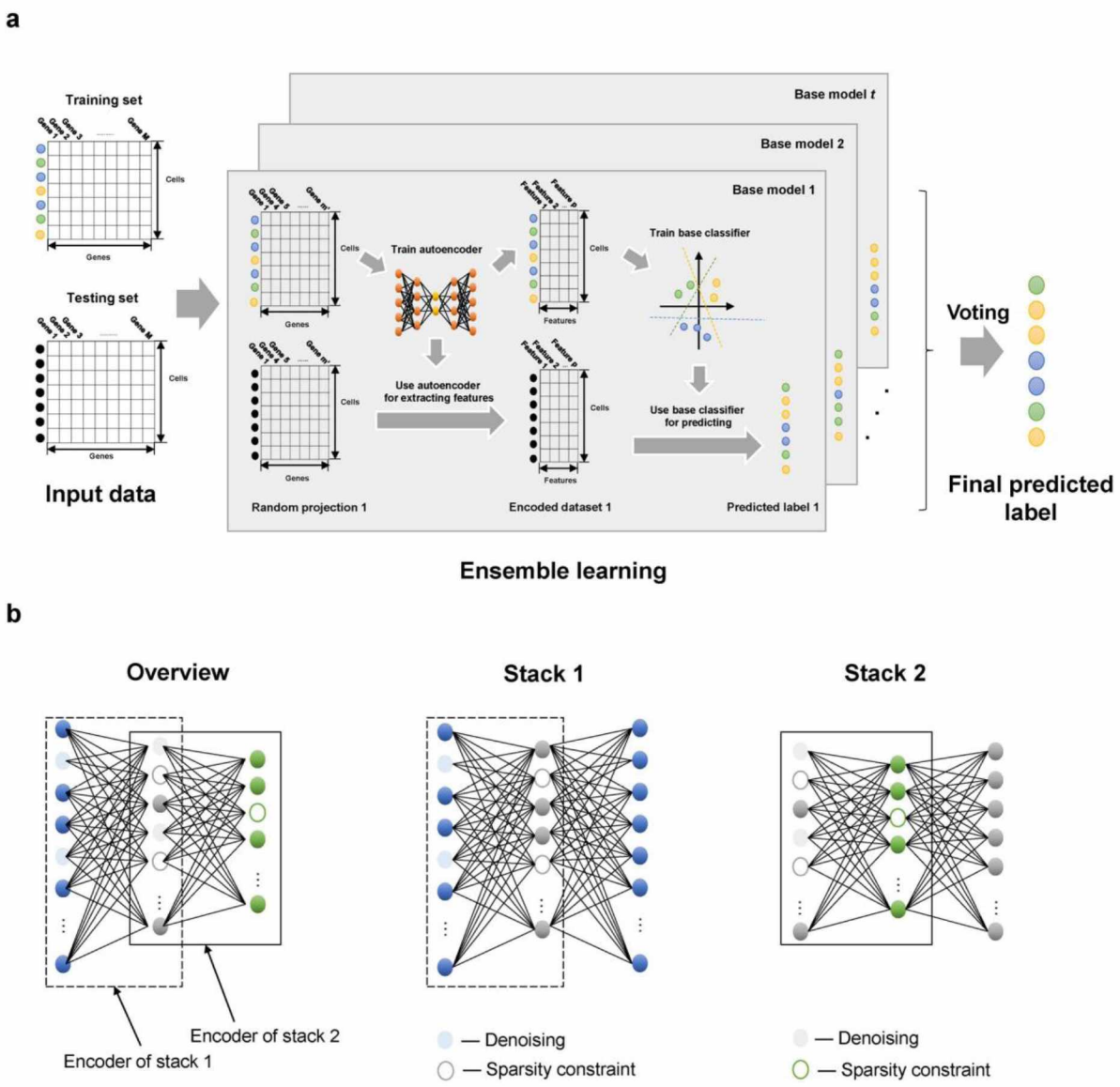
Παρακάτω παρουσιάζονται μοντέλα που αναπτύχθηκαν με σκοπό τη μείωση διαστάσεων, την ανάλυση και την εξαγωγή χαρακτηριστικών με σκοπό την κατηγοριοποίηση και οπτικοποίηση των βιολογικών δεδομένων scRNA-Seq.

4.1. Μέθοδος scIAE

Τα δεδομένα από scRNA-Seq εκφράζουν την ποσοτική ανάλυση της γονιδιακής έκφρασης, απαραίτητη για την μελέτη της κυτταρικής ετερογένειας. Η ανάλυση της κυτταρικής ετερογένειας μπορεί να φανεί ιδιαίτερα χρήσιμη στην κατασκευή ενός καθολικού ευρετηρίου, για όλους τους τύπους κυττάρων ή για περίπλοκους ιστούς και οργανισμούς και αποτελεί τη βάση για κάθε μεταγενέστερη ανάλυση των βιολογικών δεδομένων. Ακόμη η ανάλυση αυτού του είδους των δεδομένων, όταν προέρχονται από άτομα ή οργανισμούς που πάσχουν από μία ασθένεια, μπορεί να οδηγήσει στην ανακάλυψη γονιδίων που την προκαλούν, την εύρεση περισσότερο εξειδικευμένων μεθόδων θεραπείας αλλά και την ευκολότερη διάγνωση παρόμοιων περιπτώσεων ασθενειών. Ο αυτισμός είναι μια περίπλοκη αναπτυξιακή διαταραχή που εμφανίζει μεγάλη ετερογένεια ανά περίπτωση, καθώς η έκφρασή του μπορεί να διαφέρει σημαντικά από άτομο σε άτομο. Η διαφοροποίηση αυτή όμως μπορεί να αποτελέσει πρόβλημα στην έγκαιρη αντιμετώπιση και σωστή, προσωποποιημένη διάγνωσή του. Σε τέτοιες περιπτώσεις η ανάλυση των δεδομένων από scRNA-seq μπορεί πραγματικά να βελτιώσει σε μεγάλο βαθμό την διάγνωση και αντιμετώπιση των δυσκολιών που προκύπτουν λόγω της ετερογένειας που εμφανίζεται ανά περίπτωση.

Ακόμη, η κλασική μέθοδος σχολιασμού των διαφόρων κυτταρικών τύπων παρουσιάζει κάποιες δυσκολίες τόσο ως προς την ανάγκη για εφαρμογή μεθόδων ομαδοποίησης, που εκτελείται πειραματικά, μέσω της εύρεσης κατάλληλων παραμέτρων, όσο και ως προς την έλλειψη αντιστοίχισης των κυτταρικών τύπων που τελικά προκύπτουν, με κάποια αντίστοιχη τυποποιημένη ετικέτα του συγκεκριμένου κυτταρικού τύπου.

Για να αντιμετωπιστούν τα παραπάνω, προτείνεται το μοντέλο scIAE (ensemble classifier framework based on Autoencoders), (Yin et al., Jan 2022). Πιο συγκεκριμένα στο μοντέλο αυτό γίνεται χρήση denoising, sparse και integrating stacked Autoencoders, για την εξαγωγή μια συμπιεσμένης αναπαράστασης των αρχικών πολυδιάστατων δεδομένων.



Εικόνα 25 Λεπτομερής δομή του μοντέλου scIAE

Με βάση τη μέθοδο scIAE υπάρχουν κάποια βασικά επιμέρους μοντέλα, για καθένα από τα οποία εφαρμόζεται μία τυχαία προβολή, με σκοπό την απόκτηση ενός τυχαίου υποχώρου του αρχικού συνόλου δεδομένων (training και testing). Πάνω σε αυτόν τον τυχαίο υποχώρο χρησιμοποιούνται τα δεδομένα εκπαίδευσης για να μπορέσει κάθε φορά ο Autoencoder να παράξει έναν χώρο μειωμένης διάστασης. Στη συνέχεια αυτός ο χώρος χρησιμοποιείται από τον εκπαιδευμένο πλέον stacked, denoising, sparse Autoencoder, για να συμπιέσει τα δεδομένα ελέγχου. Από το χώρο αυτό με τις λιγότερες διαστάσεις ή λανθάνοντα χώρο, χρησιμοποιούνται τα δεδομένα εκπαίδευσης για να εκπαιδεύσουν έναν βασικό ταξινομητή, ο οποίος τελικά προβλέπει τις κλάσεις των δεδομένων ελέγχου του λανθάνοντος χώρου. Όλες οι προβλέψεις από τα επιμέρους μοντέλα συνενώνονται και η τελική πρόβλεψη προκύπτει μέσω συμ-

ψηφισμού από τις προβλέψεις κάθε βασικού μοντέλου. Η έξοδος που προκύπτει από τον scIAE είναι το τελικό αποτέλεσμα ή πρόβλεψη κλάσης για κάθε ένα δείγμα του μοντέλου. Οι επιλογές ανάμεσα σε βασικούς ταξινομητές που διατίθενται είναι ο KNN, SVM, DT, PLSDA. Μάλιστα για τους τρεις τελευταίους δίνεται η δυνατότητα απόρριψης της πρόβλεψης (rejection option), αν για κάποιο δείγμα προκύψει πιθανότητα σωστής πρόβλεψης ετικέτας μικρότερης από ένα συγκεκριμένο κατώφλι που μπορεί να οριστεί από το χρήστη.

Όσον αφορά τη μέθοδο που περιεγράφηκε παραπάνω, επειδή η χρήση ενός μόνο Autoencoder δεν είχε τα επιθυμητά αποτελέσματα, η εφαρμογή stacked Autoencoder θεωρήθηκε καταλληλότερη για το συγκεκριμένο μοντέλο, αφού κάθε επίπεδο απομονώνει τα πιο χρήσιμα χαρακτηριστικά. Η εφαρμογή ensemble learning methods φαίνεται να αυξάνει την ποικιλία στα δεδομένα με αποτέλεσμα να προκύπτουν καλύτερα αποτελέσματα, αφού το μοντέλο εξασκείται σε περισσότερα και διαφορετικά δείγματα.

Σε σύγκριση με άλλες μεθόδους εξαγωγής χαρακτηριστικών, το συγκεκριμένο φαίνεται να είναι ιδιαίτερα αποδοτικό και ακριβές μοντέλο, ανεξάρτητα από το πλήθος των επιθυμητών μειωμένων διαστάσεων του λανθάνοντος χώρου. Σε σύγκριση με άλλες μεθόδους, ειδικές για κατηγοριοποίηση scRNA-Seq δεδομένων, φαίνεται να έχει μεγάλη ισχύ κατηγοριοποίησης για τους τομείς της πρόβλεψης ασθενειών, κατηγοριοποίησης για όλα τα διαφορετικά είδη κυτταρικών τύπων, για τα διάφορα είδη οργανισμών και για δεδομένα που προέρχονται από διαφορετικές πλατφόρμες ανάλογα με τις συνθήκες συλλογής τους και τα πρωτόκολλα που εφαρμόστηκαν και ακολουθήθηκαν κατά την εξαγωγή τους (π.χ. κάποια δεδομένα ενεργοποιήθηκαν μέσω κάποιας διαδικασίας φθορισμού, ενώ άλλα δεδομένα μπορεί να προέκυψαν από κύτταρα που επεξεργάστηκαν μέσω κάποιας τεχνικής μικρό-ρευστοποιημένων σταγόνων).

Άλλες γνωστές μέθοδοι εξαγωγής χαρακτηριστικών (PCA, ICA, NMF, MDS,...) μπορούν να βρουν χαρακτηριστικά που εξηγούν τις διαφορές (variation) στα αρχικά δεδομένα και να χρησιμοποιηθούν για πολυάριθμες εφαρμογές όπως για την αφαίρεση θορύβου από αυτά, την αποσυνέλιξη (deconvolution) τους και τη μείωση των διαστάσεων τους, όμως δεν διατηρούν τις μη γραμμικές πληροφορίες ή χαρακτηριστικά των αρχικών δεδομένων. Ακόμη οι μη γραμμικές μέθοδοι για εξαγωγή χαρακτηριστικών όπως η Isomap και η UMAP που χρησιμοποιούνται ευρέως, είναι ιδιαίτερα ευαίσθητες όσον αφορά το θόρυβο των αρχικών δεδομένων αλλά και την επιλογή των κατάλληλων παραμέτρων, γεγονός που δεν

τις καθιστά βέλτιστες μεθόδους για δεδομένα από single-cell λόγω των dropout events, το πλήθος των μηδενικών αποτυπώσεων (dropouts) αλλά και τον θόρυβο που προκύπτει.

Μέθοδοι ειδικές για single-cell (ZIFA GLM-PCA, Sc-scope, DAVAE,…) χρησιμοποιούνται κυρίως για οπτικοποίηση και μείωση της διάστασης των δεδομένων αυτών, χωρίς όμως να μπορούν, στο συγκεκριμένο χώρο που προκύπτει, να δώσουν ουσιαστικές και χρήσιμες πληροφορίες που μπορούν να χρησιμοποιηθούν για την ταξινόμηση των κυττάρων.

Αντίθετα οι Autoencoders, που χρησιμοποιούνται στο μοντέλο που περιγράφεται, εξάγουν γραμμικές και μη γραμμικές πληροφορίες από τα αρχικά δεδομένα και η μειωμένη αναπαράσταση τους περιλαμβάνει χρήσιμες πληροφορίες που μπορούν να χρησιμοποιηθούν για την ταξινόμησή τους. Παρόλο που υπάρχουν και μέθοδοι που χρησιμοποιούν Autoencoders για ταξινόμηση scRNA-seq δεδομένων (π.χ. VASC) δεν χρησιμοποιούνται sparse και denoising Autoencoders.

Συμπερασματικά, με τη μέθοδο που παρουσιάζεται προκύπτουν καλά αποτελέσματα ανεξάρτητα από τον αριθμό των διαστάσεων που επιλέχθηκαν για τη μείωση της διάστασης των δεδομένων και ειδικότερα όταν πρόκειται για πολλές διαστάσεις που οι περισσότεροι αλγόριθμοι αποτυγχάνουν. Άρα η επιλογή και ο καθορισμός των τιμών των παραμέτρων και των διαστάσεων του λανθάνοντος χώρου δεν επηρεάζει τόσο το τελικό αποτέλεσμα όσον αφορά την ακρίβεια και την αποτελεσματικότητα της μεθόδου.

Μέσα από τη σύγκριση με διαφορετικούς βασικούς ταξινομητές, φαίνεται ότι βέλτιστη είναι η απόδοση του SVM σε σχέση με τους KNN, DT, PLSDA και με τη χρήση πάντα σταθερών παραμέτρων. Ο αλγόριθμος φαίνεται να χειρίζεται με ευκολία προβλήματα μη ισορροπημένων δεδομένων, δηλαδή δεδομένων όπου το μέγεθος του δείγματος είναι πολύ μικρό για κάποιες συγκεκριμένες κλάσεις. Επίσης εφαρμόστηκε stability test με εφαρμογή του αλγόριθμου στα ίδια δεδομένα επαναλαμβανόμενα, για κάθε πρόβλεψη, με το αποτέλεσμα να φαίνεται να παραμένει σταθερό με πολύ μικρές αποκλίσεις.

Ένας περιορισμός της συγκεκριμένης μεθόδου είναι ότι ο λανθάνων χώρος δεν είναι άμεσα ερμηνεύσιμος με αποτέλεσμα να είναι δύσκολο να βρεθούν συσχετίσεις ανάμεσα στα δεδομένα του λανθάνοντος χώρου και στα αρχικά δεδομένα.

4.2. Μέθοδος MPRV

Τα δεδομένα από scRNA-seq είναι εξαιρετικά περίπλοκα και πολυδιάστατα. Μέσα από αυτά επιτυγχάνεται βαθύτερη κατανόηση της κυτταρικής ετερογένειας και διευκολύνεται η ανακάλυψη πολλών βιοδεικτών ή biomarkers για διάφορες ασθένειες. Η μέθοδος που παρουσιάζεται είναι ιδιαίτερα αποδοτική για την επίτευξη εξαιρετικά χαμηλής διαστατικότητας. Η αύξηση των βιοϊατρικών δεδομένων τα τελευταία χρόνια έχει στέψει την έρευνα στην αναζήτηση υπολογιστικών μεθόδων χειρισμού τους, μέσω της ανάπτυξης της βιοπληροφορικής, της πληροφορικής της υγείας και της κλινικής πληροφορικής. Επιπλέον η χρήση των ωμικών δεδομένων και η ανάπτυξη του human genome project έχει ως στόχο την αντιμετώπιση και την έγκαιρη πρόβλεψη πολλών ασθενειών.

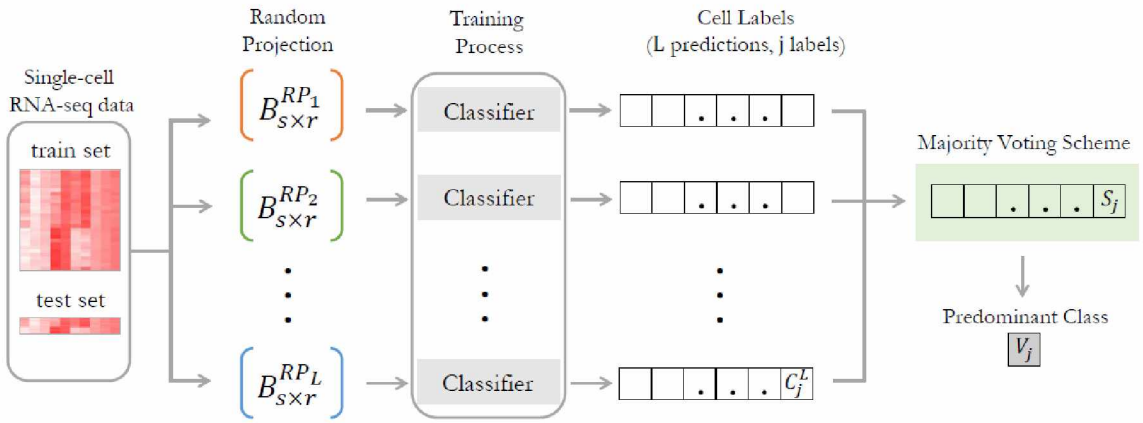
Πλέον είναι εμφανές ότι ο όγκος των δεδομένων αυξάνεται και το κόστος παραγωγής τους μειώνεται. Για αυτόν το λόγο έχει προκύψει ανάγκη για εύρεση μεθόδων μηχανικής μάθησης που μπορούν να χειριστούν αυτού του είδους τα πολυδιάστατα δεδομένα και αντιμετώπισης ενός από τα βασικότερα προβλήματα που προκύπτουν και ονομάζεται κατάρα της διαστατικότητας ή “curse of dimensionality”. Ο χειρισμός αυτών των δεδομένων είναι εμφανές ότι αποτελεί μία σημαντική υπολογιστική πρόκληση.

