ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ - ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΤΟΜΕΑΣ ΛΟΓΙΚΟΥ ΤΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΕΡΓΑΣΤΗΡΙΟ ΑΝΑΛΥΣΗΣ ΠΟΛΥΔΙΑΣΤΑΤΩΝ

ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΔΙΑΧΕΙΡΙΣΗΣ ΓΝΩΣΗΣ

Προπτυχιακή Διπλωματική Εργασία

Κατηγοριοποίηση Ιστοπαθολογικών Εικόνων με χρήση Μάθησης Πολλαπλών Στιγμιότυπων

Ονοματεπώνυμο: Ευθύμιος Μενύχτας

Αριθμός Μητρώου: 235585

Επιβλέπων: Καθηγητής Μεγαλοοικονόμου Βασίλειος

Μέλη Επιτροπής: Καθηγητής Μπερμπερίδης Κωνσταντίνος

Αναπληρωτής Καθηγητής Μακρής Χρήστος

ΠΑΤΡΑ, Δεκέμβριος 2019

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή της παρούσας διπλωματικής εργασίας κ. Βασίλειο Μεγαλοοικονόμου, τόσο για την εμπιστοσύνη που μου έδειξε μέσα από την συνεργασία μας όσο και για την πολύτιμη καθοδήγηση του κατά την διάρκεια εκπόνησης της εργασίας.

Επίσης ευχαριστώ πολύ θερμά τον υποψήφιο διδάκτορα κ. Θωμά Παπαστεργίου για την υπομονή που έδειξε και την πολύτιμη υποστήριξη που μου παρείχε με τις γνώσεις του, βοηθώντας με στις πιο δύσκολες στιγμές της διαδικασίας αυτής.

Φυσικά θέλω να ευχαριστήσω και την οικογένεια μου για όλες τις θυσίες τους αυτά τα δύσκολα χρόνια της φοίτησης και περισσότερο τον αδερφό μου και τελειόφοιτο του τμήματος Βασίλειο Μενύχτα ο οποίος κατά κάποιο τρόπο χάραξε τον δρόμο και για μένα.

Περίληψη

Το αντικείμενο της παρούσας διπλωματικής εργασίας είναι η μελέτη κάποιων αλγορίθμων μηχανικής μάθησης που ανήκουν στην μάθηση πολλαπλών στιγμιότυπων (multiple instance learning) πάνω σε ιστοπαθολογικές εικόνες νεοπλασιών του μαστού. Πιο συγκεκριμένα, η μέθοδος που υλοποιήθηκε στην εργασία αυτή κάνει χρήση πολλαπλών one-class SVM σε επίπεδο στιγμιότυπων (instance level) μαζί με κάποιες μετρικές επιπέδου αντικειμένων (bag level) για την κατηγοριοποίηση τους.

Στα πλαίσια της εργασίας επίσης μελετήθηκε η κανονικοποίηση της απόχρωσης των λεγόμενων H&E ιστοπαθολογικών εικόνων, μία μέθοδος κατακερματισμού εικόνας για την παραγωγή πολλαπλών στιγμιότυπων, η χρήση του CPD για την εξαγωγή χαρακτηριστικών από εικόνες, η χρήση της PCA για την μείωση του αριθμού των χαρακτηριστικών και δύο μέθοδοι ώστε οι τιμές απόφασης διαφορετικών one-class SVM να είναι συγκρίσιμες μεταξύ τους.

Ή μέθοδος που αναπτύξαμε έχει αρκετές υπερ-παραμέτρους για τις οποίες έγιναν συγκριτικά πειράματα ώστε να μελετηθούν. Στην συνέχεια, για την αξιολόγηση της επιλέχθηκαν οι βέλτιστες από αυτές σύμφωνα με τα πειράματα και δοκιμάστηκαν πάνω σε δυο σύνολα δεδομένων με καλοήθεις και κακοήθεις νεοπλασίες του μαστού. Τέλος, έγιναν μερικές συγκρίσεις με αντίστοιχα αποτελέσματα για τα δεδομένα αυτά από την σχετική βιβλιογραφία.

ΠΕΡΙΕΧΟΜΕΝΑ

Κεφαλαιο 1: Εισαγωγη

1.1. Περιγραφή του Προβλήματος

Η παρούσα διπλωματική εργασία εστιάζει στο πρόβλημα της διάγνωσης των νεοπλασιών του μαστού και πιο συγκεκριμένα στην εφαρμογή μεθόδων μάθησης πολλαπλών στιγμιότυπων για την κατηγοριοποίηση ιστοπαθολογικών εικόνων τέτοιων νεοπλασιών σε καλοήθεις και κακοήθεις.

Σύμφωνα με έρευνα του Παγκόσμιου Οργανισμού Υγείας το 2014 [1], ο καρκίνος του μαστού είναι η πιο κοινή μορφή καρκίνου σε γυναίκες παγκοσμίως με υψηλά επίπεδα θνησιμότητας. Αποτελεί το 25.2% των περιστατικών καρκίνου με θνησιμότητα που φτάνει το 14.7%, δεύτερος μετά τον καρκίνο του πνεύμονα. Περίπου μισό εκατομμύριο γυναίκες έχουν πεθάνει από αυτόν και σχεδόν 1.7 εκατομμύρια νέες περιπτώσεις προκύπτουν κάθε χρόνο δείχνοντας αυξητική τάση με την πάροδο των χρόνων.

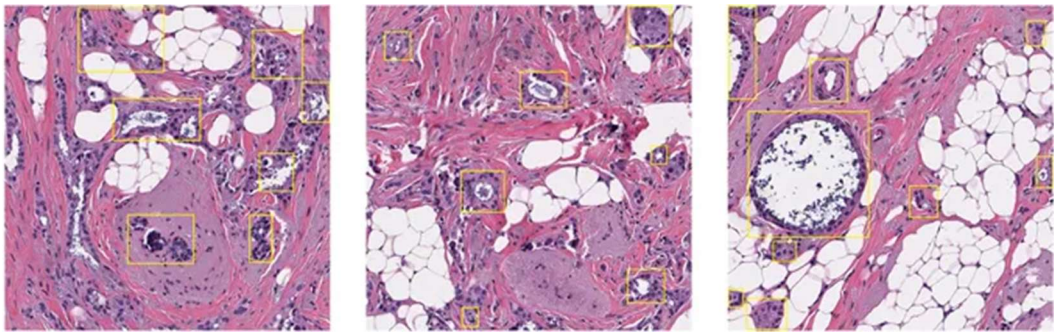
Η εδραιωμένη μέθοδος διάγνωσης αυτής της μορφής καρκίνου είναι με την χρήση ιστοπαθολογικών εικόνων, ιστού βαμμένου με αιματοξυλίνη και εωσίνη, που ονομάζονται H&E stains [2]. Η αιματοξυλίνη βάφει σκούρο μπλε τους πυρήνες των κυττάρων και η εωσίνη βάφει ροζ το κυτόπλασμα και την εξωκυτταρική μήτρα, με άλλες κυτταρικές δομές να παίρνουν κάποια απόχρωση των δύο αυτών βασικών χρωμάτων. Το αποτέλεσμα είναι οι κυτταρικές δομές να είναι πιο ευδιάκριτες με αυτόν τον τρόπο. Η μέθοδος των H&E stains έχει εδραιωθεί λόγω του πολύ χαμηλού κόστους, της απλότητας και της σύντομης διάρκειας της διαδικασίας.

Η έγκαιρη και σωστή διάγνωση αυτής της μορφής καρκίνου μπορεί να οδηγήσει σε καλύτερα προγράμματα θεραπείας και να μειώσει δραστικά την θνησιμότητα του. Η διάγνωση τέτοιων εικόνων από παθολογοανατόμους όμως παρουσιάζει σημαντικές δυσκολίες για τους εξής λόγους: (α) είναι πολύ δύσκολο νέοι ιατροί του χώρου να αποκτήσουν τις απαιτούμενες γνώσεις και εμπειρία με αποτέλεσμα να υπάρχει έλλειψη κατάλληλου προσωπικού, (β) η διάγνωση (σε αντίθεση με την δημιουργία) των stains είναι μια χρονοβόρα διαδικασία και κοστίζει, (γ) ελλοχεύει ο κίνδυνος του ανθρώπινου λάθους λόγω κόπωσης. Είναι λοιπόν επιτακτική η ανάγκη ύπαρξης συστημάτων υποβοήθησης για την διάγνωση τέτοιων εικόνων.

1.2. Τα Σύνολα Δεδομένων

Τα σύνολα δεδομένων που θα μας απασχολήσουν στην εργασία είναι δυο, το Breast Cancer Cell (BCC) του University of California Santa Barbara (UCSB) [3] και το Breast Cancer Histopathological Database (BreaKHis) του Federal University of Parana (UFPR) [4]. Και τα δυο περιέχουν H&E ιστοπαθολογικές εικόνες νεοπλασιών του μαστού με ετικέτες που μας λένε αν η κάθε εικόνα αντιστοιχεί σε καλοήθεια ή κακοήθεια (καρκίνωμα).

Τα δεδομένα αυτά είναι ιδανικά για χρήση με μεθόδους μάθησης πολλαπλών στιγμιότυπων καθώς η πληροφορία που διαφοροποιεί τους καλοήθεις όγκους από τα καρκινώματα δεν βρίσκεται σε ολόκληρη την εικόνα ενός ιστού αλλά σε μέρη αυτής. Στην παρακάτω εικόνα αυτό γίνεται εμφανές, με τα σημεία ενδιαφέροντος να περικλείονται από τα κίτρινα κουτιά.