Προτείνεται η μέθοδος MPRV, (Vrahatis *et al.*, 2020), σκοπός της οποίας είναι η βελτιστοποίηση των ταξινομητών σε δεδομένα μεγάλου όγκου. Στην αρχή επιτυγχάνεται τυχαία προβολή σε έναν χώρο μικρότερης διαστατικότητας. Σύμφωνα με το λήμμα Johnson–Lindenstrauss, με βάση το οποίο ένα σύνολο σημείων σε έναν χώρο υψηλών διαστάσεων μπορεί να προβληθεί σε χώρο μειωμένης διάστασης με τέτοιο τρόπο ώστε οι αποστάσεις και κατά συνέπεια η πληροφορία στα δεδομένα να μπορεί να διατηρηθεί σε μεγάλο βαθμό, μπορεί να υπολογιστεί η τιμή $r < O(\log n / e^2)$. Το r αποτελεί το κάτω φράγμα, δηλαδή τις ελάχιστες διαστάσεις που μπορούν να επιλεχθούν για τη διατήρηση του μεγαλύτερου μέρους της πληροφορίας των αρχικών δεδομένων. Από τον τύπο είναι εμφανές ότι η διατήρηση της πληροφορίας δεν εξαρτάται από τον αριθμό των τελικών διαστάσεων αλλά από το πλήθος των αρχικών δειγμάτων.

Σε σχέση με την PCA μπορεί να χρησιμοποιηθεί για παράλληλη επεξεργασία και διατηρεί καλύτερα τις αποστάσεις ζευγών μεταξύ όμοιων δειγμάτων, ειδικά όσον αφορά τα έκκεντρα δεδομένα, δηλαδή αυτά που τα σημεία τους στο χώρο σχηματίζουν κυκλικές ή σχεδόν κυκλικές ομάδες.

Υποστηρίζεται ότι μερικές λιγότερο εύστοχες προβολές σε χώρο μειωμένης διάστασης δεν επηρεάζουν δραματικά την αποτελεσματικότητα ενός μοντέλου αν αυτό συνδυαστεί με ensemble learning μεθόδους και με απλή εφαρμογή ενός αλγόριθμου πλειοψηφίας.

Ο αλγόριθμος ακολουθεί τέσσερα βήματα με σκοπό την τελική πρόβλεψη της ταξινόμησης των αρχικών δεδομένων. Πρώτον παράγονται πολλαπλοί υποχώροι μέσω διαφορετικών τυχαίων προβολών. Ακολουθεί εκπαίδευση του κάθε ταξινομητή (KNN, LDA) για κάθε δείγμα σε όλους τους υποχώρους. Στη συνέχεια γίνεται πρόβλεψη για κάθε στοιχείο του δοκιμαστικού συνόλου δεδομένων και τέλος λαμβάνεται η τελική απόφαση για την πρόβλεψη της κλάσης των δειγμάτων μέσω πλειοψηφίας με βάση τα αποτελέσματα όλων των προηγούμενων ταξινομητών. Τα αποτελέσματα δείχνουν ότι ακόμη και για σταθερό αριθμό τελικών διαστάσεων το μοντέλο λειτουργεί εξίσου καλά έχοντας μεγάλη ακρίβεια.



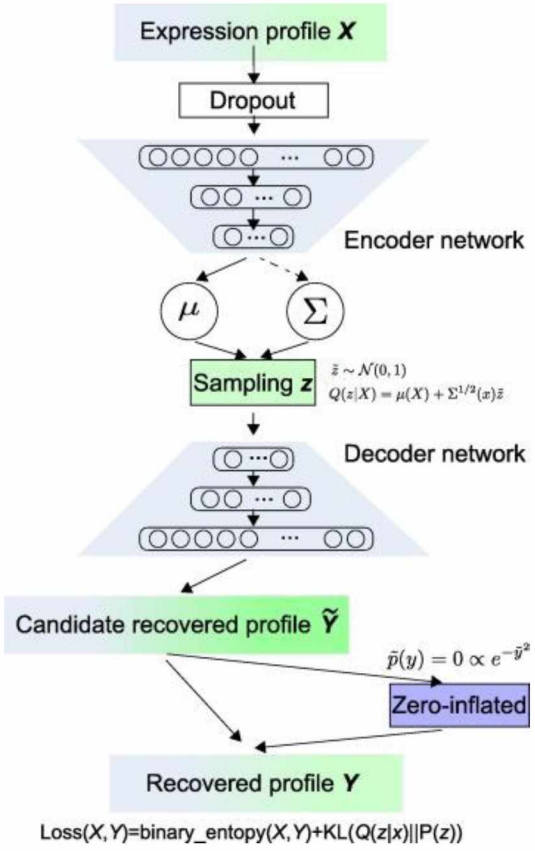
Εικόνα 26 Λεπτομερής δομή μεθόδου MPRV

Είναι εμφανές ότι η επιλογή ενός μεγάλου αριθμού τελικών διαστάσεων είναι μη πρακτικό για την βέλτιστη κατηγοριοποίηση ή και οπτικοποίηση των δεδομένων, καθώς αυξάνει την πολυπλοκότητα και το χρόνο της κατηγοριοποίησης μειώνοντας παράλληλα την ακρίβειά της. Σκοπός λοιπόν της μεθόδου που περιγράφεται είναι η αύξηση της απόδοσης των κλασικών κατηγοριοποιητών σε δεδομένα πολλών διαστάσεων και ιδιαίτερα σε βιολογικά scRNA-seq δεδομένα. Η μέθοδος αυτή φαίνεται να είναι εξαιρετικά αποδοτική ακόμη και για συγκεκριμένο αριθμό τελικών διαστάσεων, μικρότερο από εκείνο που υπολογίζεται από το λήμμα Johnson–Lindenstrauss.

4.3. Μέθοδος VASC

Μέσω της ανάλυσης των δεδομένων scRNA-seq sequencing γίνεται μελέτη της κυτταρικής ετερογένειας, αλλά και μελέτη πληθυσμών κυττάρων και γενεαλογιών. Το πρώτο βήμα είναι η αποδοτική οπτικοποίηση των δεδομένων αυτών και η μείωση της διάστασής τους. Στα βιολογικά αυτά δεδομένα είναι πολύ μεγαλύτερες οι διακυμάνσεις στην έκφραση των γονιδίων και ο μικρός αριθμός RNA μεταγραφωμάτων αυξάνει τα dropout events, με αποτέλεσμα να καθιστά αυτά τα σύνολα δεδομένων ιδιαίτερα θορυβώδη. Η μέθοδος VASC, (Wang and Gu, 2018), όπως περιγράφεται παρακάτω, χρησιμοποιείται αποδοτικά στα θορυβώδη αυτά βιολογικά, σύνολα δεδομένων.

Η μέθοδος ωVASC μοντελοποιεί τα μηδενικά ή dropout events ξεχωριστά και βρίσκει μη γραμμικές αναπαραστάσεις μικρότερης διάστασης των αρχικών δεδομένων. Λειτουργεί αποδοτικά και για τα δεδομένα που προέρχονται από διαφορετικές πηγές. Προσφέρει αποδοτικές αναπαραστάσεις για πολύ σπάνιους πληθυσμούς κυττάρων στις δύο διαστάσεις και την εύρεση υποψήφιων biomarkers ή βιοδεικτών. Όλες αυτές οι πληροφορίες δεν είναι δυνατό να βρεθούν σε ακατέργαστα δεδομένα χωρίς να γίνει απομόνωση γενετικών πληροφοριών από κάθε κύτταρο χωριστά.



Εικόνα 27 Περιγραφή δομής μεθόδου VASC

Το πρώτο βήμα της διαδικασίας είναι μείωση της διάστασης των αρχικών δεδομένων για περαιτέρω επεξεργασία. Η διαδικασία αυτή συνήθως γίνεται μέσω των μεθόδων PCA, t-tSNE και άλλων. Αν παραληφθεί το παραπάνω βήμα, τότε τα δεδομένα θα είναι πολύ θορυβώδη για να μπορέσουν να ταξινομηθούν και να επεξεργαστούν κατάλληλα. Άλλες μέθοδοι όπως η ZIFA, χρησιμοποιείται για να μοντελοποιήσει μόνο γραμμικά μοτίβα, για αυτό και η απόδοση της είναι πολύ περιορισμένη. Ακόμη δεν είναι ικανή να αντιμετωπίσει αποδοτικά τις τιμές που είναι σχεδόν μηδενικές. Στο συγκεκριμένο όμως κώδικα έχουν χρησιμοποιηθεί deep variational Autoencoders, με σκοπό την ανάλυση και οπτικοποίηση αυτών των δεδομένων. Ο λόγος που επιλέχθηκε αυτό το μοντέλο είναι γιατί μπορεί να διατηρήσει ένα πολύ μεγάλο μέρος της πληροφορίας διατηρώντας τοπολογικές πληροφορίες και εξάγοντας αποδοτικούς λανθάνοντες χώρους που αντιπροσωπεύουν τα αρχικά δεδομένα. Χρησιμοποιεί την gambel distribution για να μοντελοποιήσει ξεχωριστά τις μηδενικές ή σχεδόν μηδενικές τιμές.

Ο Autoencoder λειτουργεί αναζητώντας τις βέλτιστες παραμέτρους για να προβάλει τα πολυδιάστατα αυτά δεδομένα σε έναν λανθάνοντα χώρο μικρότερης διάστασης Z, από τον οποίο θα είναι δυνατή η ανακατασκευή του αρχικού πολυδιάστατου χώρου. Βασίζεται στην ιδέα ότι τα αρχικά δεδομένα βρίσκονται πάνω σε μία γνωστή κατανομή για παράδειγμα την κανονική, της οποίας τις παραμέτρους (μ, Σ) προσπαθεί να υπολογίσει. Όλες αυτές οι παράμετροι που υπολογίζονται με βάση τα αρχικά δεδομένα αποθηκεύονται σε μία μεταβλητή που ονομάζεται «latent variables» και σκοπός είναι η εστίαση και εξαγωγή των εσωτερικών πληροφοριών των δεδομένων ή intrinsic Information. Η κατανομή με τις παραμέτρους που υπολογίστηκε μετά την εκπαίδευση των δεδομένων χρησιμοποιείται με σκοπό την δειγματοληψία ψευδό-δειγμάτων X' από τον Decoder. Προκειμένου να υπολογιστεί η κατανομή κάθε τιμής z (latent variables) με βάση τα δεδομένα εισόδου X ή P(z|X) χρησιμοποιείται η variational κατανομή, ενώ η κατανομή (πχ κανονική), τις παραμέτρους της οποίας αναζητά το μοντέλο για να μειώσει τη διάσταση των αρχικών δεδομένων σε χώρο μικρότερης διάστασης, συμβολίζεται με Q(z|X). Σκοπός είναι η ελαχιστοποίηση του KL divergence, δηλαδή της απόστασης της κατανομής των δεδομένων που ανακατασκευάζονται με βάση τον λανθάνοντα χώρο Q(X) από εκείνη στην οποία πραγματικά ανήκουν τα αρχικά δεδομένα P(X).

Η διαδικασία αυτή αποτελείται τρία μέρη. Το πρώτο είναι αυτό του Encoder, που αποτελείται από τρία επίπεδα και παράγει τις παραμέτρους για την variational κατανομή. Το δεύτερο είναι εκείνο του Decoder, που επίσης αποτελείται από τρία επίπεδα, είναι συμμετρικός του Encoder και έχεις ως στόχο την ανακατασκευή των αρχικών δεδομένων από τον λανθάνοντα χώρο που δέχεται ως είσοδο. Τέλος

υπάρχει το ZI-layer ή Zero inflated επίπεδο το οποίο ανακτά τους γνωστούς κυτταρικούς τύπους, προσπαθώντας να μιμηθεί τα dropout events. Για το σκοπό αυτό θέτει τυχαία κάποιες τιμές στα δεδομένα, δηλαδή ορισμένα σημεία, σε μηδέν. Με τον τρόπο αυτό η μέθοδος μπορεί να χρησιμοποιηθεί για βιολογικά δεδομένα που προέρχονται από single-cell RNA sequencing, αφού λαμβάνει υπόψη τις ιδιαιτερότητές τους.

Όσον αφορά τη συνάρτηση σφάλματος, αυτή αποτελείται από δύο μέρη. Το πρώτο είναι το binary cross entropy loss και οφείλεται στο γεγονός ότι τα δεδομένα έχουν υποστεί κάποιας μορφής κλιμάκωση και το δεύτερο είναι το KL divergence loss, το οποίο οφείλεται στην απόσταση των προβλεπόμενων δεδομένων ή reconstructed data από τα αρχικά.

Στη συνέχεια για να ελεγχθεί το μοντέλο εφαρμόζεται μία τεχνική ομαδοποίησης με βάση την οποία οι δισδιάστατες αναπαραστάσεις του λανθάνοντος χώρου ομαδοποιούνται και συγκρίνονται με τις γνωστές υπάρχουσες κλάσεις. Ως k ορίζεται ο αριθμός των γνωστών κυτταρικών τύπων. Συνεπώς με αυτήν τη μέθοδο μπορούν να βρεθούν κάποιοι υποπληθυσμοί ή και άγνωστοι κυτταρικοί τύποι ενώ μπορεί να επαληθευτεί και η διαφοροποίηση των ήδη γνωστών. Όλα τα παραπάνω την καθιστούν την VASC ως μία μέθοδο ειδική για scRNA-Seq δεδομένα.

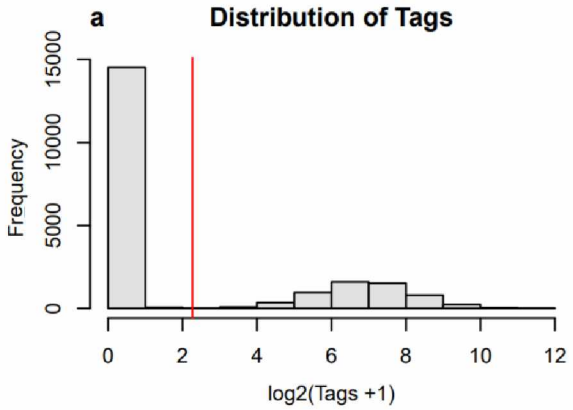
4.4. Μέθοδος CIDR

Η μέθοδος CIDR (Lin, Troup and Ho, 2017), , εξειδικεύεται στην διαχείριση των dropout events και είναι ένα υπολογιστικό εργαλείο βαριάς μορφής μοντελοποίησης δεδομένων. Είναι εξαιρετικά γρήγορο και χρησιμοποιεί για την εκπαίδευσή του πραγματικά αλλά και τεχνητά σύνολα δεδομένων. Είναι ιδιαίτερα χρήσιμο για την βελτιστοποίηση της απόδοσης ομαδοποίησης των δεδομένων.

Οι περισσότερες διαθέσιμες μέθοδοι που χρησιμοποιούνται για την επεξεργασία και ανάλυση των scRNA-Seq δεδομένων χρησιμοποιούν τεχνικές μείωσης διαστάσεων που δεν είναι εξειδικευμένες για αυτά, όπως η PCA και η t-SNE που παρουσιάζουν αδυναμίες στην αντιμετώπιση των Zero-inflated δεδομένων, δηλαδή δεδομένων με πολλές μηδενικές ή σχεδόν μηδενικές τιμές. Παράλληλα οι μέθοδοι που λαμβάνουν υπόψη τους τα ιδιαίτερα χαρακτηριστικά των scRNA-Seq δεδομένων είναι υπολογιστικά ασύμφορες, καθώς έχουν μεγάλη πολυπλοκότητα.

Για τους παραπάνω λόγους αναπτύχθηκε μια μέθοδος μείωσης διαστάσεων που βασίζεται στην PCA με σκοπό να είναι υπολογιστικά συμφέρουσα αλλά και εξειδικευμένη για το συγκεκριμένο είδος των δεδομένων. Η διαδικασία αποτελείται από πέντε στάδια με σκοπό την εύρεση ενός χώρου μειωμένης διάστασης των αρχικών δεδομένων που μπορεί να χρησιμοποιηθεί για περεταίρω ανάλυσή τους. Το πρώτο βήμα περιλαμβάνει την αναγνώριση των υποψήφιων dropout events. Στην συνέχεια γίνεται μια εκτίμηση της σχέσης ανάμεσα στο ποσοστό των υποψήφιων αυτών dropout και στα επίπεδα γονιδιακής έκφρασης του υπόλοιπου συνόλου δεδομένων. Σε επόμενο χρόνο υπολογίζεται η ανομοιότητα μεταξύ των προφίλ γονιδιακής έκφρασης για κάθε ζεύγος μεμονωμένων κυττάρων και κατασκευάζεται ένα πίνακας CIDR που ονομάζεται πίνακας ανομοιογένειας. Στον πίνακα αυτόν εφαρμόζεται στην συνέχεια PCoA, που είναι μια παραλλαγή της μεθόδου PCA για ποσοτικά ή διακριτά δεδομένα. Τέλος στον χώρο των χαρακτηριστικών που προέκυψε από τις πρώτες κύριες συνιστώσες, εφαρμόζονται μέθοδοι κατηγοριοποίησης, που επιβεβαιώνουν και την απόδοση και ορθότητα της μεθόδου.

Πιο αναλυτικά, οι διαθέσιμες ετικέτες έκφρασης γονιδίων κάθε κυττάρου ξεχωριστά, υφίστανται μια λογαριθμική μετατροπή. Η κατανομή που προκύπτει χαρακτηρίζεται από μια μεγάλη κορυφή στο μηδέν και άλλες μικρότερες τοπικές κορυφές σε άλλες θετικές μη μηδενικές τιμές, που αντιπροσωπεύουν εμφανίσεις του κάθε γονιδίου στο συγκεκριμένο κύτταρο. Θεωρούμε κατώφλι T_i, που υπολογίζεται ανάλογα με το δείγμα και διαχωρίζει τις μηδενικές τιμές από τις υπόλοιπες της κατανομής. Οι τιμές πριν από αυτό το κατώφλι είναι οι μηδενικές που είναι υποψήφιες ως dropouts και μπορεί να περιλαμβάνουν dropout events αλλά και πραγματικά μηδενικές εκφράσεις γονιδίων.

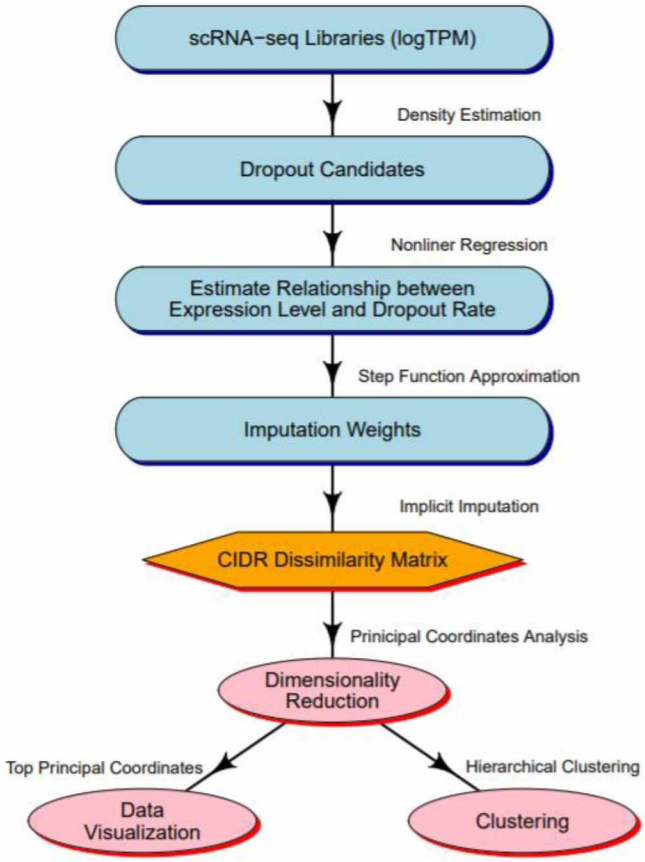


Εικόνα 28 Κατανομή ετικετών των δεδομένων μετά την λογαριθμική μετατροπή για εφαρμογή της μεθόδου CIDR

Έστω u η μη παρατηρηθείσα αληθινή έκφραση ενός γονιδίου σε ένα κύτταρο και P(u) η πιθανότητα να είναι dropout. Εμπειρικά φαίνεται η P(u) να είναι μια φθίνουσα συνάρτηση πιθανότητας, δηλαδή όσο πιο πολλές φορές φαίνεται να εκφράζεται το συγκεκριμένο γονίδιο, τόσο λιγότερο πιθανό είναι να αποτελεί dropout event και αντίστροφα. Για τον υπολογισμό αυτής της πιθανότητας χρησιμοποιείται ένα μη γραμμικό μοντέλο ελαχίστων τετραγώνων με σκοπό την προσαρμογή μιας φθίνουσας λογαριθμικής συνάρτησης στα δεδομένα. Αν χρησιμοποιηθεί ολόκληρο το σετ δεδομένων για τον υπολογισμό της πιθανότητας αυτής που συμβολίζεται ως P'(u) παρατηρείται συχνά ότι οι περισσότεροι υποψήφιοι είναι πράγματι dropout events. Είναι σημαντικό να επισημανθεί ότι αυτή η πληροφορία προέκυψε από συμπεράσματα που προκύπτουν από τα ίδια τα δεδομένα, μέσα από τις διαφορές στις εκφράσεις των γονιδίων και τη σύγκριση των κυττάρων.

Η P'(u) χρησιμοποιείται για τον υπολογισμό των τιμών του πίνακα ανομοιότητας CIDR. Οι dropout αντιμετωπίζονται ως χαμένες τιμές και ακολουθεί μια διαδικασία που ονομάζεται CIDR's pairwise implicit process. Περιληπτικά με βάση αυτή τη διαδικασία, που καταλογίζει τιμές στα δεδομένα που χάθηκαν, για κάθε ζεύγος του πίνακα ανομοιότητας, ελέγχεται η τιμή κατωφλίου T για κάθε κύτταρο. Υπάρχουν δύο περιπτώσεις εφόσον ο αλγόριθμος εφαρμόζεται μόνο στις τιμές που λείπουν και παραλείπει τις υπόλοιπες. Η πρώτη περίπτωση είναι να υπάρχει μια μόνο χαμένη τιμή στο ζεύγος που ελέγχεται και η δεύτερη να είναι και οι δύο τιμές του γονιδίου ελλιπείς. Για την πρώτη περίπτωση, χρησιμοποιείται μια απλοποιημένη μορφή W(u), που όμως δεν επηρεάζει την τελική αποτελεσματικότητα κατηγοριοποίησης, αντί για την P'(u) για τον υπολογισμό ενός σταθμισμένου μέσου όρου των δύο τιμών που θα χρησιμοποιηθεί για να συμπληρώσει τον πίνακα. Αν ισχύει η δεύτερη περίπτωση, τότε συμπληρώνεται και για τις δύο εγγραφές η τιμή μηδέν. Έτσι προκύπτει ο τελικός πίνακας CIDR στον οποίο εφαρμόζεται τελικά PCoA.

Τέλος, εφαρμόζεται μέθοδος ιεραρχικής ομαδοποίησης στις πρώτες λίγες πρωτεύουσες συνιστώσες που προκύπτουν και επιλέγεται ο αριθμός των ομάδων με βάση το Variance Ratio κριτήριο. Πρακτικά η μέθοδος που περιεγράφηκε διατηρεί τις αποστάσεις ανάμεσα στις ομάδες και μειώνει τις αποστάσεις των dropouts από τα σωστά υπολογισμένα γονίδια κάνοντας τις κλάσεις πιο ισχυρά συνδεδεμένες. Ολόκληρη η παραπάνω διαδικασία επιτυγχάνεται πολύ γρηγορότερα συγκριτικά με άλλες, ενώ έχει μεγάλη ακρίβεια στην ομαδοποίηση βιολογικών δεδομένων.



Εικόνα 29 Περίληψη βημάτων για εφαρμογή της μεθόδου CIDR

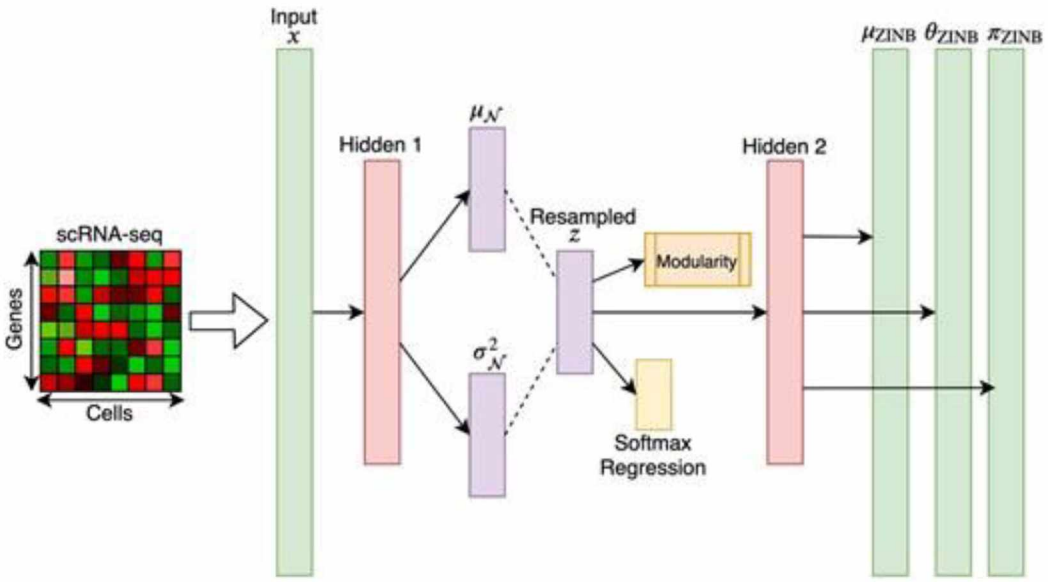
4.5. Μέθοδος netAE

Η αλληλούχιση RNA μεμονωμένου κυττάρου είναι μια πολλά υποσχόμενη μέθοδος που μπορεί να δώσει χρήσιμες πληροφορίες για την κυτταρική ετερογένεια και την αναγνώριση γνωστών ή νέων κυτταρικών τύπων και υποτύπων. Ακόμη βοηθά στον εντοπισμό μηχανισμών των οργανισμών σε επίπεδο κυττάρου, όπως την εύρεση διαφορετικών λειτουργιών των κυτταρικών τύπων ή διάφορες καταστάσεις και αλλαγές στην γονιδιακή τους έκφραση που σχετίζονται με τη φάση και τον κύκλο ζωής του κάθε κυττάρου. Η χρήση μηχανικής μάθησης για την ανάλυση των πολυδιάστατων αυτών δεδομένων μπορεί να οδηγήσει στην ανακάλυψη σημαντικών βιο-δεικτών.

Μέχρι σήμερα η διαδικασία αυτή γίνεται μέσω ομαδοποίησης και χειροκίνητης ανάθεσης ετικετών στις ομάδες που προκύπτουν. Εκτός από την μη πρακτικότητα της παραπάνω διαδικασίας, η κατηγοριοποίηση των κυττάρων δεν θεωρείται μια άκαμπτη διαδικασία που ολοκληρώνεται εύκολα, αλλά μια

διαδικασία που συνεχώς πρέπει να εξελίσσεται, καθώς προκύπτουν νέα δεδομένα, που απαιτούν τη διόρθωση και τον επανασχολιασμό των ήδη σχολιασμένων κυτταρικών τύπων, με βάση τα νέα δεδομένα ή τους νέους κυτταρικούς τύπους που προέκυψαν από κάποιο άλλο σετ δεδομένων. Για τη διαδικασία που περιγράφεται παρακάτω χρειάζεται μόνο ένα μικρό τμήμα των δεδομένων να έχουν σχολιαστεί και τα υπόλοιπα προβλέπονται από το μοντέλο.

Στο netAE μοντέλο που προτείνεται, (Dong and Alterovitz, 2021), δίνονται κάποια δεδομένα με ετικέτες, πάνω στα οποία το μοντέλο εκπαιδεύεται αρχικά και στη συνέχεια προσπαθεί να επεκτείνει αυτή τη γνώση για τα υπόλοιπα διαθέσιμα κύτταρα χωρίς ετικέτα. Αυτό λειτουργεί αποδοτικά ακόμη και για μικρό αριθμό δειγμάτων εκπαίδευσης με ετικέτα, διατηρώντας τις απαραίτητες πληροφορίες και χαρακτηριστικά των αρχικών δεδομένων. Πρόκειται για μια semi-supervised μέθοδο που μπορεί να χρησιμοποιηθεί σε βιολογικά δεδομένα που διαθέτουν λίγα μόνο δείγματα με ετικέτα / σχολιασμό. Είναι μια μέθοδος μείωσης διαστάσεων κατά την οποία ενισχύεται η πληροφορία από τα διαθέσιμα δεδομένα με ετικέτα του αρχικού χώρου δεδομένων ώστε να περάσει όσο το δυνατόν απαράλλαχτη στον χώρο μειωμένης διάστασης. Βασίζεται στους Autoencoders που σκοπός τους είναι να βρουν τις σημαντικές πληροφορίες των αρχικών δεδομένων και να δημιουργήσουν έναν χώρο μειωμένης διάστασης, με ουσιαστικά χαρακτηριστικά (features), από τον οποία θα είναι δυνατή η ανακατασκευή του αρχικού χώρου με σχετικά μεγάλη ακρίβεια. Σε αυτήν την μέθοδο στηρίζεται ο netAE προσθέτοντας κάποιες επιπλέον διεργασίες. Αρχικά απαιτείται η εύρεση ενός χώρου μειωμένης διάστασης στον οποίο η δομή των ομάδων να είναι "ισχυρή" και ως εκ τούτου τα κύτταρα ίδιου τύπου να απέχουν λιγότερο. Έτσι σχηματίζεται ένα δίκτυο στενά συνδεδεμένων κυττάρων, από το οποίο προήλθε και το όνομα αυτού του μοντέλου. Η δεύτερη επιπλέον διεργασία είναι η εξασφάλιση ότι στο νέο αυτό χώρο οι κλάσεις θα είναι εύκολα διαχωρίσιμες. Στο τέλος της διεργασίας χρησιμοποιείται ένας απλός ταξινομητής για την πρόβλεψη των κλάσεων. Με τη χρήση ίδιων μεθόδων ταξινόμησης, ο netAE ξεπερνά άλλες μεθόδους μείωσης διαστάσεων ειδικά όταν πρόκειται για δεδομένα που έχουν μικρό σύνολο εκπαίδευσης.



Εικόνα 30 Εσωτερική δομή μεθόδου netAE

Η συγκεκριμένη μέθοδος παρόλο που εστιάζει στην ενίσχυση της πληροφορίας που προέρχεται από τα δεδομένα με ετικέτα, δεν διαταράσσει τη συνολική δομή του αρχικού χώρου δεδομένων και φαίνεται να έχει και άλλες πρακτικές εφαρμογές εκτός από την ταξινόμηση κυττάρων.

Έτσι κατασκευάστηκε ένας variational Autoencoder που σε συνδυασμό με την μετρική συνδετικότητας ή modularity measure των Girvan και Newman παράγει έναν χώρο μειωμένης διάστασης, με ισχυρή δομή ομαδοποίησης (clusterization structure) και εύκολα διαχωρίσιμες κλάσεις. Το modularity measure που ενσωματώνεται στο νευρωνικό δίκτυο με τη μορφή σφάλματος, όπως θα αναλυθεί και στη συνέχεια, χρησιμοποιείται ως μέσο ποσοτικοποίησης της ισχύος της δομής των κλάσεων. Ο variational Autoencoder επιλέχθηκε καθώς βοηθά το μοντέλο να βρει μια κατανομή που ταιριάζει στα πολυδιάστατα αρχικά δεδομένα αλλά και σε άλλα παρόμοια με αυτά μέσα από κάποια μικρής διάστασης χαρακτηριστικά του λανθάνοντος χώρου. Ο τρόπος λειτουργίας variational Autoencoder δεν θα αναλυθεί περεταίρω καθώς έχει αναλυθεί ήδη προηγουμένως.

Όσον αφορά την modularity μετρική, χρησιμοποιείται για να υπολογιστεί η πυκνότητα ενός γράφου έναντι της πυκνότητας ενός άλλου τυχαίου ελεγχόμενου βαθμού. Πρακτικά, είναι μια μέθοδος υπολογισμού μια τιμής για κάθε ακμή ενός γράφου, που αντιπροσωπεύει την απόσταση ανάμεσα στους κόμβους του, ο οποίος στην προκειμένη περίπτωση σχηματίζεται αν ενωθούν όλα τα σημεία του χώρου με ακμές και κατά την διάρκεια αφαιρεθούν αυτές με τη μεγαλύτερη τιμή, έως ότου προοδευτικά παραμείνει ο επιθυμητός αριθμός ομάδων και προκύψει μη συνεκτικό γράφημα. Έτσι ο αλγόριθμος αυτός

εξασφαλίζει ότι η απόσταση ανάμεσα στα σημεία της ίδιας ομάδας είναι η μικρότερη δυνατή, αφού διατηρούνται οι ακμές με τη μικρότερη τιμή, ενώ η απόσταση ανάμεσα στις ομάδες είναι μέγιστη, αφού αφαιρούνται οι ακμές με τη μεγαλύτερη τιμή. Επίσης η μέγιστη τιμή ακμής για την οποία δύο σημεία θεωρείται ότι ανήκουν στην ίδια ομάδα μπορεί να καθοριστεί από το χρήστη. Σημαντικό είναι να αναφερθεί ότι η έννοια της συνδετικότητας (modularity) χρησιμοποιείται προκειμένου να βελτιστοποιηθεί η αναπαράσταση του λανθάνοντος χώρου, δεδομένου ενός σταθερού και γνωστού συνόλου ομάδων, σε αντίθεση με την τυπική χρήση της που είναι η βελτιστοποίηση των ομάδων χρησιμοποιώντας έναν σταθερό λανθάνοντα χώρο.