Σχήμα 1.1: Καρκινώματα σε H&E stains σημειωμένα από παθολογοανατόμους.

Όταν τα σημεία ενδιαφέροντος καταλαμβάνουν λίγο χώρο σε σχέση με το συνολικό μέγεθος της εικόνας η πληροφορία που κατέχουν είναι πολύ εύκολο να «χαθεί», με την υπόλοιπη «άχρηστη» πληροφορία να υπερισχύει κατά την εξαγωγή των χαρακτηριστικών. Κατακερματίζοντας τις εικόνες σε μικρότερα κομμάτια φέρνουμε στο προσκήνιο πιθανά τέτοια σημεία πριν την εξαγωγή των χαρακτηριστικών, ταυτόχρονα μετατρέποντας το πρόβλημα σε πολλαπλών στιγμιότυπων. Μετά την κατηγοριοποίηση αυτών, με την χρήση ευέλικτων κανόνων σύντηξης των αποτελεσμάτων μπορούμε να βγάλουμε συμπεράσματα για τις αρχικές εικόνες με μεγαλύτερο ποσοστό επιτυχίας. Η μάθηση πολλαπλών στιγμιότυπων αναλύεται περισσότερο στην ενότητα 2.4.

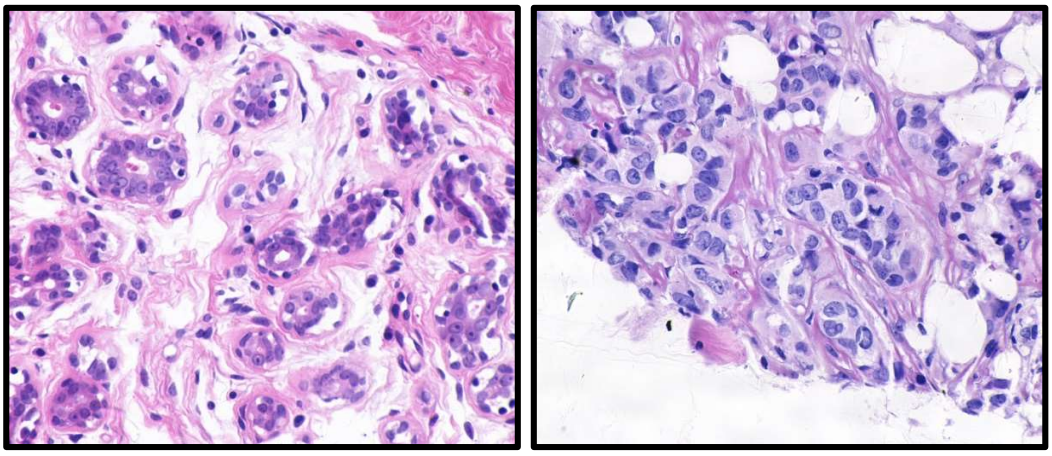
1.2.1. Breast Cancer Cell – BCC

Αυτό το σύνολο δεδομένων περιέχει συνολικά 58 εικόνες από 10 ασθενείς, 9 που πάσχουν από καρκίνο και 1 που έχει κάποια καλοήθη νεοπλασία. Κάθε εικόνα έχει μια ετικέτα καλοήθειας ή κακοήθειας. Είναι προφανές πως ένα τόσο ανισόρροπο και μικρό σύνολο δεδομένων δεν μπορεί να χρησιμοποιηθεί για την αξιολόγηση οποιουδήποτε αλγορίθμου σε επίπεδο ασθενή. Γι' αυτόν τον λόγο θα το χρησιμοποιήσουμε μόνο για την αξιολόγηση της κατηγοριοποίησης σε επίπεδο εικόνας.

Καλοήθεις	Κακοήθεις	Σύνολο
32 (1)	26 (9)	58 (10)

Πίνακας 1.1: Κατανομή εικόνων στο BCC (σε παρένθεση οι ασθενείς).

Είναι απαραίτητο λοιπόν εφόσον το σύστημα μας είναι πολλαπλών στιγμιότυπων να δημιουργήσουμε ένα πλήθος στιγμιότυπων ανά εικόνα. Αυτό το πετύχαμε με την χρήση κατακερματισμού (segmentation) όπως αυτός εξηγείται στην ενότητα 4.2. Το αποτέλεσμα είναι οι αριθμοί του πίνακα 1.1 να πολλαπλασιάζονται ανάλογα με την κλίμακα του κατακερματισμού. Για παράδειγμα με χρήση κατακερματισμού 5x5 προκύπτουν 1450 στιγμιότυπα.



Σχήμα 1.2: Δείγματα από το BCC. Καλοήθεια (αριστερά) και καρκίνωμα (δεξιά).

1.2.2. Breast Cancer Histopathological Database – BreaKHis

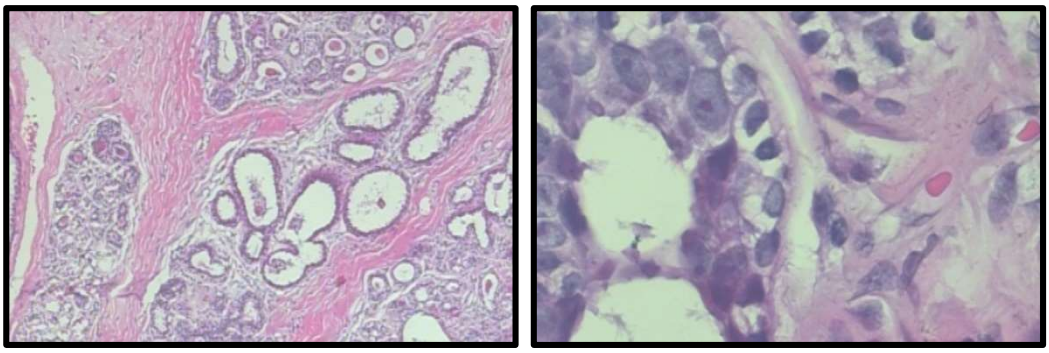
Σε αντίθεση με το προηγούμενο σύνολο δεδομένων που είναι μικρό αυτό περιέχει συνολικά 7909 εικόνες από 82 ασθενείς, 58 που πάσχουν από καρκίνο και 24 που έχουν κάποια καλοήθη νεοπλασία. Οι εικόνες είναι χωρισμένες ανάλογα με την μεγέθυνση του μικροσκοπίου σε 4 κατηγορίες: 40x, 100x, 200x και 400x.

Μεγέθυνση	Καλοήθεις	Κακοήθεις	Σύνολο
40x	652 (24)	1370 (58)	1995 (82)
100x	644 (24)	1437 (58)	2081 (82)
200x	623 (24)	1390 (58)	2013 (82)
400x	588 (24)	1232 (58)	1820 (82)
Σύνολο	2480 (24)	5429 (58)	7909 (82)

Πίνακας 1.2: Κατανομή εικόνων στο BreaKHis (σε παρένθεση οι ασθενείς).

Έχοντας κατά μέσο όρο 24 εικόνες με ετικέτα ανά ασθενή αυτό το σύνολο δεδομένων είναι ιδανικό για χρήση σε συστήματα πολλαπλών στιγμιότυπων χωρίς να κρίνεται αναγκαία η χρήση κατακερματισμού. Λόγω περιορισμών χρόνου το BreaKHis δοκιμάστηκε μόνο σε επίπεδο ασθενή και όχι σε επίπεδο εικόνας, δηλαδή δεν έγινε κατακερματισμός των εικόνων.

Επίσης οι ετικέτες αυτού του συνόλου δεδομένων δεν περιορίζονται μόνο σε καλοήθεις και κακοήθεις αλλά επεκτείνονται σε 4 υποκατηγορίες για την κάθε περίπτωση, συνολικά δηλαδή αντιστοιχίζονται σε 8 διαφορετικές κλάσεις. Ο αλγόριθμος μας δεν δοκιμάστηκε στο πρόβλημα των 8 κλάσεων καθώς κάτι τέτοιο θα αποτελούσε πιθανώς μια δεύτερη εργασία.



Σχήμα 1.3: Δείγματα από το BreaKHis, 40x (αριστερά) και 400x (δεξιά).

1.3. Προσέγγιση και Δομή της Εργασίας

Το πρώτο πρόβλημα που κληθήκαμε να αντιμετωπίσουμε ήταν αυτό της διαφορετικότητας στα χρώματα μεταξύ των ιστοπαθολογικών εικόνων [5]. Όπως εξηγείται στην ενότητα 4.1, υπάρχει ένα σύνολο από παράγοντες όπως οι διαφορές στις πρώτες ύλες (στις βαφές), στο μικροσκόπιο που τράβηξε την φωτογραφία, στον φωτισμό του εργαστηρίου κ.ο.κ. οι οποίοι οδηγούν σε σημαντικές διαφορές στα χρώματα των εικόνων. Αυτό το φαινόμενο μπορεί να δυσκολέψει έναν παθολογοανατόμο και πόσο μάλλον ένα αυτόματο σύστημα κατηγοριοποίησης. Στην ίδια ενότητα αναφέρουμε μια δημοσίευση η οποία παρουσιάζει μια σύγχρονη λύση για αυτό το πρόβλημα καθώς επίσης και τον κώδικα που χρησιμοποιήσαμε στην παρούσα εργασία.