Πιο συγκεκριμένα ο netAE χρησιμοποιεί αρχικά το σύνολο των δεδομένων και το αποτυπώνει μέσω του encoder σε δύο ξεχωριστούς πίνακες μ(x), σ(x) όπου N(μ(x),σ(x)) είναι η κανονική κατανομή που υπολογίζεται από τα δεδομένα. Στην συνέχεια λαμβάνονται δείγματα από αυτήν την κατανομή με σκοπό την εύρεση ενός λανθάνοντα χώρου, μειωμένης διάστασης, από τον οποίο ο Decoder θα μπορέσει να ανακατασκευάσει τα αρχικά δεδομένα. Ο Decoder έχει ως έξοδο τρεις πίνακες, μ(x),θ(x),π(x), προκειμένου να κατασκευαστεί ένα μοντέλο παρόμοιο με το ZINB (Greene, 1994), το οποίο φαίνεται να διαχειρίζεται σωστά τις ιδιαιτερότητες των συγκεκριμένων βιολογικών δεδομένων, παράγοντας αποτελέσματα με μεγάλη ακρίβεια. Όμως ο netAE έχει δύο ακόμα απαιτήσεις για τον λανθάνοντα χώρο που απαιτούν μεθόδους επιβλεπόμενης μάθησης και ενσωματώνονται στο παραπάνω μη επιβλεπόμενο μοντέλο. Αυτές μπορούν να ενσωματωθούν στο μοντέλο με τη μορφή όρων στην συνάρτηση σφάλματος, ένας όρος που ονομάζεται modularity loss και ένας που ονομάζεται classification loss. Αν το modularity διατυπωθεί με συγκεκριμένο τρόπο, όπως παρουσιάζεται στην εργασία που περιγράφεται, είναι διαφοροποιήσιμο, επομένως μπορεί να χρησιμοποιηθεί εύκολα ως όρος της συνάρτησης σφάλματος που εκπαιδεύει το μοντέλο. Επιπλέον, αν προστεθεί και ένα επίπεδο softmax μετά τον κωδικοποιητή και υπολογιστεί η αρνητική πιθανότητα καταγραφής του ταξινομητή softmax, μπορεί επίσης να χρησιμοποιηθεί ως όρος συνάρτησης σφάλματος, αφού εξασφαλίζει ότι ο λανθάνων χώρος είναι εύκολα διαχωρίσιμος. Η συνάρτηση κόστους του netAE φαίνεται παρακάτω:

$$L_{unsupervised} = -\mathbb{E}_{z\sim q\delta(z|x)}[log\ ZINB(x \mid\mid \mu(z), \theta(z), \pi(z))] + \kappa \mathbb{KL}(q_\delta(z|x)||p(z))$$

$$L_{netAE} = L_{unsupervised} - \lambda \mathbb{E}_{z\sim q\delta(z|x)}[Q'(z,\gamma)] - \phi \mathbb{E}_{z\sim q\delta(z|x)}[log\ f(y \mid\mid z)]$$

5. Παρουσίαση της μεθόδου κατηγοριοποίησης συλλογικής μάθησης με τη χρήση Autoencoder (scVEC)

5.1. Εισαγωγή

Η ανάγκη για μελέτη της ετερογένειας των δεδομένων από scRNA-seq και της κατανόησης του γονιδιώματος και της λειτουργίας της γονιδιακής έκφρασης των οργανισμών, οδήγησε στην αναζήτηση μεθόδων μηχανικής μάθησης που μπορούν να βρουν μοτίβα και να βοηθήσουν στην έρευνα. Όλοι οι παραπάνω τρόποι εξαγωγής χαρακτηριστικών και μείωσης διαστάσεων μπορούν να χρησιμοποιηθούν, άλλοτε εξάγοντας περισσότερο και άλλοτε λιγότερο ποιοτικά και ουσιαστικά συμπεράσματα. Το μεγαλύτερο όμως πρόβλημα των περισσότερων μεθόδων είναι ότι δεν μπορούν να διατηρήσουν τα μη γραμμικά χαρακτηριστικά που διαθέτουν τα δεδομένα, καθώς στους υποχώρους που προκύπτουν συνήθως αναπαριστώνται τα γραμμικά χαρακτηριστικά που διατηρήθηκαν, με αποτέλεσμα να χάνεται η μη γραμμική πληροφορία των αρχικών δεδομένων.

Ακόμη, τα ιδιαίτερα χαρακτηριστικά των βιολογικών δεδομένων από αλληλούχιση μεμονωμένου κυττάρου, όπως αναφέρθηκαν ήδη σε προηγούμενο κεφάλαιο, μειώνουν την αποτελεσματικότητα των διαθέσιμων μεθόδων. Τα δεδομένα αυτά είναι πολλών διαστάσεων και εξαιρετικά αραιά, λόγω των φαινομένων dropout. Επομένως, είναι σημαντικό για την εξαγωγή ουσιαστικής αναπαράστασης, η εύρεση αποτελεσματικότερων και πιο ειδικών μεθόδων για την μείωση τα διάστασης και την κατηγοριοποίησή των δεδομένων αυτών.

Οι Autoencoders, είναι ένα νευρωνικό δίκτυο που, όπως έχει αναφερθεί, έχει πολλές πρακτικές εφαρμογές, λόγω της συμπιεσμένης αναπαράστασης των αρχικών δεδομένων εισόδου στο εσωτερικό του. Αυτή η αναπαράσταση των δεδομένων έχει βρεθεί ότι μπορεί να συντηρήσει τόσο γραμμικές όσο και μη γραμμικές πληροφορίες και χαρακτηριστικά και εξάγει ουσιαστικές πληροφορίες που μπορούν να χρησιμοποιηθούν όχι μόνο για την οπτικοποίηση των δεδομένων αλλά και ως επιπλέον πληροφορίες που βοηθούν την έρευνα, καθώς προκύπτουν δεδομένα χαμηλότερης διάστασης που μπορούν να ερμηνευτούν.

Όσον αφορά την κατηγοριοποίηση των κυττάρων υπάρχουν επίσης αρκετές τεχνικές είτε επιβλεπόμενης είτε μη επιβλεπόμενης μάθησης. Χαρακτηριστικά παραδείγματα των προηγούμενων κατηγοριών είναι ο knn και ο k-means που έχουν εφαρμοστεί αρκετές φορές σε δεδομένα από scRNA-seq. Γενικά υπάρχουν πολύ συγκεκριμένες περιπτώσεις στις οποίες έχει γίνει προσπάθεια συνδυασμού

των Autoencoder με κάποια μέθοδο κατηγοριοποίησης για τέτοιου είδους δεδομένα, μερικές από τις οποίες έχουν αναλυθεί ήδη στο κεφάλαιο της βιβλιογραφικής ανασκόπησης.

Στο πλαίσιο αυτό παρουσιάζεται μια μεθοδολογία που συνδυάζει τις δύο παραπάνω μεθόδους και πιο συγκεκριμένα τους Autoencoders και τους συλλογικούς ταξινομητές ή Ensemble Classifiers με σκοπό την αποτελεσματική κατηγοριοποίηση των παραπάνω βιολογικών δεδομένων και ονομάζεται scVEC (Variational Autoencoder and Ensemble Classifiers Framework for scRNA-seq data Classification). Είναι σημαντικό να αναφερθεί ότι πρόσφατα δημοσιεύτηκε παρόμοια μέθοδος που επιβεβαιώνει την υψηλή αποτελεσματικότητα της παρούσας μεθόδου που αναπτύχθηκε. Η μέθοδος αυτή είναι η scIAE, που περιεγράφηκε νωρίτερα και έχει ως στόχο την κατηγοριοποίηση των βιολογικών scRNA-seq δεδομένων, με τη χρήση Autoencoder και μεθόδων συλλογικής μάθησης. Στην παρούσα εργασία επιλέχθηκε να χρησιμοποιηθεί ένα συγκεκριμένο είδος Autoencoder που ονομάζεται Variational, σε αντίθεση με την επιλογή stacked, denoising, sparse Autoencoders της αντίστοιχης παρόμοιας, προαναφερθείσας μεθόδου. Ο λόγος που επιλέχθηκε ο συγκεκριμένος, είναι ότι μέσω αυτού δίνεται η δυνατότητα στο νευρωνικό δίκτυο να μάθει ομαλές, συμπιεσμένης διάστασης αναπαραστάσεις των δεδομένων εισόδου. Αντίθετα, οι τυπικοί Autoencoders, χρειάζεται απλώς να μάθουν μια κωδικοποίηση που τους επιτρέπει να αναπαράγουν την είσοδο στην έξοδο, ενώ οι stacked, denoising, sparse Autoencoders (Meng, Ding *et al.*, 2018) χρησιμοποιούνται διότι οι στοίβες εκπαίδευσης μπορούν να διατηρήσουν τα πιο χρήσιμα χαρακτηριστικά σε κάθε επίπεδο, χωρίζοντας σημαντικά τμήματα πληροφορίας σε κάθε ένα από αυτά.

Το μοντέλο που εφαρμόζεται φαίνεται να έχει επιτυχία σε δεδομένα scRNA-seq, καθώς έχει τη δυνατότητα να βρει και να κωδικοποιήσει εσωτερικές πληροφορίες των δεδομένων (intrinsic Information), επιτυγχάνοντας πιο εύστοχη συμπιεσμένη αναπαράστασή τους και εύρεση περισσότερων μη γραμμικών πληροφοριών σε αυτά, γεγονός που επαληθεύεται και από την εφαρμογή της αντίστοιχης μεθόδου scIAE. Ο Variational Autoencoder στηρίζεται στην ιδέα ότι τα δεδομένα εισόδου έχουν κάποια κατανομή πιθανότητας, της οποίας τις παραμέτρους υπολογίζει. Έτσι επιλέχθηκε γιατί μπορεί να ανακαλύψει πληροφορίες για τα δεδομένα εισόδου αλλά και για άλλα δεδομένα όμοια με αυτά, χωρίς να έχει εκπαιδευτεί σε όλα, καθώς μαθαίνει εξάγει εσωτερικές πληροφορίες για τα δεδομένα με βάση την κατανομή στην οποία υποθέτει ότι ανήκουν.

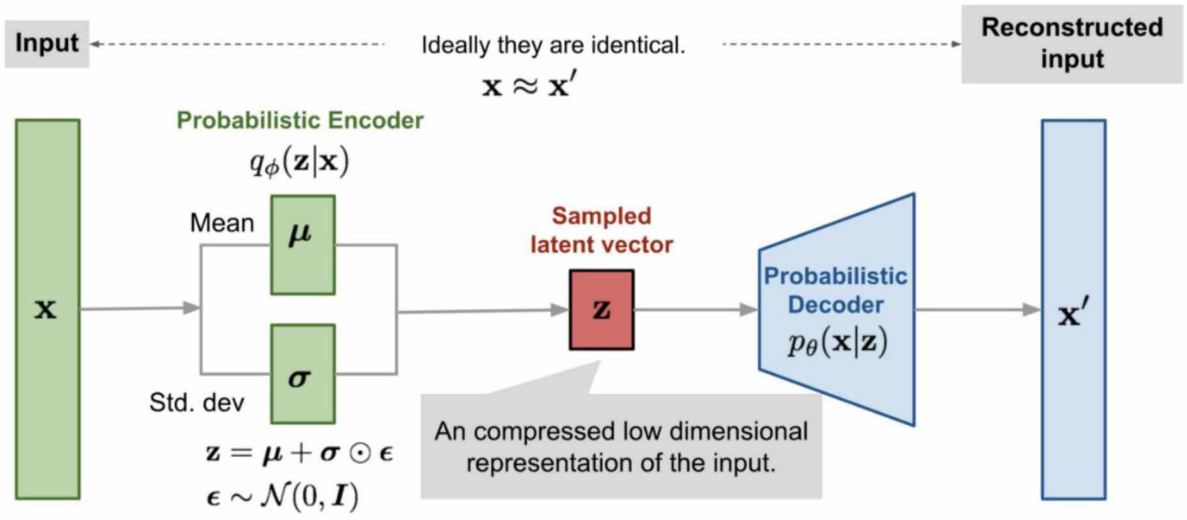
Πρίν γίνει εκτενής αναφορά στη μέθοδο που υλοποιήθηκε, παρουσιάζονται ξεχωριστά οι μέθοδοι που χρησιμοποιήθηκαν για την κατασκευή και τον έλεγχο της ορθότητας του μοντέλου, προκειμένου να

γίνει ευκολότερα κατανοητή η επιλογή τους και ο λόγος που περιλαμβάνονται στη μέθοδο που αναπτύχθηκε. Πιο συγκεκριμένα στο κεφάλαιο αυτό περιγράφεται η λειτουργία των Variational Autoencoder (Kingma and Welling, 2019) και ο λόγος που το νευρωνικό αυτό δίκτυο επιλέχθηκε σε σχέση με έναν απλό ή Vanilla Autoencoder (Hinton and Salakhutdinov, 2006). Στην συνέχεια γίνεται αναφορά στην χρήση της μεθόδου KNN και στον τρόπο λειτουργίας της. Ακόμη, παρουσιάζεται η μέθοδος k - Fold Cross Validation και η τεχνική της συλλογικής μάθησης ή Ensemble Learning.

5.1.1. Variational Autoencoders

Ο Variational (Kingma and Welling, 2019) είναι ένα είδος Autoencoder που χρησιμοποιείται για να ομαλοποιήσει τον λανθάνοντα χώρο. Πιο συγκεκριμένα, αντί για την επιστροφή σημείων ή δειγμάτων μειωμένης διάστασης το νευρωνικό δίκτυο επιστρέφει μια κατανομή. Με αυτόν τον τρόπο, ο λανθάνων χώρος γίνεται συνεχής, κάνοντας ευκολότερη την τυχαία δειγματοληψία στα δεδομένα. Ακόμα είναι ευκολότερη η εξαγωγή εσωτερικών ή εγγενών πληροφοριών (intrinsic information) από αυτά, δηλαδή πληροφοριών που εξάγονται από κάθε ένα από τα δεδομένα και όχι από την συσχέτισή αυτού με τα υπόλοιπα. Αυτή είναι και η διαφορά του συγκεκριμένου τύπου Autoencoder σε σχέση με τον απλό ή Vanilla, δηλαδή η διαφορά στην κατασκευή του λανθάνοντα χώρου που τελικά οδηγεί στην εύρεση σημαντικότερων μοτίβων.

Για να επιτευχθεί το παραπάνω, το δίκτυο του κωδικοποιητή έχει ως έξοδο δύο πίνακες ίδιου μεγέθους (μειωμένης διάστασης). Ο πρώτος περιγράφει την μέση τιμή «μ» και ο δεύτερος την τυπική απόκλιση «σ» μιας κανονικής κατανομής. Το μ επηρεάζει το κέντρο του σημείου μειωμένης διάστασης στον λανθάνοντα χώρο και το σ την απόσταση από το κέντρο, δηλαδή την περιοχή γύρω από το κέντρο, που είναι πιθανό να βρίσκεται ένα δείγμα. Από αυτήν την κατανομή στην συνέχεια γίνεται δειγματοληψία με σκοπό την εξαγωγή ενός πίνακα σημείων μειωμένης διάστασης που τελικά θα αποτελέσουν τον λανθάνοντα χώρο. Αυτό σημαίνει ότι ακόμη και με την ίδια είσοδο, όσο οι παράμετροι μ και σ παραμένουν ίδιες, η πραγματική έξοδος του του κωδικοποιητή διαφέρει κάθε φορά απλά και μόνο λόγω της δειγματοληψίας (*Intuitively Understanding Variational Autoencoders | by Irhum Shafkat | Towards Data Science*, Feb. 2018). Έτσι, το νευρωνικό δίκτυο επιτυγχάνει καλύτερη γενίκευση (generalization), αφού το δίκτυο δεν μαθαίνει μόνο ότι το συγκεκριμένο δείγμα ανήκει σε μια κλάση, αλλά και ότι όλα τα υπόλοιπα σημεία με τις ίδιες παραμέτρους της κατανομής (που βρίσκονται μέσα στον ίδιο κυκλικό χώρο) ανήκουν σε αυτήν.



Εικόνα 31 Δομή του Variational Autoencoder

5.1.1.1. Περιγραφή τρόπου λειτουργίας και δομής των VAE

Όπως ήδη αναφέρθηκε, σκοπός των Variational Autoencoders είναι να κανονικοποιήσουν τον λανθάνοντα χώρο όσο το δυνατόν περισσότερο, να αποτρέψουν την υπερπροσαρμογή και να εξασφαλίσουν ότι τα δεδομένα μειωμένης διάστασης έχουν τις βέλτιστες δυνατές αναπαραστάσεις και βοηθούν την εξαγωγή ουσιαστικών συμπερασμάτων (Intuitively Understanding Variational Autoencoders | by Irhum Shafkat | Towards Data Science n.d.).

Όπως και ένας απλός Autoencoder έτσι και ο VAE αποτελείται από ένα επίπεδο κωδικοποίησης και ένα επίπεδο αποκωδικοποίησης και σκοπός είναι η ανακατασκευή των αρχικών δεδομένων εισόδου στην έξοδο. Η διαφορά έγκειται στο γεγονός ότι η έξοδος του κωδικοποιητή για κάθε ένα από τα δείγματα εισόδου, είναι μια κατανομή και όχι ένα σημείο. Αρχικά, η είσοδος συμπιέζεται και κωδικοποιείται με τη μορφή παραμέτρων μιας κατανομής στον λανθάνοντα χώρο. Στην συνέχεια, από αυτόν το χώρο, γίνεται δειγματοληψία από κάθε κατανομή που ορίζεται από κάθε ένα σημείο του. Το νέο σημείο δειγματοληψίας αποκρυπτογραφείται από τον decoder, ενώ ταυτόχρονα υπολογίζεται το σφάλμα ανακατασκευής του. Τελικά, η τιμή του σφάλματος χρησιμοποιείται για οπισθοδιάδωση δια μέσου του νευρωνικού δικτύου. Μαθηματικά αυτή η διαδικασία μπορεί να μοντελοποιηθεί ως εξής:

$$input(\boldsymbol{x}) \rightarrow latent\ distribution\ \boldsymbol{p}(\boldsymbol{z}|\boldsymbol{x}) \rightarrow sampled\ latent\ representation\ \boldsymbol{z} \sim \boldsymbol{p}(\boldsymbol{z}|\boldsymbol{x})$$
$$\rightarrow input\ reconstruction\ \boldsymbol{d}(\boldsymbol{z})$$

Αντίστοιχα, για έναν Vanilla AE:

$$input(\boldsymbol{x}) \rightarrow latent\ representation\ \boldsymbol{z} = \boldsymbol{e}(\boldsymbol{x}) \rightarrow input\ reconstruction\ \boldsymbol{d}(\boldsymbol{z})$$

5.1.1.2. Υπολογισμός συνάρτησης απώλειας σε VAE

Ιδανικά για βέλτιστο διαχωρισμό ανάμεσα στις κλάσεις χρειάζεται αυτές αρχικά να επικαλύπτονται μεταξύ τους, ώστε το δίκτυο, σε επόμενο βήμα, να μάθει να τις διαχωρίζει καλύτερα. Διαφορετικά υπάρχει μεγάλη πιθανότητα το νευρωνικό δίκτυο να κάνει «overfit», δηλαδή να μάθει να διαχωρίζει κάθε κλάση ξεχωριστά, ώστε για κάθε ένα από τα δεδομένα εκπαίδευσης να επιτυγχάνεται άριστη ακρίβεια, όμως στα δεδομένα ελέγχου η ακρίβεια είναι ελάχιστη. Αυτό συμβαίνει επειδή οι Autoencoders στην πραγματικότητα εκπαιδεύονται στην κρυπτογράφηση και αποκρυπτογράφηση των δεδομένων εισόδου προσπαθώντας να ελαχιστοποιήσουν μόνο την συνάρτηση κόστους, χωρίς να έχει σημασία πώς είναι οργανωμένος ο λανθάνων χώρος (*Understanding Variational Autoencoders (VAEs) | by Joseph Rocca | Towards Data Science*, Sep. 2019).

Για να πετύχουμε την ιδανική περίπτωση, χρειάζεται μια τροποποίηση στον υπολογισμό της συνάρτησης απώλειας που χρησιμοποιείται. Αρχικά ο όρος KL divergence πρέπει να ελαχιστοποιηθεί ώστε να μειωθούν οι αποστάσεις ανάμεσα στις διαφορετικές κατανομές, αφού από αυτόν εξαρτάται η συγκέντρωση όλων των συμπιεσμένων δειγμάτων στο κέντρο του λανθάνοντα χώρου. Αυτό όμως από μόνο του οδηγεί σε αύξηση του θορύβου και ο αποκωδικοποιητής (decoder) αδυνατεί να ανακατασκευάσει τα δεδομένα. Για αυτό είναι απαραίτητη η προσθήκη του όρου της απώλειας ανακατασκευής ή reconstruction loss, που ενισχύει το βέλτιστο διαχωρισμό των κλάσεων. Άρα ισχύει ότι:

$$\text{Final}\_Loss = KL + ReconstructLoss.$$

Όμως, όταν το reconstruction loss χρησιμοποιηθεί για οπισθοδιάδωση, επειδή η διαδικασία δειγματοληψίας από μια κατανομή είναι στοχαστική και όχι ντετερμινιστική, δηλαδή παρόλο που τα μ και σ είναι σταθερά δεν προκύπτει πάντα η ίδια έξοδος, έχει χρησιμοποιηθεί το reparameterization trick ή τέχνασμα παραμετροποίησης (Jang, 2016; Jang, Gu and Poole, 2017) Με το τέχνασμα αυτό οι παράμετροι μ και σ συνδυάζονται σε μια ντετερμινιστική έκφραση με τον τυπικό κανονικό θόρυβο,

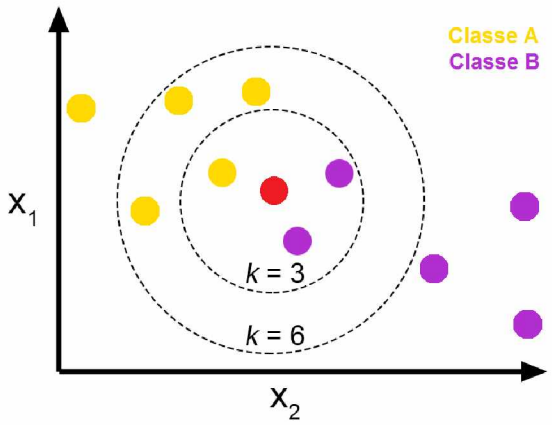
μέσω της οποίας είναι δυνατό να γίνει οπισθοδιάδωση. Πλέον ισχύει η έκφραση N(μ, σ) == N(0, 1) * σ + μ όπου το N(0,1) ορίζεται ως μια μεταβλητή e η οποία είναι γνωστή και σταθερή και μπορεί να βελτιστοποιηθεί μέσω του reconstruction loss. Στον κώδικα η αντίστοιχη συνάρτηση έχει αναπαρασταθεί ως εξής:

```
def reparameterize (self, mu, logvar):
std = logvar.mul (0.5). exp_ ()
eps = Variable (std.data.new (std. size ()). normal_ ())
return eps.mul(std). add_(mu)
```

Όπου το mu είναι ο μέσος όρος μ, logvar η λογαριθμική διακύμανση που χρησιμοποιείται για να εξασφαλιστεί ότι η τυπική απόκλιση θα είναι θετικός αριθμός, std η τυπική απόκλιση σ και eps το e όπως έχει αναφερθεί νωρίτερα.

5.1.2. Αλγόριθμος k Κοντινότερων Γειτόνων (KNN)

Μια τεχνική ταξινόμησης που χρησιμοποιείται ευρέως και βασίζεται σε δεδομένα αποστάσεων μεταξύ σημείων, είναι ο αλγόρθμος των K κοντινότερων γειτόνων ή KNN. Βασική προϋπόθεση για την χρήση αυτής είναι η ύπαρξη όλων των ετικετών των δειγμάτων εκπαίδευσης, άρα θεωρείται τεχνική επιβλεπόμενης μάθησης. Ο KNN ονομάζεται επίσης και «οκνηρός» αλγόριθμος μάθησης επειδή δεν μαθαίνει από το σύνολο εκπαίδευσης, αλλά αποθηκεύει το σύνολο δεδομένων και τη στιγμή της ταξινόμησης, απλώς εκτελεί μια ενέργεια σε αυτό (*K-Nearest Neighbor(KNN) Algorithm for Machine Learning - Javatpoint*, no date).



Εικόνα 32 Στιγμιότυπο αλγόριθμου KNN

Ο αλγόριθμος λειτουργεί έχοντας αρχικά υπολογίσει τις αποστάσεις των κλάσεων για τα δεδομένα εκπαίδευσης και όταν εμφανιστεί ένα νέο στοιχείο ελέγχου τότε υπολογίζει την απόστασή αυτού από τα δεδομένα εκπαίδευσης για να αποφασίσει την κλάση του. Αφού υπολογιστούν όλες οι αποστάσεις του νέου σημείου από τα υπόλοιπα του συνόλου εκπαίδευσης, λαμβάνονται υπόψη μόνο οι αποστάσεις των K κοντινότερων για την πρόβλεψη της κλάσης του νέου στοιχείου. Ιδιαίτερη σημασία για την μέθοδο αυτή έχει η επιλογή του κατάλληλου K. Συνήθως για την επιλογή του ισχύει ότι:

$$K = \sqrt{N}$$

Όπου N το πλήθος των δεδομένων εκπαίδευσης.

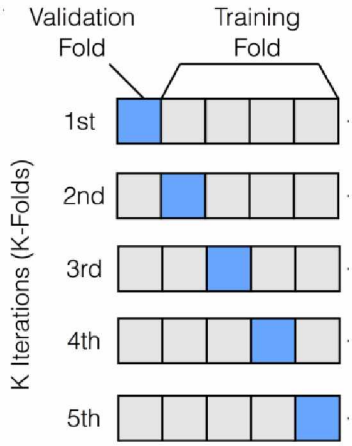
Πιο συγκεκριμένα, αφού επιλεχθεί ο αριθμός των K γειτόνων, υπολογίζεται η απόσταση κάθε σημείου από τα υπόλοιπα (Ευκλείδεια, Manhattan…). Στην συνέχεια επιλέγονται οι K κοντινότεροι και υπολογίζεται πόσα από τα κοντινότερα σημεία ανήκουν σε κάθε κλάση. Στο τέλος σε κάθε σημείο ανατίθεται η κλάση με τον μέγιστο αριθμό γειτόνων που ανήκουν σε αυτή.

Ο KNN είναι εύκολος στη χρήση αλγόριθμος και έχει καλή γενίκευση. Μπορεί να επιτύχει μεγάλη ακρίβεια ακόμη και σε δεδομένα με μεγάλο θόρυβο και η απόδοσή του βελτιώνεται όσο μεγαλύτερο είναι το σύνολο εκπαίδευσης. Όμως, η επιλογή του K μπορεί να επηρεάσει σημαντικά την απόδοσή του αλγόριθμου. Άλλο ένα αρνητικό του είναι το υψηλό υπολογιστικό κόστος, λόγω της ανάγκης υπολογισμού όλων των αποστάσεων ανάμεσα στα δεδομένα / σημεία. Αυτός είναι και ο λόγος που ο KNN δεν είναι αποδοτικός όταν τα δεδομένα είναι πολυδιάστατα, γιατί εκτός από το υπολογιστικό κόστος επηρεάζεται δραματικά και η ακρίβειά του. Συνεπώς για να εφαρμοστεί σε πολυδιάστατα δεδομένα χρειάζεται να εφαρμοστεί σε αυτά νωρίτερα κάποια τεχνική μείωσης διαστάσεων ή να συνδυαστεί με κάποια άλλη μέθοδο ή συνδυασμό μεθόδων (Ensemble Learning, Ensemble Classifiers).

5.1.3. Μέθοδος διασταυρωμένης επικύρωσης (Cross Validation)

Στη μέθοδο k–fold Cross Validation το σύνολο των δεδομένων ανακατανέμεται τυχαία και διαχωρίζεται σε k ομάδες. Για κάθε μια από αυτές χρησιμοποιούνται οι υπόλοιπες ως δεδομένα εκπαίδευσης. Ένα μοντέλο εκπαιδεύεται στα k-1 δεδομένα εκπαίδευσης και ελέγχεται στα δεδομένα «hold-out». Έτσι κάθε φορά υπολογίζεται μια βαθμολογία από την διαίρεση των σωστά

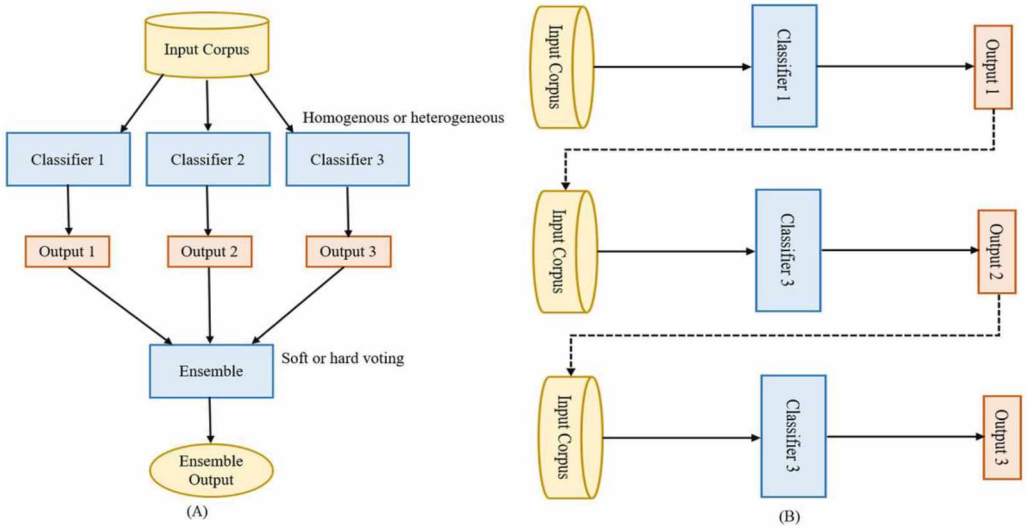
κατηγοριοποιημένων δειγμάτων με το σύνολό τους. Τέλος χρησιμοποιώντας αυτήν είναι δυνατός ο υπολογισμός της ακρίβειας του συγκεκριμένου μοντέλου, δεδομένων των παραμέτρων που επιλέχθηκαν. Γενικά όλα τα δείγματα ελέγχονται μια φορά όταν ανήκουν στην hold-out ομάδα και χρησιμοποιούνται ως δεδομένα εκπαίδευσης k-1 φορές .



Εικόνα 33 Μέθοδος διασταυρωμένης επικύρωσης δεδομένων

5.1.4. Συλλογική Μάθηση (Ensemble Learning)

Η συλλογική μάθηση (M.A. Ganaie, Minghui Hu *et al.*, 2021) είναι τεχνική μηχανικής μάθησης που συνδυάζει πολλά βασικά μοντέλα προκειμένου να παραχθεί ένα βέλτιστο μοντέλο πρόβλεψης, στο οποίο μεγιστοποιείται η σταθερότητα των αποτελεσμάτων, ελαττώνεται η πιθανότητα υπερπροσαρμογής στα δεδομένα εκπαίδευσης και αυξάνεται η ισχύς της ακρίβειας πρόβλεψης (*Stacking Ensemble Machine Learning With Python*, Dec. 2018). Υπάρχουν τρεις μέθοδοι συλλογικής μάθησης. Η μέθοδος που ελαχιστοποιεί την διακύμανση είναι η bagging, για την ελαχιστοποίηση της προκατάληψης (bias) χρησιμοποιείται η μέθοδος boosting και για την βελτιστοποίηση των προβλέψεων η stacking.



Εικόνα 34 Μέθοδος Ensemble Learning ή Συλλογικής μάθησης

Στην πραγματικότητα, αντί να εστιάσει κάποιος στην ανάπτυξη ενός περίπλοκου μοντέλου, με υψηλές απαιτήσεις χρόνου ή πολυπλοκότητας, συνδυάζει ήδη υπάρχουσες απλές και αποτελεσματικές μεθόδους μηχανικής μάθησης. Κάθε μέθοδος έχει διαφορετικά πλεονεκτήματα και σκοπός είναι ο συνδυασμός τους με στόχο την πλήρη εκμετάλλευση της αποδοτικότητάς τους, για εξαγωγή βέλτιστων αποτελεσμάτων. Έρευνες έχουν αποδείξει ότι η τελική ακρίβεια των μοντέλων συλλογικής μάθησης υπερισχύει σε σχέση με άλλες μεθόδους ενεργητικής μάθησης.

5.2. Μεθοδολογία και υλοποίηση μεθόδου scVEC

Σκοπός της μεθόδου scVEC που αναπτύχθηκε στα πλαίσια της παρούσας εργασίας είναι η μείωση της διάστασης πολυδιάστατων δεδομένων από scRNA-seq και η εξόρυξη πληροφορίας και χαρακτηριστικών, μέσω της ταξινόμησης των κυττάρων, με βάση τα γονίδια που εκφράζονται σε αυτά. Η μέθοδος αυτή μπορεί να οδηγήσει στην εύρεση νέων και την ταξινόμηση ήδη γνωστών κυτταρικών τύπων, να ενισχύσει τη μελέτη της κυτταρικής ετερογένειας, αλλά θέτει και τις βάσεις για τη χρήση μεθόδων συλλογικής μάθησης σε βιολογικά πολυδιάστατα δεδομένα. Ειδικά για το συγκεκριμένο τύπο δεδομένων, η ενίσχυση των υπαρχόντων μεθόδων εξαγωγής χαρακτηριστικών κρίνεται αναγκαία, καθώς η απόδοση των μεμονωμένων τεχνικών φαίνεται να παρουσιάζει μειωμένη ακρίβεια. Ακόμη αν συνδυαστεί με πιο εξειδικευμένες τεχνικές που βασίζονται στις ιδιαιτερότητες αυτού του είδους των δεδομένων όπως η ειδική διαχείριση των μηδενικών τιμών, μπορεί να βελτιωθεί σημαντικά η ακρίβεια της κατηγοριοποίησης και της εξαγωγής χαρακτηριστικών. Στο μοντέλο αυτό γίνεται μια προσπάθεια διαχείρισης των μηδενικών τιμών με την προσθήκη ενός επιπέδου τυχαίων μηδενισμών εκφράσεων

γονιδίων στα δεδομένα εισόδου (dropout layer) πρίν το επίπεδο του encoder, όπως έχει ήδη υποδειχθεί με τη μέθοδο VASC που περιεγράφηκε σε προηγούμενο κεφάλαιο (Wang and Gu, 2018). Τα δεδομένα που μπορεί να διαχειριστεί το μοντέλο μπορεί να είναι σε μορφή «.csv» ή «.mat», καθώς έχουν αναπτυχθεί μέθοδοι διαχείρισης και των δύο ΄τύπων δεδομένων. Χρειάζεται απλά η επιλογή της κατάλληλης μεθόδου στην κλάση «Databuilder» όπως φαίνεται αργότερα στον κώδικα στο αντίστοιχο παράρτημα.