Στην συνέχεια έπρεπε να επιλέξουμε μια μέθοδο για την εξαγωγή χαρακτηριστικών. Ένα σύστημα κατηγοριοποίησης εικόνων (εκτός από αυτά που χρησιμοποιούν νευρωνικά δίκτυα βαθιάς εκμάθησης) πρέπει να διαθέτει μια μέθοδο με την οποία θα εξάγει χαρακτηριστικά από αυτές, τα οποία στην συνέχεια θα τροφοδοτούνται σε κάποιο μοντέλο μηχανικής μάθησης. Η συνήθης λύση είναι η χρήση περιγραφέων όπως οι Parameter Free Threshold Adjacency Statistics (PFTAS) [6]-[7], Scale Invariant Feature Transform (SIFT) [8], Gray Level Co-occurrence Matrix (GLCM) [9], Histogram of Oriented Gradients (HOG) [10], κ.ο.κ.. Η μέθοδος που επιλέξαμε κάνει χρήση της γνωστής παραγοντοποίησης για τανυστές Canonical Polyadic Decomposition (CPD) ή αλλιώς PARAFAC ή CANDECOMP [11]. Όπως εξηγούμε και σε βάθος στην ενότητα 4.3, μέσω αυτής της παραγοντοποίησης ενός τανυστή που περιέχει τις εικόνες του προβλήματος δημιουργούμε ένα «λεξικό» μονόχρωμων φιγούρων και κάθε εικόνα εκφράζεται ως ένας γραμμικός συνδυασμός αυτών. Οι συντελεστές αυτών των γραμμικών συνδυασμών είναι και τα χαρακτηριστικά που εξάγει η μέθοδος μας.

Το επόμενο κομμάτι της εργασίας ήταν η δημιουργία ενός multiple instance learning (MIL) κατηγοριοποιητή που ανήκει στην κατηγορία των instance space [12] αλγορίθμων. Όπως εξηγούμε στην υποενότητα 2.4.1 ένας τέτοιος κατηγοριοποιητής αποτελείται από έναν ή και περισσότερους κατηγοριοποιητές σε επίπεδο στιγμιότυπων και από κάποια μέθοδο σύντηξης των αποτελεσμάτων του/τους για την δημιουργία συμπερασμάτων σε επίπεδο αντικειμένων. Η δική μας μέθοδος δημιουργεί τόσους κατηγοριοποιητές σε επίπεδο στιγμιότυπων όσες είναι και οι κλάσεις του προβλήματος (δοκιμάστηκε μόνο σε δυαδικά προβλήματα) και αναπτύχθηκαν τρεις μετρικές επιπέδου αντικειμένων για την σύντηξη των αποτελεσμάτων τους. Το βασικό μας εργαλείο είναι τα one-class SVM [13] που αναφέρονται και στο κεφάλαιο 3 και ο κατηγοριοποιητής μας αναλύεται στο κεφάλαιο 5.

Στο πλαίσιο ανάπτυξης του παραπάνω κατηγοριοποιητή μελετήθηκαν επίσης η χρήση της Principal Component Analysis (PCA) [14] για την μείωση της διαστατικότητας των δεδομένων και η βαθμονόμηση των τιμών απόφασης που προκύπτουν από τα one-class SVM ώστε να μπορέσουν τα αποτελέσματα τους να γίνουν συγκρίσιμα.

Στην είσοδο του κατηγοριοποιητή μας υπάρχουν αρκετές υπερ-παράμετροι τις οποίες βελτιστοποιούμε κάνοντας χρήση ενός Bayesian optimizer [15]. Για την αξιολόγηση της επίδοσης του αλγορίθμου που αναπτύξαμε έπρεπε να χρησιμοποιήσουμε μια περίπλοκη μορφή cross validation ώστε οι υπερ-παράμετροι αυτές να βελτιστοποιούνται ξεχωριστά για κάθε fold. Η μέθοδος που ακολουθήσαμε για αυτόν τον σκοπό είναι αυτή του εμφωλευμένου cross validation (nested cross validation) που εκτελεί ένα επιπλέον cross validation στο training set του κάθε fold. Ο τρόπος λειτουργίας εξηγείται περισσότερο στην ενότητα 4.6.

Τέλος, στο κεφάλαιο 6 τρέξαμε μια σειρά από πειράματα πάνω στα σύνολα δεδομένων που αναφέραμε προηγουμένως και συγκρίναμε τα αποτελέσματα μας με αντίστοιχα αποτελέσματα από την βιβλιογραφία.

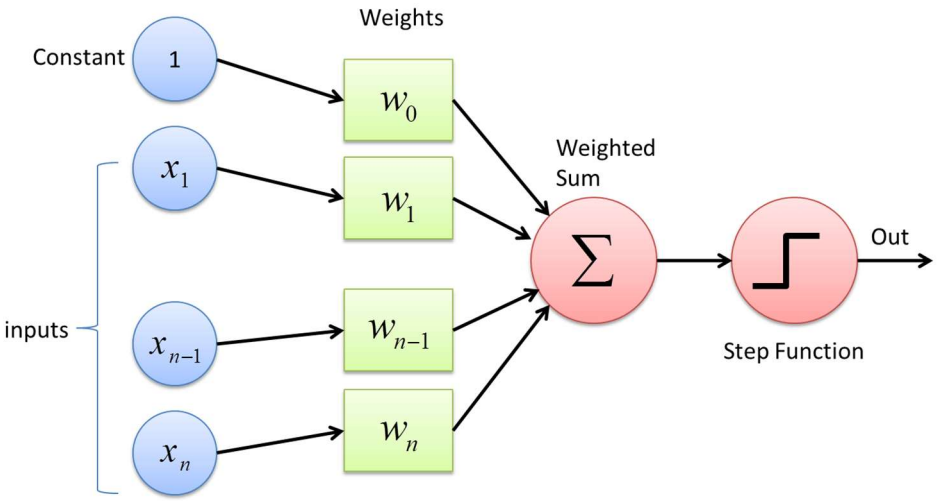
ΚΕΦΑΛΑΙΟ 2: ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

2.1. Γενικά

Η μηχανική μάθηση (machine learning) είναι ένα πεδίο το οποίο ανήκει στην τεχνητή νοημοσύνη. Ενώ δεν υπάρχει σαφής ορισμός για το τι εστί τεχνητή νοημοσύνη, η μηχανική μάθηση είναι καλά ορισμένη. Όπως μας προϊδεάζει το όνομα, αφορά αλγορίθμους μέσω των οποίων μια μηχανή μπορεί και μαθαίνει από δεδομένα.

Πολλά προβλήματα είναι δύσκολο ή και χρονοβόρο να λυθούν με την χρήση αυστηρά ορισμένων μαθηματικών μοντέλων. Η μηχανική μάθηση, χωρίς την χρήση ρητά προκαθορισμένων κανόνων ή μοντέλων, αλλά με τον εντοπισμό μοτίβων στα δεδομένα, μπορεί να δώσει λύσεις σε τέτοια προβλήματα πολύ γρήγορα και με ικανοποιητική ακρίβεια. Ένας αλγόριθμος μηχανικής μάθησης επομένως για να δουλέψει χρειάζεται έναν σχετικά μεγάλο όγκο δεδομένων, πράγμα που στις μέρες μας δεν αποτελεί πρόβλημα αφού τα ψηφιακά αρχεία και αισθητήρες κάθε είδους συναντώνται κυριολεκτικά παντού και φυσικά με την βοήθεια του διαδικτύου η πρόσβαση σε μεγάλους όγκους δεδομένων δεν ήταν ποτέ ευκολότερη.

Τα δεδομένα σαν σύνολο πρέπει να αποτελούνται από διανύσματα χαρακτηριστικών (feature vectors), ένα ανά αντικείμενο. Αν μιλούσαμε για κρασιά π.χ. ένα διάνυσμα χαρακτηριστικών θα μπορούσε να περιέχει την απόχρωση του κρασιού, την περιεκτικότητα του σε αλκοόλ, και ό,τι άλλο θα επιθυμούσαμε να συμπεριλάβουμε.



Σχήμα 2.1: Η βασική δομή ενός single-layer Perceptron.

Ας πάρουμε για παράδειγμα την βασική δομή ενός δικτύου single-layer Perceptron που αποτελεί έναν δυαδικό κατηγοριοποιητή. Όπως φαίνεται στο σχήμα 2.1 το δίκτυο αποτελείται από n+1 εισόδους, τις x_1 έως x_n στις οποίες εισάγονται τα διανύσματα χαρακτηριστικών και μια σταθερά. Κάθε μια από αυτές συνοδεύεται από ένα βάρος w_o έως w_n τα οποία προκύπτουν όλα μαζί μέσω επαναληπτικής εκπαίδευσης με ανατροφοδότηση και δρουν σαν συντελεστές για τις εισόδους. Τα αποτελέσματα των πολλαπλασιασμών των εισόδων με τα βάρη τους αθροίζονται σε μια τιμή (συνεχή μεταβλητή) η οποία στην συνέχεια μετατρέπεται σε 0 ή 1 ανάλογα με την βηματική συνάρτηση που θα χρησιμοποιηθεί πριν την έξοδο. Το συγκεκριμένο μοντέλο περιορίζεται σε δυο τιμές εξόδου, 0 και 1, πράγμα που το καθιστά κατάλληλο μόνο για πολύ αυστηρή κατηγοριοποίηση. Οι νευρώνες που χρησιμοποιούνται στα μοντέρνα νευρωνικά δίκτυα διαφέρουν με το single-layer Perceptron στην βηματική συνάρτηση. Ενώ το Perceptron μετατρέπει την τιμή του αθροίσματος σε 0 ή 1 μέσω μιας βηματικής συνάρτησης, στους σύγχρονους νευρώνες μπορούμε να διαλέξουμε ανάμεσα σε πολλές δοκιμασμένες συναρτήσεις «ενεργοποίησης».