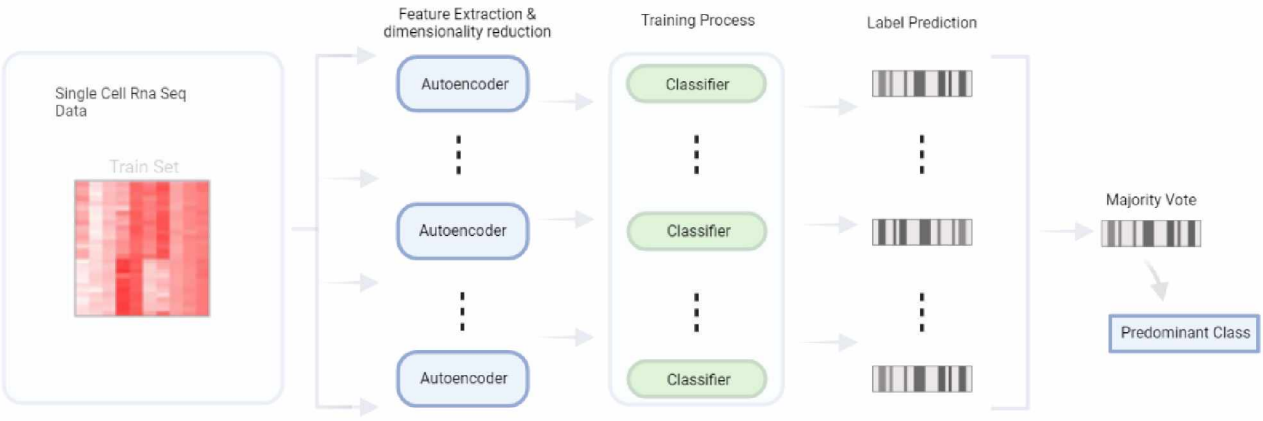
Η μείωση της διάστασης των πολυδιάστατων βιολογικών δεδομένων γίνεται με την χρήση Variational Autoencoder, για τους λόγους που έχουν ήδη αναφερθεί. Κάθε φορά η έξοδος που προκύπτει από το νευρωνικό δίκτυο ενός Autoencoder είναι διαφορετική, καθώς το δίκτυο μαθαίνει και διαφορετική συμπιεσμένη αναπαράσταση των αρχικών δεδομένων. Για τον σκοπό αυτό έγινε χρήση 'x' διαφορετικών Variational Autoencoder με σκοπό την εξαγωγή 'x' διαφορετικών χώρων μειωμένης διάστασης. Ο αριθμός τους μπορεί να επιλεχθεί από το χρήστη.

Σε κάθε έναν από αυτούς τους χώρους εφαρμόστηκε ο αλγόριθμος ταξινόμησης KNN ή k-nearest neighbor μέσω μιας μεθόδου 10-fold cross validation. Σκοπός της μεθόδου αυτής είναι η εφαρμογή του KNN σε κάθε έναν από τους δέκα επιμέρους υποχώρους μειωμένης διάστασης που προκύπτουν, έχοντας χρησιμοποιήσει ως σύνολο εκπαίδευσης την γνώση από τους υπόλοιπους εννέα και έχοντας ελέγξει κάθε δείγμα του συνόλου δεδομένων. Έτσι εξασφαλίζεται ότι κάθε παρατήρηση από το αρχικό σύνολο δεδομένων έχει την ευκαιρία να εμφανιστεί τόσο στο σύνολο εκπαίδευσης όσο και στο σύνολο δεδομένων δοκιμής (*Why and how to Cross Validate a Model? | by Sanjay.M | Towards Data Science*, Nov. 2018), (*A Gentle Introduction to k-fold Cross-Validation*, May 2018). Με αυτόν τον τρόπο η πρόβλεψη για κάθε ένα από τα δείγματα είναι όσο το δυνατόν πιο αντικειμενική και περιορίζεται η πιθανότητα να προκύψει το πρόβλημα επιλογής ενός «προβληματικού συνόλου εκπαίδευσης» αυξάνοντας την ακρίβεια της ταξινόμησης. Για την εφαρμογή της μεθόδου χρησιμοποιήθηκε η μέθοδος της βιβλιοθήκης sklearn, που ανήκει στις βιβλιοθήκες μηχανικής μάθησης της Python. Συνήθως αυτή η μέθοδος χρησιμοποιείται για την εκτίμηση της απόδοσης ενός μοντέλου όταν διατίθεται μικρός αριθμός δειγμάτων ή δεδομένων εκπαίδευσης, ή για την επιλογή ενός βέλτιστου συνδυασμού παραμέτρων (parameter tuning).

Όλες οι προβλέψεις ταξινόμησης κάθε κυττάρου / δείγματος αποθηκεύονται. Μετά το πέρας των 'x' επαναλήψεων και αφού έχουν αποθηκευτεί οι προβλέψεις των μειωμένης διάστασης δεδομένων, από όλους τους Autoencoders, εφαρμόζεται μέθοδος συλλογικής μάθησης (Ensemble Learning)

(B.Meshram and M. Shinde, 2015). Η ensemble μέθοδος αποτελείται από πολλούς Variational Autoencoders σε καθέναν από τους οποίους στην συνέχεια, εφαρμόζονται δέκα διαφορετικοί ταξινομητές, αφού το κάθε σύνολο δεδομένων εισόδου, χωρίζεται σε δέκα μικρότερα και ισόποσα (με ίδιο πλήθος δειγμάτων). Στο τέλος, έχοντας την παραπάνω πληροφορία από τις προβλέψεις της κάθε επανάληψης, το μοντέλο χρησιμοποιεί πλειοψηφική μέθοδο (majority) για την οριστική επιλογή της κλάσης που ανήκει κάθε ένα από τα παραπάνω δείγματα.

Με αυτόν τον τρόπο δημιουργείται ένα μοντέλο κατηγοριοποίησης που συνδυάζει νευρωνικά δίκτυα κρυπτογράφησης – αποκρυπτογράφησης που ο λανθάνων, συμπιεσμένος χώρος τους χρησιμοποιείται από ένα σύνολο κατηγοριοποιητών, για πρόβλεψη της κλάσης στην οποία ανήκει κάθε δείγμα. Πιο συγκεκριμένα με αυτόν τον τρόπο, όλες οι προβλέψεις (από κάθε βασικό κατηγοριοποιητή και για κάθε λανθάνοντα χώρο) εξάγονται από ένα «συλλογικό μοντέλο» που παίρνει την τελική απόφαση για την πρόβλεψη της κλάσης κάθε δείγματος πλειοψηφικά, επιλέγοντας αυτή που προτάθηκε από τους περισσότερους λανθάνοντες χώρους ή συμπιεσμένες αναπαραστάσεις. Η ακρίβεια της μεθόδου καθορίζεται από την σύγκριση της πρόβλεψης της μειωμένης αναπαράστασης με την αρχική κλάση στην οποία πραγματικά ανήκουν τα δεδομένα.



Εικόνα 35 Αναπαράσταση της μεθόδου scVEC

Η παραπάνω μέθοδος ανήκει στην κατηγορία bagging, καθώς χρησιμοποιούνται μοντέλα ίδιου τύπου, στα οποία το ίδιο σύνολο δεδομένων χρησιμοποιείται πολλές φορές. Από κάθε ένα μοντέλο προκύπτει και μια διαφορετική πρόβλεψη και στο τέλος όλες οι προβλέψεις συλλέγονται και αποθηκεύονται. Για

την επιλογή της τελικής κλάσης, που ανήκει το κάθε κύτταρο ή δείγμα, εφαρμόζεται μέθοδος voting μέσω πλειοψηφίας (majority vote).

5.3. Αποτελέσματα και Συμπεράσματα της μεθόδου

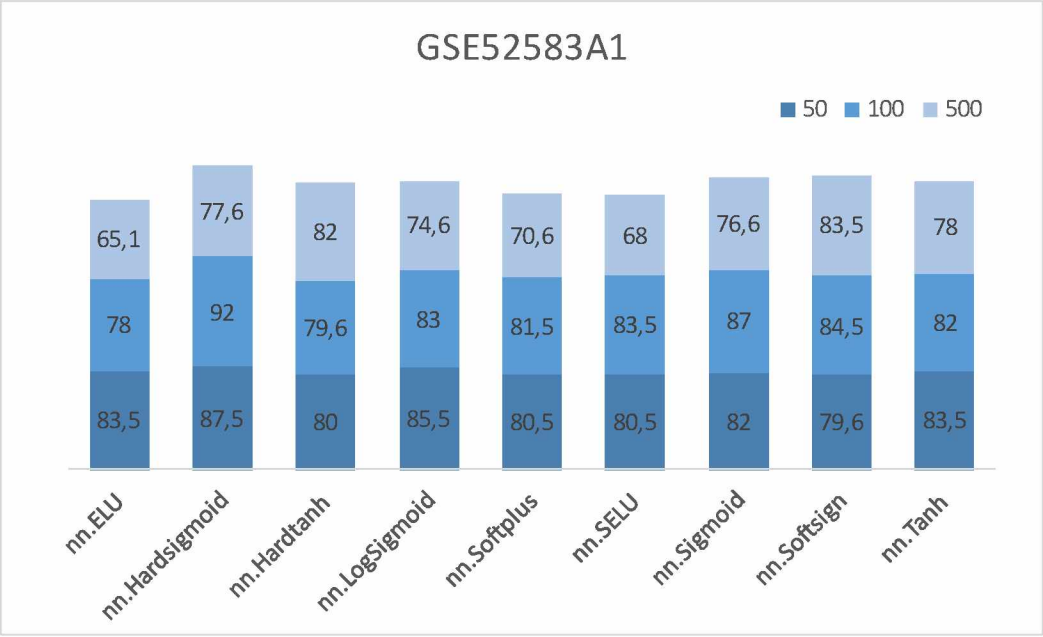
Για την διαπίστωση της αποτελεσματικότητας της μεθόδου scVEC που περιεγράφηκε, ακολουθήθηκαν κάποιες διαδικασίες σύγκρισης με πιθανές εναλλακτικές μεθόδους με σκοπό την κατηγοριοποίηση των κυττάρων με βάση την γονιδιακή τους έκφραση. Στα αποτελέσματα είναι εμφανές ότι ο συγκεκριμένος συνδυασμός τεχνικών, που αποτελούν τη μέθοδο που υλοποιήθηκε, παράγει τα βέλτιστα αποτελέσματα, καθώς φαίνεται ότι η ακρίβεια της μεθόδου είναι μεγάλη, όπως και η ισχύς της κατηγοριοποίησης των διαθέσιμων συνόλων δεδομένων που χρησιμοποιήθηκαν. Τα αποτελέσματα αυτά είναι σημαντικό να αναφερθεί ότι επιβεβαιώνονται και επαληθεύονται από την παρόμοια μέθοδο scIAE που δημοσιεύτηκε τον Δεκέμβρη του 2020 και που παρουσιάζεται στη βιβλιογραφική ανασκόπηση της εργασίας. Η διαφορά της συγκεκριμένης μεθόδου με αυτή που παρουσιάζεται στο πλαίσιο της εργασίας είναι ότι η δεύτερη χρησιμοποιεί Variational Autoencoder, ενώ η πρώτη χρησιμοποιεί stacked denoising sparse Autoencoder. Στην ενότητα αυτή παρουσιάζονται τα αποτελέσματα των παραπάνω συγκρίσεων αλλά και η διαδικασία επιλογής βέλτιστων παραμέτρων με σκοπό την καλύτερη απόδοση του μοντέλου.

5.3.1. Ρύθμιση παραμέτρων

Με σκοπό τη βελτιστοποίηση της απόδοσης του μοντέλου χρειάστηκε να ρυθμιστούν κάποιες βασικές παράμετροι. Οι παράμετροι περιλαμβάνουν την εύρεση της κατάλληλης συνάρτησης ενεργοποίησης και του αριθμού των διαστάσεων του λανθάνοντος χώρου, στον οποίο διατηρείται όσο μεγαλύτερο ποσοστό πληροφορίας των αρχικών δεδομένων.

Με βάση τα αποτελέσματα είναι εμφανές ότι οι συναρτήσεις ενεργοποίησης με τη μεγαλύτερη ακρίβεια είναι οι ELU, Hardsigmoid, Hardtanh, LogSigmoid, Softplus, SELU, Sigmoid, Softsign και η Tanh με περισσότερο αποδοτική την Hardsigmoid. Παρόλο που έγινε μια προσπάθεια ερμηνείας της καλύτερης απόδοσης της συγκεκριμένης συνάρτησης, συγκριτικά με τις υπόλοιπες, δεν προέκυψε κάποιο συμπέρασμα. Όσον αφορά τη διαστατικότητα, τη χειρότερη απόδοση είχαν οι 500 διαστάσεις στον λανθάνοντα χώρο με τις 50 και 100 να είναι σχεδόν εξίσου αποτελεσματικές, ανάλογα με το σύνολο δεδομένων εισόδου.

1	GSE52583									
2		nn.ELU	nn.Hardsigmoid	nn.Hardtanh	nn.LogSigmoid	nn.SELU	nn.Sigmoid	nn.Softplus	nn.Softsign	nn.Tanh
3	**50 dim**	83,5	87,5	80	85,5	80,5	82	80,5	79,6	83,5
4	**100 dim**	78	92	79,6	83	83,5	87	81,5	84,5	82
5	**500 dim**	65,1	77,6	82	74,6	68	76,6	70,6	83,5	78
6										
7	GSE86469									
8		nn.ELU	nn.Hardsigmoid	nn.Hardtanh	nn.LogSigmoid	nn.SELU	nn.Sigmoid	nn.Softplus	nn.Softsign	nn.Tanh
9	**50 dim**	82,9	86,3	85,8	74,6	85,2	86,2	73	86	85
10	**100 dim**	85	88,5	86	83	86,8	89	82	86	86
11	**500 dim**	84	85	86,2	85,5	84,6	85	85	87	85,7
12										
13	E-MTAB-2805									
14		nn.ELU	nn.Hardsigmoid	nn.Hardtanh	nn.LogSigmoid	nn.Softplus	nn.SELU	nn.Sigmoid	nn.Softsign	nn.Tanh
15	**50 dim**	85	88	80,5	83	85	86	87	84	85
16	**100 dim**	83	90	86	86	86	86	87,8	89,5	83
17	**500 dim**	47	74,6	84	44	45	55	56,5	84	81
18										
19	GSE67120									
20		nn.ELU	nn.Hardsigmoid	nn.Hardtanh	nn.LogSigmoid	nn.Softplus	nn.SELU	nn.Sigmoid	nn.Softsign	nn.Tanh
21	**50 dim**	74,5	73	76,7	66,8	63	76	81,7	83	83
22	**100 dim**	71,8	75	78	76	74	72	77	79	78
23	**500 dim**	55	69,6	73	53,5	55,8	53	58	72	69
24										

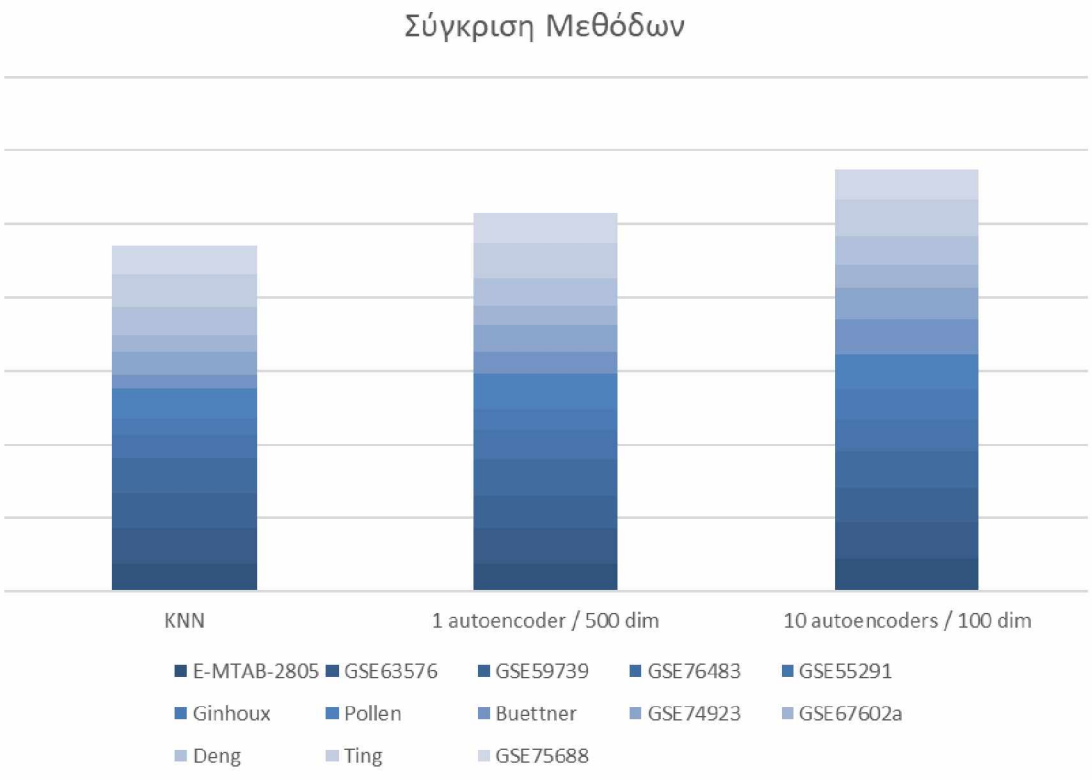


Εικόνα 36 Αποτελέσματα ρύθμισης παραμέτρων για 4 σύνολα δεδομένων

5.3.2. Αποτελέσματα Συγκρίσεων

Για την εξαγωγή συμπερασμάτων απόδοσης του μοντέλου χρησιμοποιήθηκαν 12 σύνολα δεδομένων από scRNA-seq. Εφαρμόστηκε σύγκριση ανάμεσα στην εφαρμογή μεθόδου KNN απευθείας στα δεδομένα εισόδου, στην εφαρμογή ενός μοντέλου που αποτελείται από έναν Autoencoder με 500 διαστάσεις στον λανθάνοντα χώρο και του μοντέλου που υλοποιήθηκε και περιλαμβάνει 10

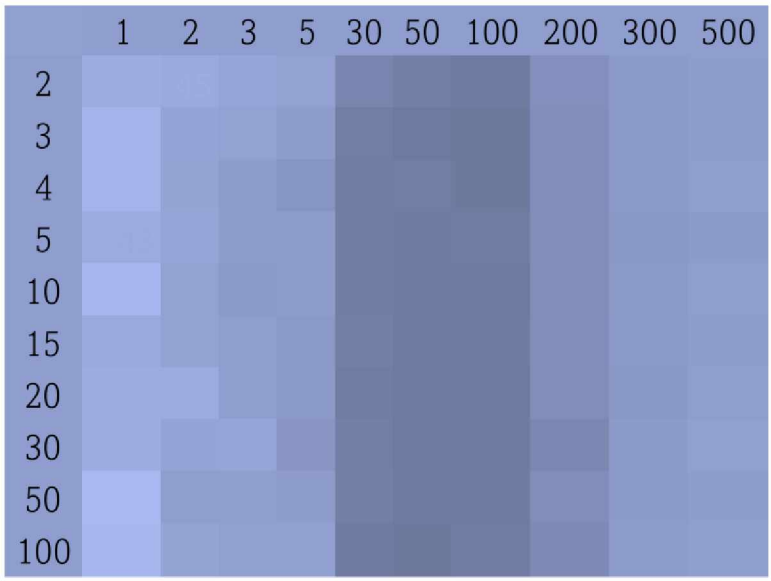
Autoencoders και 100 διαστάσεις. Τα αποτελέσματα, όπως φαίνονται στην Εικόνα 37 αποδεικνύουν αρχικά ότι ο συνδυασμός του Autoencoder με τον KNN αποδίδει καλύτερα σε σχέση με την εφαρμογή μόνο του KNN. Στην συνέχεια είναι επίσης έκδηλο ότι με τη χρήση περισσότερων Autoencoders είναι δυνατή η μεγαλύτερη μείωση των αρχικών διαστάσεων, κρατώντας μεγαλύτερο ποσοστό της πληροφορίας των αρχικών δεδομένων και έχοντας καλύτερα αποτελέσματα σε σύγκριση με μια μικρότερη μείωση της διάστασης των αρχικών δεδομένων, αλλά τη χρήση ενός μόνο Autoencoder. Μάλιστα στα 10 από τα 12 σύνολα που χρησιμοποιήθηκαν η διαφορά ανάμεσα στην απόδοση με τις άλλες μεθόδους είναι σημαντική. Παρόλα αυτά χρειάζεται να επισημανθεί ότι κάθε σύνολο δεδομένων είναι διαφορετικό και μπορεί να χρειαστεί διαφορετικό συνδυασμό παραμέτρων για να έχει μεγαλύτερη ακρίβεια. Όμως η ακρίβεια του μοντέλου φαίνεται να είναι αρκετά μεγάλη ακόμη και για σταθερό αριθμό διαστάσεων και ίδιες παραμέτρους κάθε φορά. Ακόμη χρειάζεται να αναφερθεί ότι έγινε διαφορετική επεξεργασία στα δεδομένα για την χρήση τους απευθείας στον KNN καθώς εφαρμόστηκε η preprocessing.normalize() της βιβλιοθήκης sklearn για να προκύψουν καλύτερα αποτελέσματα, ενώ για τις άλλες δύο μεθόδους χρησιμοποιήθηκε η preprocesing.StandardScaler().



Εικόνα 37 Αποτελέσματα σύγκρισης των τριών μεθόδων. Σε 10 από τα 13 σύνολα δεδομένων που χρησιμοποιήθηκαν η μέθοδος έχει πολύ καλύτερη ακρίβεια σε σχέση με τις υπόλοιπες.

5.3.3. Πρόσθετα συμπεράσματα

Χρησιμοποιήθηκαν δύο από τα διαθέσιμα σύνολα δεδομένων (Buettner, Ginhoux) για να υπολογιστεί η ακρίβεια του μοντέλου έχοντας 3,10,30,100 Autoencoders, από κάθε έναν από τους οποίους προκύπτει χώρος με 5,50,100 και 500 τελικές διαστάσεις. Από τα αποτελέσματα προκύπτει ότι στις ακραίες τιμές διαστάσεων η ακρίβεια του μοντέλου ελαχιστοποιείται και η απόδοση είναι βέλτιστη όταν οι διαστάσεις είναι 50 ή 100. Για τους αριθμούς των Autoencoders που χρησιμοποιήθηκαν δεν παρατηρήθηκαν σημαντικές διαφορές που επηρεάζουν την απόδοση του μοντέλου, οπότε η επιλογή του κατάλληλου αριθμού χρειάζεται περισσότερη διερεύνηση.



Εικόνα 38 Heatmap που κατασκευάστηκε από τις τιμές που προέκυψαν για το σύνολο δεδομένων Buettner. Στον άξονα x είναι οι τιμές των διαστάσεων που ελέγχθηκαν και στον άξονα y ο αριθμός των Autoencoders που χρησιμοποιήθηκαν.

5.4. Παράρτημα Κώδικα

5.4.1. Εισαγωγή

Για την υλοποίηση της μεθόδου scVEC χρησιμοποιήθηκαν βιβλιοθήκες της Python, της Pytorch και της sklearn. Οι βιβλιοθήκες αυτές θεωρήθηκαν καταλληλότερες και πιο εύχρηστες για την εφαρμογή μεθόδων μηχανικής μάθησης καθώς περιλαμβάνουν αρκετές από τις απαραίτητες διαδικασίες που χρειάστηκαν όπως η μέθοδος Cross Validation και ο αλγόριθμος KNN.

Στον κώδικα που παρουσιάζεται στο κεφάλαιο αυτό περιλαμβάνεται η λίστα με όλες τις βιβλιοθήκες που χρησιμοποιήθηκαν καθώς και οι κλάσεις χειρισμού των δεδομένων εισόδου, εκτός από την

υλοποίηση της μεθόδου. Ακόμη, δίνεται η δυνατότητα χρήσης της πλατφόρμας CUDA σε περίπτωση που διαθέτει ο υπολογιστής καλύτερη κάρτα γραφικών για γρηγορότερη επεξεργασία. Ο αλγόριθμος μπορεί να επεξεργαστεί mat αλλά και csv αρχεία μέσω της επιλογής της κατάλληλης μεθόδου στην κλάση «Databuilder».

Ο αλγόριθμος χρησιμοποιεί ολόκληρο το σύνολο δεδομένων προκειμένου να εκπαιδεύσει το μοντέλο και στη συνέχεια γίνεται έλεγχος μέσω μιας 10 fold cross validation μεθόδου. Η επιλογή των κατάλληλων παραμέτρων γίνεται στο τέλος του κώδικα και περιλαμβάνει την επιλογή του αριθμού των επαναλήψεων ή αριθμού των autoencoders, την επιλογή του αριθμού των τελικών διαστάσεων των δεδομένων, του αριθμού των εποχών εκπαίδευσης, της επιθυμητής συνάρτησης ενεργοποίησης και του συνόλου δεδομένων εισόδου.

5.4.2. Κώδικας

```
import numpy as np
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import os
import scipy.io
import torch
import torch.nn as nn
import torch.nn.functional as F

from torch.utils.data import Dataset, DataLoader
from torch import nn, optim
from torch.autograd import Variable
from sklearn import preprocessing
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import KFold
from sklearn.neighbors import KNeighborsClassifier
from sklearn import metrics

device = torch.device('cuda' if torch.cuda.is_available() else 'cpu')

def major_class(a):
    mc = a[0]
    for i in range(len(a)):
        for j in range(len(a)):
            k = len(a)-1-j
```

```
            if(k)>i:
                if a[i] == a[k]:
                    mc = a[i]
    return mc

#processing of two types of input data
def load_mat_data(path):
    mat = scipy.io.loadmat(DATA_PATH)
    x,y = mat['data'], mat['class']
    x = x.astype('float32')
    # stadardize values
    standardizer = preprocessing.StandardScaler()
    x = standardizer.fit_transform(x)
    return x, standardizer, y

def load_csv_data(path):
    # read in from csv
    df = pd.read_csv(DATA_PATH, sep=',', header=0)
    # replace nan with -99
    df = df.fillna(-99)
    df_base = df.iloc[:, 1:]
    # get assigned_cluster Label
    y = df.iloc[:,0].values
    x = df_base.values.reshape(-1, df_base.shape[1]).astype('float32')
    # stadardize values
    standardizer = preprocessing.StandardScaler()
    x = standardizer.fit_transform(x)
    return x, standardizer, y

def numpyToTensor(x):
    x_train = torch.from_numpy(x).to(device)
    return x_train

class DataBuilder(Dataset):
    def __init__(self, path):
        # self.x, self.standardizer, self.y = load_csv_data(DATA_PATH)
        self.x, self.standardizer, self.y = load_mat_data(DATA_PATH)
        self.x = numpyToTensor(self.x)
        self.len=self.x.shape[0]
        self.y = numpyToTensor(self.y)
```

```
    def __getitem__(self,index):
        return (self.x[index], self.y[index])
    def __len__(self):
        return self.len

class Autoencoder(nn.Module):
    def __init__(self,act_func, D_in,H=512,H2=200,latent_dim=100):

        #Encoder
        super(Autoencoder,self).__init__()
        self.dropout = nn.Dropout(p=0.5)
        self.linear1=nn.Linear(D_in,H)
        self.lin_bn1 = nn.BatchNorm1d(num_features=H)
        self.linear2=nn.Linear(H,H2)
        self.lin_bn2 = nn.BatchNorm1d(num_features=H2)
        self.linear3=nn.Linear(H2,H2)
        self.lin_bn3 = nn.BatchNorm1d(num_features=H2)

         # Latent vectors mu and sigma
        self.fc1 = nn.Linear(H2, latent_dim)
        self.bn1 = nn.BatchNorm1d(num_features=latent_dim)
        self.fc21 = nn.Linear(latent_dim, latent_dim)
        self.fc22 = nn.Linear(latent_dim, latent_dim)

         # Sampling vector
        self.fc3 = nn.Linear(latent_dim, latent_dim)
        self.fc_bn3 = nn.BatchNorm1d(latent_dim)
        self.fc4 = nn.Linear(latent_dim, H2)
        self.fc_bn4 = nn.BatchNorm1d(H2)

         # Decoder
        self.linear4=nn.Linear(H2,H2)
        self.lin_bn4 = nn.BatchNorm1d(num_features=H2)
        self.linear5=nn.Linear(H2,H)
        self.lin_bn5 = nn.BatchNorm1d(num_features=H)
        self.linear6=nn.Linear(H,D_in)
        self.lin_bn6 = nn.BatchNorm1d(num_features=D_in)

        self.relu = act_func()
```

```
    def encode(self, x):
        lin1 = self.relu(self.lin_bn1(self.linear1(x)))
        lin2 = self.relu(self.lin_bn2(self.linear2(lin1)))
        lin3 = self.relu(self.lin_bn3(self.linear3(lin2)))

        fc1 = F.relu(self.bn1(self.fc1(lin3)))

        r1 = self.fc21(fc1)
        r2 = self.fc22(fc1)

        return r1, r2

    def reparameterize(self, mu, logvar):
        if self.training:
            std = logvar.mul(0.5).exp_()
            eps = Variable(std.data.new(std.size()).normal_())
            return eps.mul(std).add_(mu)
        else:
            return mu

    def decode(self, z):
        fc3 = self.relu(self.fc_bn3(self.fc3(z)))
        fc4 = self.relu(self.fc_bn4(self.fc4(fc3)))

        lin4 = self.relu(self.lin_bn4(self.linear4(fc4)))
        lin5 = self.relu(self.lin_bn5(self.linear5(lin4)))
        return self.lin_bn6(self.linear6(lin5))

    def forward(self, x):
        x = self.dropout(x)
        mu, logvar = self.encode(x)
        z = self.reparameterize(mu, logvar)
        return z, self.decode(z), mu, logvar

#training of the autoencoder using the dataset
def train(epochs,model, optimizer, loss_mse):
    model.train()
    train_loss = 0
    train_losses = []
    for epoch in range(epochs):
```

```
        for batch_idx, data in enumerate(trainloader):
            d = data[0].to(device)
            targets = data[1].to(device)

            optimizer.zero_grad()
            z, recon_batch, mu, logvar = model(d)

            loss = loss_mse(recon_batch, d, mu, logvar)
            loss.backward()
            train_loss += loss.item()

            optimizer.step()

        if epoch % 100 == 0:
            print(f'Epoch: {epoch}')
            print(f'Train Loss = {loss.item():.4f}')

        train_losses.append(train_loss / len(trainloader.dataset))

    return z

class customLoss(nn.Module):
    def __init__(self):
        super(customLoss, self).__init__()
        self.mse_loss = nn.MSELoss(reduction="sum")

    def forward(self, x_recon, x, mu, logvar):
        loss_MSE = self.mse_loss(x_recon, x)
        loss_KLD = -0.5 * torch.sum(1 + logvar - mu.pow(2) - logvar.exp())

        return loss_MSE + loss_KLD

# takes in a module and applies the specified weight initialization
def weights_init_uniform_rule(m):
    classname = m.__class__.__name__
    # for every Linear layer in a model..
    if classname.find('Linear') != -1:
        # get the number of the inputs
        n = m.in_features
        y = 1.0/np.sqrt(n)
```

```
        m.weight.data.uniform_(-y, y)
        m.bias.data.fill_(0)

def print_predictions(samples, pred_rpknn, major, classes):
    #Prints all predictions for the testing dataset
    count = 0
    for k in range(samples):
            #print(pred_rpknn[k], major[k], classes[k])
            if major[k] == classes[k]:
                count = count+1
    acc = count/samples
    print("Accuracy is", acc)

def kf_knn(data, classes, rp, epochs, act_func, dim):
    samples = len(classes)
    pred_rpknn = [[0] * samples for i in range(rp)]

    n_splits = 10
    kf = KFold(n_splits, shuffle=True, random_state=7)

    #Create new KNN Classifier
    knn = KNeighborsClassifier(n_neighbors=5)

    loss_mse = customLoss()

    for d in dim:
        for i in act_func:
            print(d, i, rp)
            for r in range(rp):

                D_in = data.shape[1]
                model = Autoencoder(i, D_in, 512, 200, d).to(device)
                model.apply(weights_init_uniform_rule)
                optimizer = optim.Adam(model.parameters(), lr=1e-3)
                latent_space = train(epochs, model, optimizer, loss_mse)

                pred_knn = [0]*samples

                #k_fold indexes generated for k (x/y) train (x/y) test iterations
                for train_index, test_index in kf.split(data):
```

```
                    x_train, x_test = latent_space[train_index], la-
tent_space[test_index]
                    y_train, y_test = classes[train_index], classes[test_index]
                    knn.fit(x_train.detach().numpy(), y_train.ravel())
                    pred_knn = knn.predict(x_test.detach().numpy())
                    #print("Accuracy:",metrics.accuracy_score(y_test, pred_knn))
                    c = 0
                    for idx in test_index:
                        pred_rpknn[r][idx] = pred_knn[c]
                        c +=1

            pred_rpknn_tr = np.transpose(pred_rpknn)

            #find major class from the predictions
            major = [0]*samples
            for i in range(samples):
                major[i] = major_class(pred_rpknn_tr[i])

            print_predictions(samples, pred_rpknn_tr, major, classes)

#parameter tuning & runs the algorithm
epochs = 100
rp = [3] #number of autoencoders used / repeats

act_func = [nn.Hardsigmoid]
dim = [50, 500, 100] #n0 of latent space's dimensions

DATA_PATHS = ['/home/user/Desktop/GSE52583.mat', '/home/user/Desktop/GSE86469.mat']

for DATA_PATH in DATA_PATHS:
    for r in rp:
        print(DATA_PATH)
        data_set=DataBuilder(DATA_PATH)

        trainloader=DataLoader(dataset=data_set,batch_size=data_set.y.shape[0],
num_workers=5)
        kf_knn(data_set.x, data_set.y, r, epochs, act_func, dim)
```

6. Βιβλιογραφία

A Gentle Introduction to k-fold Cross-Validation (no date). Available at: https://machinelearningmastery.com/k-fold-cross-validation/ (Accessed: December 2, 2021).

A Gentle Introduction to LSTM Autoencoders (no date). Available at: https://machinelearningmastery.com/lstm-autoencoders/ (Accessed: December 2, 2021).

Arumugam, R., Uli, J.E. and Annavi, G. (2019) "A Review of the Application of Next Generation Sequencing (NGS) in Wild Terrestrial Vertebrate Research," *Annual Research & Review in Biology* [Preprint]. doi:10.9734/arrb/2019/v31i530061.

Bellman R.E. (1961) "Adaptive Control Processes."

B.Meshram, S. and M. Shinde, S. (2015) "A Survey on Ensemble Methods for High Dimensional Data Classification in Biomedicine Field," *International Journal of Computer Applications*, 111(11). doi:10.5120/19580-1162.