Με την χρήση τέτοιων αλγορίθμων μπορεί να επιτευχθεί η πρόβλεψη κάποιας μετρικής ή η κατηγοριοποίηση ενός συνόλου αντικειμένων, γι' αυτό και ο συγκεκριμένος κλάδος είναι πολύ στενά συνδεδεμένος με την διαδικασία της λήψης αποφάσεων (decision making). Μιλώντας πάλι για κρασιά, ένα παράδειγμα θα μπορούσε να είναι η πρόβλεψη κάποιας ποιότητας ενός βαρελιού έναν χρόνο μετά την έναρξη της ωρίμανσης, σύμφωνα με δεδομένα παλαιότερων χρόνων. Οι τομείς στους οποίους μπορεί να χρησιμεύσει η μηχανική μάθηση είναι πάρα πολλοί, όπως η γεωπονία, η οικονομία, η υγεία, η ρομποτική, κ.ο.κ.. Είναι επομένως κρίσιμης σημασίας για πολλές επιχειρήσεις αλλά και υπηρεσίες με άπειρες εφαρμογές σε πραγματικά προβλήματα.

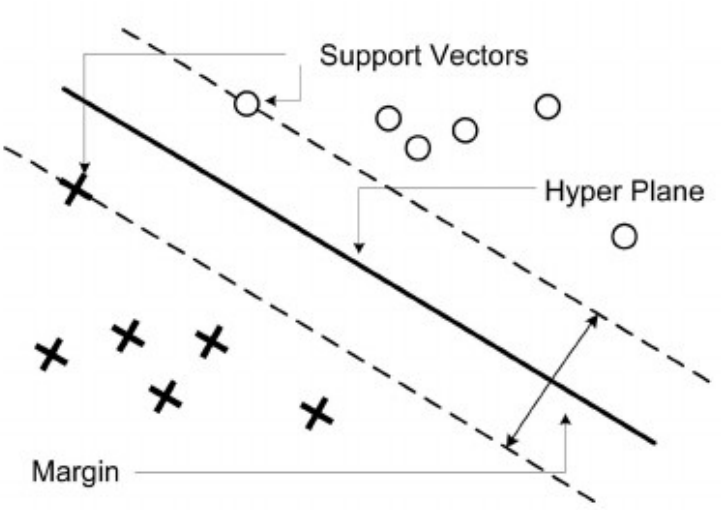
2.2. Εποπτευόμενη Μάθηση

Η πιο κλασσική μορφή μηχανικής μάθησης είναι η εποπτευόμενη μάθηση (supervised learning). Τα προβλήματα στα οποία απευθύνεται η εποπτευόμενη μάθηση είναι δυο: (α) ο υπολογισμός μιας συνεχούς τιμής βάσει κάποιων ανεξάρτητων μεταβλητών ή αλλιώς «παλινδρόμηση» (regression) και (β) η κατηγοριοποίηση αντικειμένων (classification) χρησιμοποιώντας ένα σύνολο προκαθορισμένων ετικετών (labels).

Σε αυτήν την μορφή μηχανικής μάθησης τα διανύσματα χαρακτηριστικών των δεδομένων εκπαίδευσης συνοδεύονται από ετικέτες που θεωρούνται αληθείς (ground truth). Για παράδειγμα, αν ο σκοπός ενός αλγορίθμου είναι η αυτόματη αναγνώριση χειρόγραφων χαρακτήρων από εικόνες, τα δεδομένα εκπαίδευσης πρέπει να είναι πολλές τέτοιες εικόνες από χειρόγραφους χαρακτήρες. Σε αυτή την περίπτωση τα διανύσματα χαρακτηριστικών θα ήταν οι ίδιες οι εικόνες σε μορφή pixels ή κάποια χαρακτηριστικά εξαγμένα από αυτές και οι ετικέτες θα ήταν ο χαρακτήρας που απεικονίζεται από κάθε εικόνα.

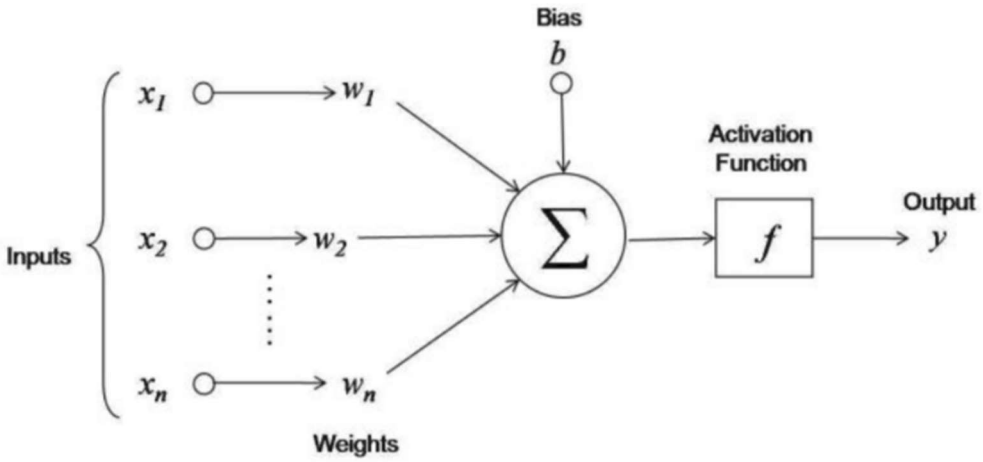
Αυτοί οι αλγόριθμοι, μέσω μιας επαναληπτικής διαδικασίας μάθησης δημιουργούν ένα σύνολο από κανόνες - που ως επί το πλείστον δεν είναι κατανοητοί από εμάς - και τους εφαρμόζουν επάνω σε νέα διανύσματα χαρακτηριστικών για να δώσουν μια εκτίμηση.

Ένα μοντέλο εποπτευόμενης μάθησης που χρησιμοποιείται κατά κόρων είναι το SVM (Support Vector Machine). Αυτό είναι και το βασικό εργαλείο που χρησιμοποιήσαμε σε αυτήν την εργασία. Στην βασική του μορφή το SVM είναι ένας δυαδικός κατηγοριοποιητής ο οποίος ορίζει ένα όριο απόφασης (decision boundary) μέσω της εκπαίδευσης, το οποίο δεν είναι κάτι άλλο παρά ένα υπερ-επίπεδο που χωρίζει τον διανυσματικό χώρο των διανυσμάτων χαρακτηριστικών στα δυο. Ύστερα, προβάλλοντας νέα διανύσματα χαρακτηριστικών σε αυτόν τον χώρο, ανάλογα με την μεριά του υπερ-επιπέδου (ορίου απόφασης) στην οποία ανήκουν γίνεται μια εκτίμηση για την κατηγορίας στην οποία ανήκει το καθένα. Το SVM και οι μορφές του θα αναλυθούν περισσότερο στο κεφάλαιο 3.

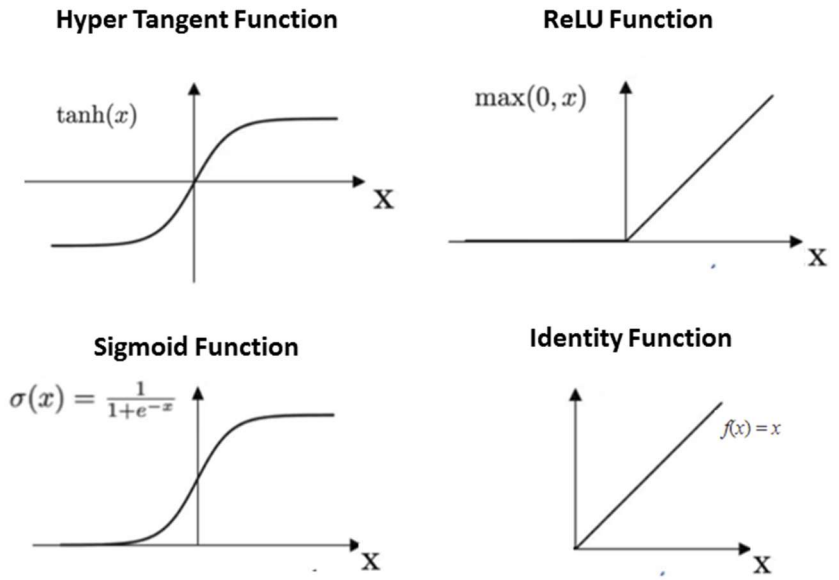


Σχήμα 2.2: Τα βασικά στοιχεία ενός linear soft-margin SVM.

Ακόμα ένα πολύ σημαντικό μοντέλο εποπτευόμενης μάθησης είναι τα νευρωνικά δίκτυα. Θα μιλήσουμε για την απλή τους μορφή και δεν θα επεκταθούμε στα νευρωνικά δίκτυα βαθιάς εκμάθησης τα οποία μπορούν να περάσουν και στην κατηγορία της μη-εποπτευόμενης μάθησης. Τα νευρωνικά δίκτυα αποτελούνται, όπως μας λέει και το όνομα τους, από τεχνητούς νευρώνες. Κάθε νευρώνας μπορεί να παρομοιαστεί με το single-layer Perceptron που περιγράψαμε στην ενότητα 2.1. Η διαφορά τους όπως αναφέρθηκε και νωρίτερα βρίσκεται στην συνάρτηση πριν την έξοδο, με το Perceptron να έχει αποκλειστικά την βηματική συνάρτηση ενώ οι νευρώνες των νευρωνικών δικτύων υποστηρίζουν οποιαδήποτε συνάρτηση ενεργοποίησης.

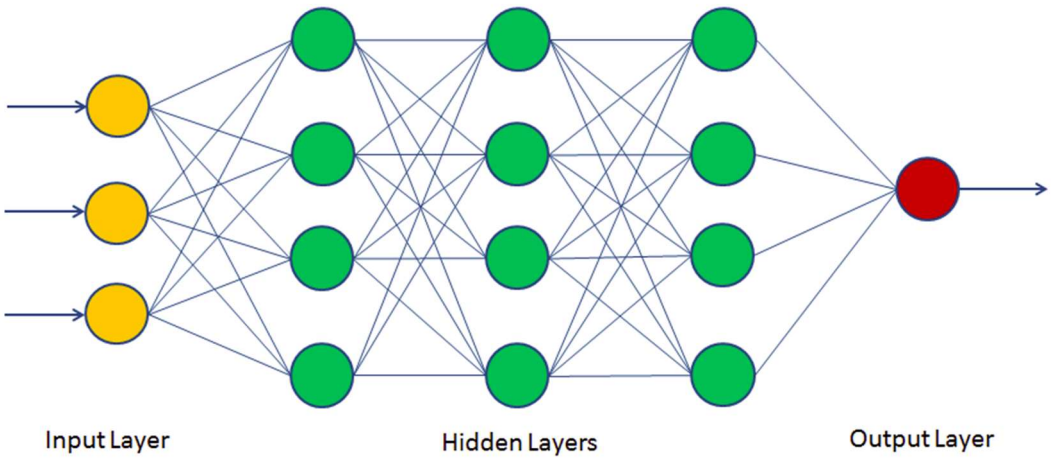


Σχήμα 2.3: Η δομή ενός νευρώνα από νευρωνικό δίκτυο.



Σχήμα 2.4: Μερικές συναρτήσεις ενεργοποίησης για νευρώνες.

Τα νευρωνικά δίκτυα έχουν τους νευρώνες τους οργανωμένους σε επίπεδα. Συνήθως αποτελούνται από ένα επίπεδο εισόδου, ένα επίπεδο εξόδου και έναν μικρό αριθμό (αλλιώς μιλάμε για βαθιά εκμάθηση) από «κρυφά» επίπεδα τα οποία βρίσκονται ανάμεσα στα άλλα δυο. Οι νευρώνες εισόδου έχουν μια είσοδο ενώ οι νευρώνες εξόδου μια έξοδο. Κάθε νευρώνας (εκτός από αυτούς της εισόδου) δέχεται στην είσοδο του τις εξόδους ενός αριθμού από νευρώνων του προηγούμενου επιπέδου, ανάλογα με την αρχιτεκτονική του ενίοτε δικτύου.



Σχήμα 2.5: Η δομή ενός μικρού νευρωνικού δικτύου.

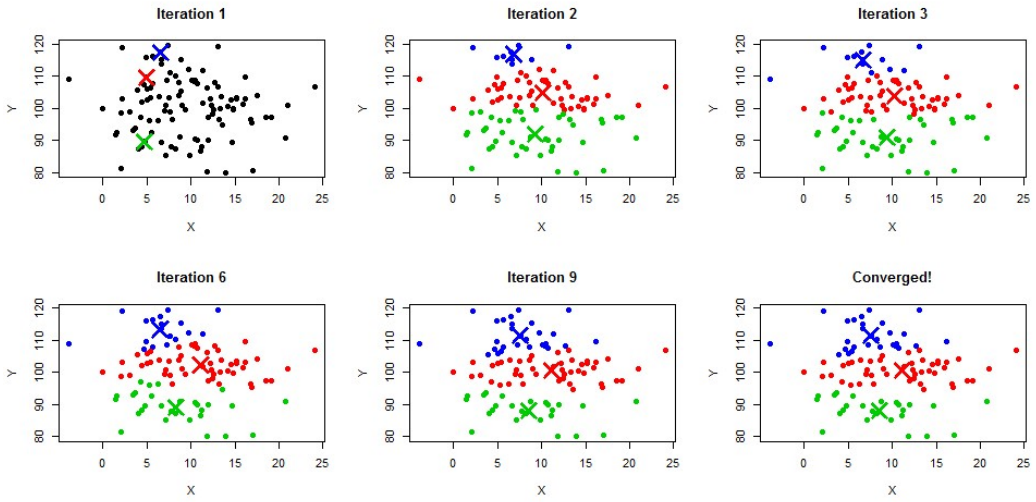
Στο παραπάνω σχήμα φαίνεται ένα πλήρως συνδεδεμένο νευρωνικό δίκτυο, δηλαδή στο οποίο κάθε νευρώνας τροφοδοτείται από τις εξόδους όλων των νευρώνων του προηγούμενου επιπέδου. Κάθε νευρώνας έχει τα δικά του βάρη στις εισόδους, το δικό του bias και την δική του συνάρτηση ενεργοποίησης. Τα βάρη και τα bias είναι προϊόντα της εκπαίδευσης του δικτύου. Στην επιβλεπόμενη μάθηση τα νευρωνικά δίκτυα εκπαιδεύονται με την πίσω διάδοση λάθους η οποία εμπεριέχει αρκετά μαθηματικά και δεν θα αναλυθεί εδώ. Αυτό που προσπαθεί να κάνει είναι να βρει μέσω κάποιων παραγώγων το «μερίδιο» της ευθύνης κάθε νευρώνα για το λάθος στην έξοδο και να κάνει μια διόρθωση στα βάρη του ανάλογα με το μέγεθος αυτού.

Εκτός από τα SVM και τα νευρωνικά δίκτυα που μόλις περιγράψαμε, φυσικά υπάρχουν και άλλα μοντέλα εποπτευόμενης μάθησης όπως τα Linear/Logistic Regression, Naive Bayes, Linear Discriminant Analysis, Decision Trees και k-NN (k-nearest neighbors). Αυτή η μορφή μηχανικής μάθησης είναι που θα μας απασχολήσει στην παρούσα εργασία και συγκεκριμένα ένας τύπος αυτής, η μάθηση πολλαπλών στιγμιότυπων (multiple instance learning) που θα αναλυθεί στην ενότητα 2.4.

2.3. Μη-Εποπτευόμενη Μάθηση

Στην μη-εποπτευόμενη μάθηση (un-supervised learning) τα δεδομένα που υπάρχουν στην διάθεση μας δεν έχουν ετικέτες. Αυτό συμβαίνει συχνά σε προβλήματα που περιγράφονται από μεγάλο όγκο δεδομένων, όπου στις περισσότερες περιπτώσεις δεν υπάρχει η δυνατότητα της ανθρώπινης παρέμβασης για την δημιουργία τους λόγω χρόνου ή και κόστους.

Το πρόβλημα που καλείται να λύσει η μη-εποπτευόμενη μάθηση είναι αυτό της συσταδοποίησης (clustering). Ο στόχος είναι να βρεθεί ένα βέλτιστο σύνολο από συστάδες στις οποίες να μπορούν να καταταχθούν τα αντικείμενα που έχουν μεγάλη ομοιότητα - ως προς κάποιο/α χαρακτηριστικό/ά - μεταξύ τους. Το πλήθος αυτών των συστάδων μπορεί να είναι ρητά προκαθορισμένο ή και όχι. Επιπλέον, μαζί με την συσταδοποίηση συχνά γίνεται και ο εντοπισμός – αφαίρεση ακραίων τιμών (outliers), αντικειμένων δηλαδή που περισσότερο μπερδεύουν ένα σύστημα παρά προσφέρουν χρήσιμη πληροφορία.



Σχήμα 2.6: Παράδειγμα χρήσης του k-means σε βήματα.

Ίσως ο πιο γνωστός και ευρέως χρησιμοποιούμενος αλγόριθμος μη-εποπτευόμενης μάθησης για συσταδοποίηση είναι ο k-means (διαφορετικός από τον k-nearest neighbors / k-NN). Ο αλγόριθμος αυτός ανήκει στην κατηγορία των αλγορίθμων συσταδοποίησης που παίρνουν σαν παράμετρο τον αριθμό k των συστάδων από τον χρήστη. Αρχικά ορίζει τυχαία k από τα αντικείμενα ως «κέντρα» στον διανυσματικό χώρο των διανυσμάτων χαρακτηριστικών και στην συνέχεια εκτελεί επαναληπτικά δυο βήματα έως ότου να συγκλίνει: (α) κατάταξη κάθε αντικειμένου στο κοντινότερο κέντρο σύμφωνα με την Ευκλείδεια

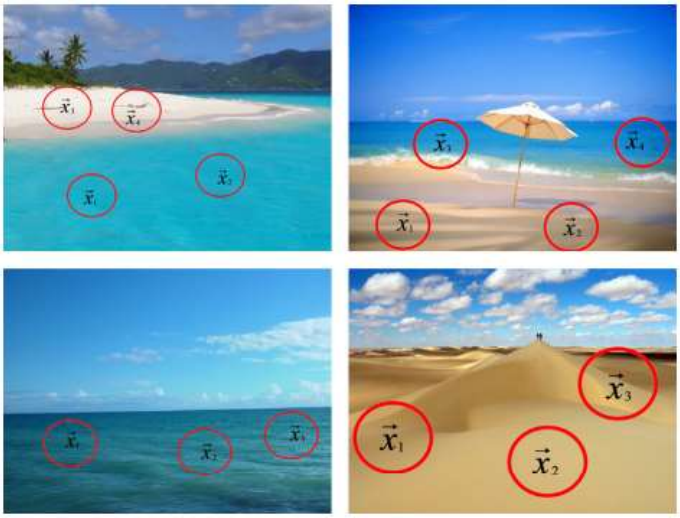
απόσταση και (β) επανυπολογισμός του κάθε κέντρου ως το μέσο των διανυσμάτων που ανήκουν σε αυτό.