Clarke, L. *et al.* (2017) "The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data," *Nucleic Acids Research*, 45(D1). doi:10.1093/nar/gkw829.

Complete Guide to Factor Analysis (Updated 2022) - Qualtrics (no date). Available at: https://www.qualtrics.com/experience-management/research/factor-analysis/ (Accessed: March 18, 2022).

Dash, S. *et al.* (2019) "Big data in healthcare: management, analysis and future prospects," *Journal of Big Data*, 6(1). doi:10.1186/s40537-019-0217-0.

Deng, Y. *et al.* (2019a) "Scalable analysis of cell type composition from single-cell transcriptomics using deep recurrent learning," *Nature methods*, 16(4), p. 311. doi:10.1038/S41592-019-0353-7.

Deng, Y. *et al.* (2019b) "Scalable analysis of cell-type composition from single-cell transcriptomics using deep recurrent learning," *Nature Methods*, 16(4). doi:10.1038/s41592-019-0353-7.

Dong, Z. and Alterovitz, G. (2021) "netAE: semi-supervised dimensionality reduction of single-cell RNA sequencing to facilitate cell labeling," *Bioinformatics*, 37(1), pp. 43–49. doi:10.1093/BIOINFORMATICS/BTAA669.

Donoho, D.L. (2000) "Aide-Memoire. High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality," *American Math. Society Lecture-Math Challenges of the 21st Century* [Preprint].

Doug Laney (2001) *3D Data Management: Controlling DataVolume, Velocity, and Variety*.

Du, B., Kong, X. and Feng, X. (2020) "Generalized Principal Component Analysis-Based Subspace Decomposition of Fault Deviations and Its Application to Fault Reconstruction," *IEEE Access*, 8. doi:10.1109/ACCESS.2020.2971507.

Durif, G. *et al.* (2019) "Probabilistic count matrix factorization for single cell expression data analysis," *Bioinformatics*, 35(20). doi:10.1093/bioinformatics/btz177.

Eraslan, G. *et al.* (2019a) "Single-cell RNA-seq denoising using a deep count autoencoder," *Nature Communications 2019 10:1*, 10(1), pp. 1–14. doi:10.1038/s41467-018-07931-2.

Eraslan, G. *et al.* (2019b) "Single-cell RNA-seq denoising using a deep count autoencoder," *Nature Communications*, 10(1). doi:10.1038/s41467-018-07931-2.

Ganaie, M.A. *et al.* (2021) "Ensemble deep learning: A review." Available at: http://arxiv.org/abs/2104.02395.

van de Geer, S.A. and van Houwelingen, H.C. (2004) "High-dimensional data: P ≫ n in mathematical statistics and bio-medical applications," *Bernoulli*, 10(6). doi:10.3150/bj/1106314843.

Haque, A. *et al.* (2017) "A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications," *Genome Medicine*. doi:10.1186/s13073-017-0467-4.

Hastie, T. and Tibshirani, R. (2003) "Expression arrays and the p≫ n problem," *See< http://www-stat. stanford. edu/~ hastie/Papers/pgtn. pdf* [Preprint].

Hinton, G.E. and Salakhutdinov, R.R. (2006) "Reducing the dimensionality of data with neural networks," *Science*, 313(5786). doi:10.1126/science.1127647.
Hunter, D.R. and Lange, K. (2004) "A Tutorial on MM Algorithms," *American Statistician*, 58(1). doi:10.1198/0003130042836.
Intuitively Understanding Variational Autoencoders | by Irhum Shafkat | Towards Data Science (no date). Available at: https://towardsdatascience.com/intuitively-understanding-variational-autoencoders-1bfe67eb5daf (Accessed: December 2, 2021).
Jang, E. (2016) *Tutorial: Categorical Variational Autoencoders using Gumbel-Softmax, Web Page*.
Jang, E., Gu, S. and Poole, B. (2017) "Categorical reparameterization with gumbel-softmax," in *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*.
Johnstone, I.M. and Titterington, D.M. (2009) "Statistical challenges of high-dimensional data," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. doi:10.1098/rsta.2009.0159.
Kingma, D.P. and Welling, M. (2019) "An introduction to variational autoencoders," *Foundations and Trends in Machine Learning*. doi:10.1561/2200000056.
K-Nearest Neighbor(KNN) Algorithm for Machine Learning - Javatpoint (no date). Available at: https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning (Accessed: December 2, 2021).
Landgraf, A.J. and Lee, Y. (2020) "Dimensionality reduction for binary data through the projection of natural parameters," *Journal of Multivariate Analysis*, 180. doi:10.1016/j.jmva.2020.104668.
Li, Z. *et al.* (2012) "Computational intelligence and intelligent systems: 6th International Symposium, ISICA 2012 Wuhan, China, October 27-28, 2012 Proceedings," *Communications in Computer and Information Science*, 316. doi:10.1007/978-3-642-34289-9.
Lin, P., Troup, M. and Ho, J.W.K. (2017) "CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data," *Genome Biology*, 18(1), pp. 1–11. doi:10.1186/S13059-017-1188-0/FIGURES/5.
Lundstrom, M. (2003) "Moore's Law Forever?," *Science*, 299(5604). doi:10.1126/science.1079567.
McInnes, L. *et al.* (2018) "UMAP: Uniform Manifold Approximation and Projection," *Journal of Open Source Software*, 3(29). doi:10.21105/joss.00861.
Meng, L. *et al.* (2018) "Research of stacked denoising sparse autoencoder," *Neural Computing and Applications*, 30(7), pp. 2083–2100. doi:10.1007/S00521-016-2790-X.
Next-Generation Sequencing (NGS) | Explore the technology (no date). Available at: https://emea.illumina.com/science/technology/next-generation-sequencing.html (Accessed: December 2, 2021).
Non-Negative Matrix Factorization (no date). Available at: https://docs.oracle.com/database/121/DMCON/GUID-76F89641-E1D3-4B11-8319-4A152389D510.htm#DMCON058 (Accessed: December 2, 2021).
Pierson, E. and Yau, C. (2015) "ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis," *Genome Biology*, 16(1). doi:10.1186/s13059-015-0805-z.
Porte, J.D. la and Herbst, B. (2008) "An introduction to diffusion maps," *... Sciences, University of ...* [Preprint].
Q, Y. *et al.* (2021) "scIAE: an integrative autoencoder-based ensemble classification framework for single-cell RNA-seq data," *Briefings in bioinformatics* [Preprint]. doi:10.1093/BIB/BBAB508.
Risso, D. *et al.* (2018) "A general and flexible method for signal extraction from single-cell RNA-seq data," *Nature Communications*, 9(1). doi:10.1038/s41467-017-02554-5.
Saul, L. and Roweis, S. (2000) "An introduction to locally linear embedding," *unpublished. Available at: http://www. cs. toronto. ...* [Preprint].
scRNA-Seq (no date). Available at: https://emea.illumina.com/science/sequencing-method-explorer/kits-and-arrays/scrna-seq.html (Accessed: December 2, 2021).

Shawe-Taylor, J. and Cristianini, N. (2004) *Kernel Methods for Pattern Analysis*, *Kernel Methods for Pattern Analysis*. doi:10.1017/cbo9780511809682.
Shortliffe, E.H. and Barnett, G.O. (2014) "Biomedical data: Their acquisition, storage, and use," in *Biomedical Informatics: Computer Applications in Health Care and Biomedicine: Fourth Edition*. doi:10.1007/978-1-4471-4474-8_2.
Stacking Ensemble Machine Learning With Python (no date). Available at: https://machinelearningmastery.com/stacking-ensemble-machine-learning-with-python/ (Accessed: December 2, 2021).
Statistical and Mathematical Concepts behind PCA | by Rukshan Pramoditha | Data Science 365 | Medium (no date). Available at: https://medium.com/data-science-365/statistical-and-mathematical-concepts-behind-pca-a2cb25940cd4 (Accessed: December 2, 2021).
Topic Modeling Articles with NMF. Extracting topics is a good... | by Rob Salgado | Towards Data Science (no date). Available at: https://towardsdatascience.com/topic-modeling-articles-with-nmf-8c6b2a227a45 (Accessed: December 2, 2021).
Torgerson, W.S. (1958) *Theory and methods of scaling. Wiley.* Edited by W.S. Torgerson.
Townes, F.W. *et al.* (2020) "Erratum: Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model (Genome Biology (2019) 20 (295) DOI: 10.1186/s13059-019-1861-6)," *Genome Biology*. doi:10.1186/s13059-020-02109-w.
t-SNE: Behind the Math. Being one of the most talked about... | by Sushanth Sreenivasa | Towards Data Science (no date). Available at: https://towardsdatascience.com/t-sne-behind-the-math-4d213b9ebab8 (Accessed: December 2, 2021).
t-SNE clearly explained. An intuitive explanation of t-SNE... | by Kemal Erdem (burnpiro) | Towards Data Science (no date). Available at: https://towardsdatascience.com/t-sne-clearly-explained-d84c537f53a (Accessed: December 2, 2021).
Understanding Variational Autoencoders (VAEs) | by Joseph Rocca | Towards Data Science (no date). Available at: https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73 (Accessed: December 2, 2021).
Using Linear Discriminant Analysis (LDA) for data Explore: Step by Step. | Blog (no date). Available at: https://www.apsl.net/blog/2017/07/18/using-linear-discriminant-analysis-lda-data-explore-step-step/ (Accessed: December 3, 2021).
Vailati-Riboni, M., Palombo, V. and Loor, J.J. (2017) "What are omics sciences?," in *Periparturient Diseases of Dairy Cows: A Systems Biology Approach*. doi:10.1007/978-3-319-43033-1_1.
Vrahatis, A.G. *et al.* (2020) "Ensemble Classification through Random Projections for single-cell RNA-seq data," *undefined* [Preprint]. doi:10.1101/2020.06.24.169136.
Wang, D. and Gu, J. (2018) "VASC: Dimension Reduction and Visualization of Single-cell RNA-seq Data by Deep Variational Autoencoder," *Genomics, Proteomics & Bioinformatics*, 16(5), pp. 320–331. doi:10.1016/J.GPB.2018.08.003.
What is Artificial Intelligence (AI)? | IBM (no date). Available at: https://www.ibm.com/cloud/learn/what-is-artificial-intelligence (Accessed: December 2, 2021).
Why and how to Cross Validate a Model? | by Sanjay.M | Towards Data Science (no date). Available at: https://towardsdatascience.com/why-and-how-to-cross-validate-a-model-d6424b45261f (Accessed: December 2, 2021).
Xiu, L. (2019) "Time Moore: Exploiting Moore's Law from the Perspective of Time," *IEEE Solid-State Circuits Magazine*. doi:10.1109/MSSC.2018.2882285.

7. Πηγές εικόνων

ΗΜΕΡΟΜΗΝΙΑ ΑΝΑΚΤΗΣΗΣ: 22/1/2022

Εικόνα 1 Διαδικασία εργαστηριακής μεθόδου αλληλούχισης RNA μεμονωμένου κυττάρου

https://learn.gencore.bio.nyu.edu/single-cell-rnaseq/

Εικόνα 2 Κατηγορίες μεθόδων μηχανικής μάθησης και εφαρμογές σε βιοϊατρικά δεδομένα

https://www.nature.com/articles/s41582-020-0377-8

Εικόνα 3 Εφαρμογές μεθόδων μηχανικής μάθησης

https://www.flickr.com/photos/184632966@N04/49599262071

Εικόνα 4 Διάκριση ανάμεσα στις έννοιες μηχανική μάθηση, βαθιά μάθηση και τεχνητή νοημοσύνη

https://www.edureka.co/blog/ai-vs-machine-learning-vs-deep-learning/

Εικόνα 5 Στόχοι του προγράμματος χαρτογράφησης του ανθρώπινου γονιδιώματος

https://geneticeducation.co.in/the-human-genome-project-aims-objectives-techniques-and-outcomes/

Εικόνα 6 Τα χαρακτηριστικά των μεγάλων δεδομένων του Douglas Laney

https://www.techentice.com/the-data-veracity-big-data/

Εικόνα 7 Η κατάρα της διαστατικότητας

https://deepai.org/machine-learning-glossary-and-terms/curse-of-dimensionality

Εικόνα 8 Κεντρικό δόγμα της μοριακής βιολογίας

https://quizlet.com/154819566/joshua-smith-bms110-central-dogma-of-molecular-biology-ch-20i-ch-1s-ch-12-s-study-guide-flash-cards/

Εικόνα 9 Διαφορά NGS με scRNA-seq

https://www.technologynetworks.com/genomics/articles/recent-advances-in-single-cell-genomics-techniques-324695

Εικόνα 10 Σύγκριση μεθόδων μείωσης διαστάσεων των μεγάλων δεδομένων

https://quantdare.com/what-is-the-difference-between-feature-extraction-and-feature-selection/

Εικόνα 11 Εφαρμογή PCA σε δεδομένα τριών διαστάσεων

https://www.analyticsvidhya.com/blog/2016/03/pca-practical-guide-principal-component-analysis-python/

Εικόνα 12 Εφαρμογή PCA σε δεδομένα τριών διαστάσεων

https://www.spss-tutorials.com/spss-factor-analysis-tutorial/

Εικόνα 13 Παράδειγμα ανάλυσης παραγόντων / Κατασκευάστηκε στο Word

Εικόνα 14 Εύρεση βέλτιστης ευθείας για προβολή των σημείων, ώστε να επιτευχθεί η μέγιστη διαχωρισιμότητα των κλάσεων

https://www.quora.com/What-is-an-intuitive-explanation-for-linear-discriminant-analysis-LDA

Εικόνα 15 Απεικόνιση παραγοντοποίησης μη αρνητικού πίνακα (NMF)

https://en.wikipedia.org/wiki/Non-negative_matrix_factorization

Εικόνα 16 Εφαρμογή PCA πυρήνα για μη γραμμικά διαχωρίσιμα δεδομένα
https://stats.stackexchange.com/questions/94463/what-are-the-advantages-of-kernel-pca-over-standard-pca
Εικόνα 17 Απεικόνιση της Isomap σε ένα ελβετικό σύνολο δεδομένων ρολού ή swiss roll dataset
https://www.science.org/doi/10.1126/science.295.5552.7a
Εικόνα 17 Γεωδαισιακή απόσταση ενός ζεύγους σημείων
https://dsp.stackexchange.com/questions/54826/what-is-different-between-euclidean-distance-and-the-geodesic-distance/54827
Εικόνα 19 Εφαρμογή της t-SNE στις δύο διαστάσεις
https://towardsdatascience.com/t-sne-behind-the-math-4d213b9ebab8
Εικόνα 20 Εφαρμογή του χάρτη διάχυσης για μείωση διάστασης του ελβετικού συνόλου δεδομένων ρολού
https://towardsdatascience.com/unwrapping-the-swiss-roll-9249301bd6b7
Εικόνα 21 Σύγκριση της LLE με άλλες μεθόδους μείωσης διαστάσεων
https://www.astroml.org/book_figures/chapter7/fig_S_manifold_PCA.html
Εικόνα 22 Δημιουργία συνάψεων δεδομένων με βάση τις αποστάσεις ανάμεσά τους
https://pair-code.github.io/understanding-umap/supplement.html
Εικόνα 23 Λεπτομερής περιγραφή της εσωτερικής δομής ενός Autoencoder
https://towardsdatascience.com/applied-deep-learning-part-3-autoencoders-1c083af4d798
Εικόνα 24 Δομή του scScope
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6774994/
Εικόνα 25 Δομή του Variational Autoencoder
https://towardsdatascience.com/an-introduction-to-variational-auto-encoders-vaes-803ddfb623df
Εικόνα 26 Στιγμιότυπο αλγόριθμου KNN
https://towardsdatascience.com/knn-k-nearest-neighbors-1-a4707b24bd1d
Εικόνα 27 Μέθοδος διασταυρωμένης επικύρωσης δεδομένων
https://androidkt.com/pytorch-k-fold-cross-validation-using-dataloader-and-sklearn/
Εικόνα 28 Μέθοδος Ensemble Learning ή Συλλογικής μάθησης
https://peerj.com/articles/cs-425/
Εικόνα 29 Λεπτομερής δομή του μοντέλου scIAE
https://pubmed.ncbi.nlm.nih.gov/34913057/
Εικόνα 30 Λεπτομερής δομή μεθόδου MPRV
https://www.semanticscholar.org/paper/Ensemble-Classification-through-Random-Projections-Vrahatis-Tasoulis/42787bc85dc34be6dbc6472006c4f427b1bfe77c
Εικόνα 31 Περιγραφή δομής μεθόδου VASC
https://www.sciencedirect.com/science/article/pii/S167202291830439X
Εικόνα 32 Κατανομή ετικετών των δεδομένων μετά την λογαριθμική μετατροπή για εφαρμογή της μεθόδου CIDR & Εικόνα 33 Περίληψη βημάτων για εφαρμογή της μεθόδου CIDR
https://static-content.springer.com/esm/art%3A10.1186%2Fs13059-017-1188-0/MediaObjects/13059_2017_1188_MOESM1_ESM.pdf
Εικόνα 34 Εσωτερική δομή μεθόδου netAE
https://academic.oup.com/bioinformatics/article/37/1/43/5877940?login=true
Εικόνα 35 Αναπαράσταση της μεθόδου scVEC
Created with BioRender.com
Εικόνα 36 Αποτελέσματα ρύθμισης παραμέτρων για ένα σύνολο δεδομένων
Εικόνα 37 Αποτελέσματα σύγκρισης των τριών μεθόδων
Εικόνα 38 Heatmap που κατασκευάστηκε από τις τιμές που προέκυψαν για το σύνολο Buettner