Ένα παράδειγμα στο οποίο η συσταδοποίηση μπορεί να φανεί χρήσιμη είναι τα διαδικτυακά μαγαζιά. Με τα αμέτρητα δεδομένα που μπορεί να μαζέψει ένα τέτοιο μαγαζί για τους πελάτες του μπορεί μέσω της συσταδοποίησης τους να προβάλλει προτάσεις για σχετικά προϊόντα και στοχευμένες διαφημίσεις που να είναι όντως ελκυστικές.

Οι αλγόριθμοι μη-εποπτευόμενης μάθησης δεν θα μας απασχολήσουν στην διπλωματική εργασία αυτή.

2.4. Μάθηση Πολλαπλών Στιγμιότυπων

Η μάθηση πολλαπλών στιγμιότυπων είναι ένας τύπος εποπτευόμενης μάθησης με την εξής ιδιαιτερότητα: τα δεδομένα έρχονται σε μορφή αντικειμένων (bags) που αποτελούνται από πολλαπλά στιγμιότυπα (instances) και υπάρχουν ετικέτες μόνο για τα αντικείμενα σαν ενιαίες οντότητες. Με λίγα λόγια για κάθε αντικείμενο αντί για ένα feature vector υπάρχουν περισσότερα.



Σχήμα 2.7: Παράδειγμα πολλαπλών στιγμιότυπων.

Στον χώρο της κατηγοριοποίησης με χρήση αλγορίθμων μηχανικής μάθησης υπάρχει η εξής δυσκολία που είναι πιο εύκολα κατανοητή στις εφαρμογές που αφορούν εικόνες. Πολλές φορές, η πληροφορία με βάση την οποία είναι επιθυμητό να γίνει η κατηγοριοποίηση βρίσκεται σε ένα μικρό κομμάτι του αντικειμένου – για παράδειγμα σε ένα μικρό μέρος μιας εικόνας – με το υπόλοιπο να περιέχει πληροφορία που δεν μας αφορά. Αυτό έχει σαν αποτέλεσμα οι κλασσικοί αλγόριθμοι που παίρνουν ολόκληρα τα αντικείμενα σαν ενιαία στιγμιότυπα να δίνουν ίδια ή και περισσότερη σημασία σε πληροφορία που δεν θα έπρεπε, όπως για παράδειγμα στο φόντο. Είναι επιτακτική επομένως η ανάγκη να μπορεί ένας αλγόριθμος να μεταχειρίζεται τα αντικείμενα με έναν πιο ευέλικτο τρόπο.

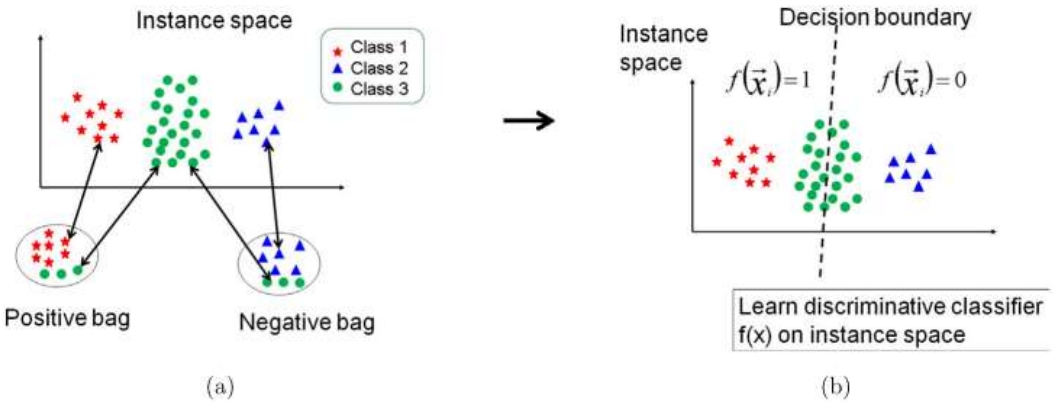
Στο σχήμα 2.7 είναι ένα παράδειγμα χρήσης πολλαπλών στιγμιότυπων για κατηγοριοποίηση εικόνας. Στις πάνω δυο εικόνες απεικονίζονται δυο παραλίες ενώ κάτω φαίνεται ένας ωκεανός και μια έρημος. Αυτό που ξεχωρίζει κυρίως τις παραλίες από τα άλλα δυο είναι η ύπαρξη θάλασσας μαζί με άμμο. Οι αλγόριθμοι μηχανικής μάθησης χρειάζονται χαρακτηριστικά τα οποία συνήθως,

όπως αναφέρουμε στην ενότητα 4.3, εξάγονται με την χρήση περιγραφέων. Ένας περιγραφέας είναι πολύ δύσκολο να περιγράψει την ύπαρξη δυο συγκεκριμένων διαφορετικών οντοτήτων σε μια εικόνα. Το αποτέλεσμα; Πολύ πιθανό είναι να κατηγοριοποιούνταν όλες οι εικόνες ως παραλίες. Εξάγοντας πολλά feature vectors με περιγραφείς από διαφορετικά σημεία κάθε εικόνας μπορούμε να κατηγοριοποιήσουμε ξεχωριστά αυτά τα σημεία. Σε μια τέτοια περίπτωση, ένας κατηγοριοποιητής MIL θα απαιτούσε την ύπαρξη και θάλασσας και άμμου, που θα μπορούσε όμως να αναγνωρίσει με μεγαλύτερη ευκολία ξεχωριστά.

Ο πρώτος που διατύπωσε μια ιεραρχία για τους αλγορίθμους πολλαπλών στιγμιότυπων ήταν ο J. Amores [12] το 2013. Για λόγους απλότητας θεωρούμε δυαδική κατηγοριοποίηση, δηλαδή θετικό – αρνητικό. Σύμφωνα με την δουλειά του, οι αλγόριθμοι αυτοί μπορούν να χωριστούν σε τρεις κατηγορίες: (α) instance space, (β) bag space και (γ) embedded space.

2.4.1. Instance Space

Στην κατηγορία του Instance Space (IS), οι αλγόριθμοι θεωρούν πως η χρήσιμη πληροφορία για την κατηγοριοποίηση βρίσκεται αποκλειστικά σε κάποια από τα επιμέρους στιγμιότυπα και όχι στα αντικείμενα συνολικά. Όλα τα στιγμιότυπα κληρονομούν την ετικέτα του αντικειμένου στο οποίο ανήκουν και δημιουργείται ένας κατηγοριοποιητής (ή και περισσότεροι), έστω $f(\vec{x}) \in [0,1]$, επιπέδου στιγμιότυπων που εκτιμά την πιθανότητα ένα στιγμιότυπο να είναι θετικό.



Σχήμα 2.8: Παράδειγμα Instance Space μοντέλου.

Κατά την πρόβλεψη κάθε στιγμιότυπο λαμβάνει μια πιθανότητα θετικότητας $f(\overrightarrow{x_i})$ από τον παραπάνω κατηγοριοποιητή και ύστερα από αυτές τις τιμές πρέπει να προκύψει μία τιμή $F(X) \in [0,1]$ για το αντικείμενο, που να υποδηλώνει την πιθανότητα αυτό να είναι θετικό. Οι αλγόριθμοι αυτοί διαφοροποιούνται μεταξύ τους σε μια υπόθεση (assumption) που πρέπει να κάνουν, η οποία αφορά τον τρόπο σύντηξης των πιθανοτήτων $f(\overrightarrow{x_i})$ ώστε να προκύψει η $F(X)$ για το αντικείμενο με το οποίο συνδέονται. Υπάρχουν πολλές υποθέσεις που μπορούν να γίνουν. Οι κύριες από αυτές είναι οι ακόλουθες.

- <u>Standard (SMI) Assumption:</u>

Αν έστω και ένα στιγμιότυπο θεωρηθεί θετικό τότε και το αντικείμενο θεωρείται θετικό. Με αλλά λόγια, η πιθανότητα ένα αντικείμενο να είναι θετικό θεωρείται ίση με την μέγιστη πιθανότητα θετικότητας των επιμέρους στιγμιότυπων του. Συνεπώς δεν συνεισφέρουν όλα τα στιγμιότυπα στην

απόφαση για το αντικείμενο αλλά μόνο ένα. Η υπόθεση αυτή είναι ασύμμετρη, που σημαίνει πως αν αντιστραφούν οι σημασίες των ετικετών το αποτέλεσμα θα είναι διαφορετικό. Περιγράφεται από την σχέση:

$$F(X) = \max_{\vec{x} \in X} f(\vec{x}) \qquad (2.1)$$

- Collective Assumption:

Αυτή η υπόθεση θεωρεί πως όλα τα στιγμιότυπα ενός αντικειμένου πρέπει να συνεισφέρουν με τον ίδιο βαθμό στο συμπέρασμα για ολόκληρο το αντικείμενο. Δεν είναι κάτι άλλο παρά ο αριθμητικός μέσος των $f(\overrightarrow{x_i})$ όπως φαίνεται από τον ακόλουθο τύπο:

$$F(X) = \frac{1}{|X|} \sum_{\vec{x} \in X} f(\vec{x}) \qquad (2.2)$$

Η υπόθεση αυτή υπάρχει και σε μορφή που χρησιμοποιεί διαφορετικά βάρη για τις διάφορες τιμές $f(\overrightarrow{x_i})$:

$$F(X) = \frac{1}{\sum_{\vec{x} \in X} w(\vec{x})} \sum_{\vec{x} \in X} w(\vec{x}) f(\vec{x}) \qquad (2.3)$$

- Presence/Threshold/Count Based Assumption:

Με αυτές τις υποθέσεις ένα αντικείμενο θεωρείται θετικό αν ο αριθμός των στιγμιότυπων που εκτιμώνται θετικά ή και αρνητικά (ξεχωριστά) τηρεί μια (διπλή) ανισότητα. Δεν χρησιμοποιούνται τόσο συχνά όσο οι προηγούμενες δυο.

2.4.2. Bag Space

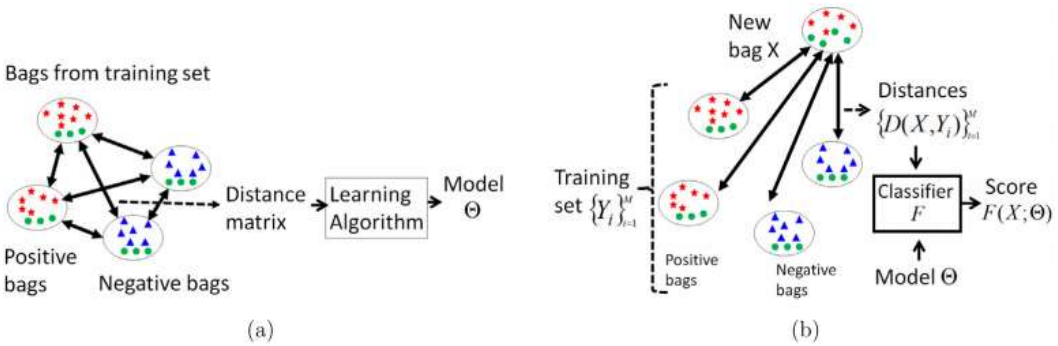
Σε αντίθεση με την κατηγορία του instance space, οι αλγόριθμοι στο bag space (BS) θεωρούν πως η πληροφορία για την κατηγορία ενός αντικειμένου δεν μπορεί να εξαχθεί από το κάθε στιγμιότυπο μεμονωμένα. Η εκτίμηση για την ετικέτα ενός αντικειμένου γίνεται μέσω της σύγκρισης του με άλλα αντικείμενα.

Τα αντικείμενα στα MIL προβλήματα είναι μη-διανυσματικές οντότητες. Αποτελούνται από πολλά διανύσματα και στην ουσία είναι πίνακες. Για αυτόν τον λόγο είναι αναγκαία η χρήση κάποιας συνάρτησης απόστασης ή ομοιότητας μεταξύ των αντικειμένων ώστε να προκύψουν μετρικές που να είναι συμβατές με τα μοντέλα μηχανικής μάθησης που υπάρχουν. Τέτοιου είδους συναρτήσεις είναι για παράδειγμα η minimal Hausdorff απόσταση (εξίσωση 2.1) και η απόσταση Chamfer (εξίσωση 2.2).

$$D(X,Y) = \min_{\vec{x}\in X,\vec{y}\in Y} \|\vec{x}-\vec{y}\| \tag{2.4}$$

$$D(X,Y) = \frac{1}{|X|}\sum_{\vec{x}\in X}\min_{\vec{y}\in Y}\|\vec{x}-\vec{y}\| + \frac{1}{|Y|}\sum_{\vec{y}\in Y}\min_{\vec{x}\in X}\|\vec{x}-\vec{y}\| \tag{2.5}$$

Αφού δημιουργηθεί ένα μητρώο αποστάσεων σύμφωνα με μια τέτοια συνάρτηση μεταξύ των αντικειμένων του training set εκπαιδεύουμε έναν αλγόριθμο και δημιουργούμε το μοντέλο μας. Στην συνέχεια για κάθε αντικείμενο του test set υπολογίζονται οι αποστάσεις του από τα αντικείμενα της εκπαίδευσης και σύμφωνα με το μοντέλο που δημιουργήθηκε νωρίτερα γίνεται η κατηγοριοποίηση του.

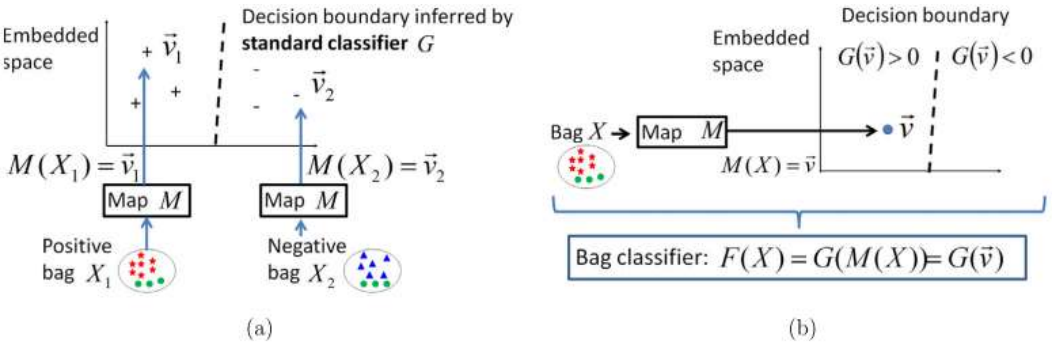


Σχήμα 2.9: Παράδειγμα Bag Space μοντέλου.

2.4.3. Embedded Space

Η κατηγορία του embedded space (ES) μοιάζει με αυτήν του bag space στο ότι χρησιμοποιούν bag-level πληροφορία σε αντίθεση με τους instance space αλγορίθμους. Η διαφορά τους είναι πως ενώ στο bag space η σύγκριση των bags γίνεται μέσω μιας συνάρτησης απόστασης, στο embedded space τα αντικείμενα γίνονται map σε διανύσματα και η σύγκριση γίνεται μέσω αυτών.

Αυτό που γίνεται στην ουσία είναι πως από τα επιμέρους στιγμιότυπα των αντικειμένων εξάγονται καινούρια χαρακτηριστικά μέσω μιας συνάρτησης, έστω *M(X)* και το πρόβλημα μετατρέπεται σε ένα κλασσικό πρόβλημα εποπτευόμενης μάθησης ενιαίων στιγμιότυπων. Στην συνέχεια το πρόβλημα επιλύεται με την δημιουργία ενός κλασσικού κατηγοριοποιητή, έστω *G($\vec{v}$),* στον νέο «embedded» διανυσματικό χώρο που δημιούργησε η *M(X)*.



Σχήμα 2.10: Παράδειγμα Embedded Space μοντέλου.

Οι αλγόριθμοι που ανήκουν στην κατηγορία του Embedded Space μπορούν να χωριστούν σε δυο περεταίρω υποκατηγορίες, ανάλογα με το αν κάνουν χρήση «λεξικού» ή όχι. Δεν θα εμβαθύνουμε σε αυτό το κομμάτι καθώς είναι εκτός του πεδίου της εργασίας.

ΚΕΦΑΛΑΙΟ 3: SUPPORT VECTOR MACHINE

3.1. Γενικά

Το Support Vector Machine (SVM) και οι παραλλαγές του αποτελούν από τα πιο βασικά, απλά και ταυτόχρονα ισχυρά εργαλεία της μηχανικής μάθησης για το πρόβλημα της κατηγοριοποίησης δεδομένων. Είναι κατά βάση ένα γεωμετρικό μοντέλο κατηγοριοποίησης που εφόσον κατασκευαστεί έχει πολύ κατανοητό τρόπο λειτουργίας.

Τα πλεονεκτήματα του που το καθιστούν χρήσιμο μέχρι και σήμερα είναι μεταξύ άλλων η πολύ καλή επίδοση σε διανυσματικούς χώρους μεγάλης τάξης (αρκεί να υπάρχουν εξίσου πολλά δεδομένα εκπαίδευσης) [16], οι πολύ χαμηλές απαιτήσεις για μνήμη μετά την εκπαίδευση κατά την χρήση του μοντέλου, καθώς και η δυνατότητα προσαρμογής σε μη γραμμικά διαχωρίσιμα δεδομένα.

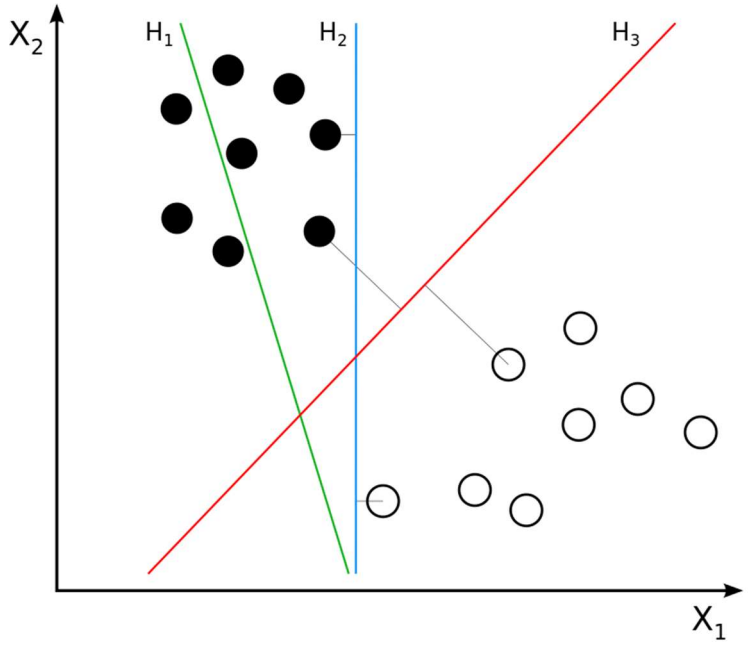
Όπως είναι αναμενόμενο φυσικά έχει και αυτό τα μειονεκτήματα του, με το βασικότερο να είναι το γεγονός ότι μιλάμε για έναν μη-πιθανοτικό κατηγοριοποιητή, που δηλαδή δεν παράγει αποτελέσματα με έναν βαθμό σιγουριάς. Αυτό δημιουργεί προβλήματα σε τεχνικές που χρησιμοποιούν πάνω από ένα SVM καθώς δεν υπάρχει σαφής τρόπος σύγκρισης των αποτελεσμάτων τους. Στο πλαίσιο αυτό έχουν εξελιχθεί διάφορες τεχνικές στις οποίες θα αναφερθούμε στην ενότητα 4.5.

Τα SVM που θα χρησιμοποιήσουμε στην εργασία μας είναι τα one-class SVM που αναφέρονται στην ενότητα 3.6 και η συγκεκριμένη υλοποίηση είναι αυτή του LibSVM [17] για την MATLAB.

3.2. Hard-Margin

Το SVM δημιουργήθηκε από τους Vladimir N. Vapnik και Alexey Ya. Chervonenkis το 1963 [18]-[19]. Σε εκείνη του τη μορφή ήταν ένας απλός δυαδικός γραμμικός κατηγοριοποιητής, δηλαδή δούλευε μόνο για δεδομένα τα οποία ήταν εντελώς γραμμικά διαχωρίσιμα και ανήκαν μόνο σε δυο κλάσεις.

Για λόγους απλότητας ας θεωρήσουμε δυσδιάστατα δεδομένα, δηλαδή που αποτελούνται από διανύσματα $\vec{x}_i$ δύο χαρακτηριστικών και τις ταμπέλες τους y_i. Ένα σύνολο τέτοιων δεδομένων $(\vec{x}_i, y_i)$ μπορεί να αναπαρασταθεί από σημεία δύο χρωμάτων σε έναν δυσδιάστατο χώρο. Σκοπός του hard-margin SVM, ή αλλιώς maximum margin classifier (MMC), είναι να βρει ένα υπερ-επίπεδο το οποίο ονομάζεται όριο απόφασης (decision boundary), στο παράδειγμά μας μια γραμμή, που να χωρίζει με βέλτιστο τρόπο τα σημεία που ανήκουν στην μια κλάση από αυτά που ανήκουν στην δεύτερη. Ως βέλτιστο όριο απόφασης ορίζεται αυτό που έχει την μέγιστη δυνατή απόσταση από το κοντινότερο σημείο κάθε κλάσης, δηλαδή το μεγαλύτερο margin, γι' αυτό και ονομάζεται maximum margin hyperplane (MMH). Η εκπαίδευση ενός τέτοιου SVM δεν είναι κάτι άλλο παρά ένα πρόβλημα μεγιστοποίησης αυτού του margin, με αποτέλεσμα το υπερ-επίπεδο και τα support vectors, τα σημεία δηλαδή που «ζουν» επάνω στα margins.



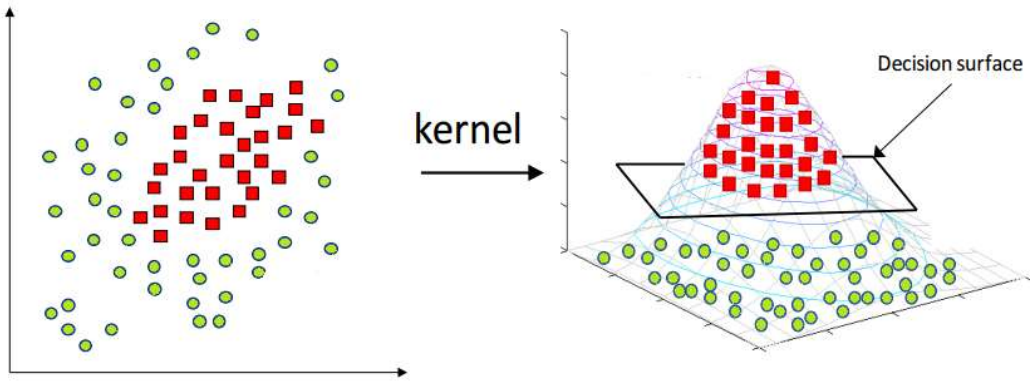
Σχήμα 3.1: Παράδειγμα hard-margin SVM.

Στο σχήμα 3.1 έχουμε δεδομένα δύο χαρακτηριστικών (x_1,x_2) τα οποία ανήκουν σε δυο γραμμικά διαχωρίσιμες κλάσεις, την μαύρη και την λευκή. Οι τρείς γραμμές (υπερ-επίπεδα) είναι τρία πιθανά όρια απόφασης. Η H_1 δεν διαχωρίζει τις κλάσεις, η H_2 τις διαχωρίζει αλλά όχι με βέλτιστο τρόπο και η H_3 είναι το ΜΜΗ, αφού όπως είναι προφανές απέχει το μέγιστο δυνατόν και από τις δυο κλάσεις. Στην περίπτωση της H_3 τα support vectors είναι τα δύο σημεία (διανύσματα) που απέχουν την μικρότερη απόσταση από το υπερ-επίπεδο.

Το σημαντικότερο πρόβλημα με αυτήν την υλοποίηση (hard-margin) είναι πως αν τα δεδομένα εκπαίδευσης δεν είναι εντελώς γραμμικά διαχωρίσιμα δεν δουλεύει, καθόλου. Για την υπέρβαση αυτού του προβλήματος προτάθηκαν δυο διαφορετικές λύσεις οι οποίες χρησιμοποιούνται μέχρι και σήμερα, το kernel trick και τα soft-margin SVM.

3.3. Kernel Trick

Η πρώτη λύση ήρθε από τους Bernhard E. Boser, Isabelle M. Guyon και Vladimir N. Vapnik το 1992 [20] με την μορφή του λεγόμενου kernel trick. Όταν τα δεδομένα δεν είναι γραμμικά διαχωρίσιμα μπορούν, μέσω της εφαρμογής ενός μη γραμμικού μετασχηματισμού, να μεταφερθούν σε έναν άλλον διανυσματικό χώρο όπου να είναι γραμμικά διαχωρίσιμα. Με την χρήση αυτού του κόλπου τα hard-margin SVM μπορούν να χειριστούν και μη γραμμικά διαχωρίσιμα δεδομένα, αρκεί να γίνει χρήση κάποιου από τους πολλούς διαθέσιμους μη γραμμικούς μετασχηματισμούς (polynomial, RBF, sigmoid, κτλ).



Σχήμα 3.2: Παράδειγμα χρήσης RBF kernel.

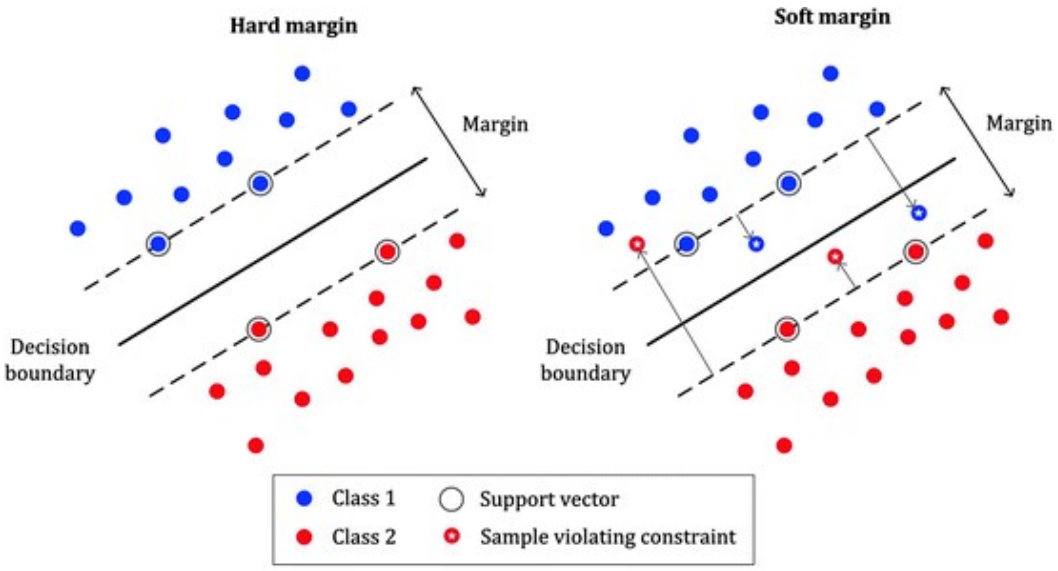
Στο σχήμα 3.2 φαίνεται ένα παράδειγμα χρήσης του radial basis function (RBF) kernel (εξίσωση 3.1) σε κάποια μη γραμμικά διαχωρίσιμα δεδομένα. Τα δυσδιάστατα δεδομένα προβάλλονται σε έναν νέο τρισδιάστατο χώρο όπου μπορούν να διαχωριστούν πλήρως με γραμμικό τρόπο με την χρήση ενός υπερ-επιπέδου. Στην ουσία πρόκειται ακόμα για έναν γραμμικό κατηγοριοποιητή, αλλά στον αρχικό χώρο των δεδομένων το όριο απόφασης θα εμφανιστεί σαν έλλειψη.

$$K(x, x') = \exp\left(-gamma\, \|x - x'\|^2\right) \qquad (3.1)$$

Το κακό με αυτήν την μέθοδο είναι πως το σφάλμα της γενίκευσης (generalization error) γίνεται συγκριτικά μεγαλύτερο από αυτό του γραμμικού κατηγοριοποιητή αφού είναι πολύ πιο εύκολο να υπάρξει over-fitting στα δεδομένα εκπαίδευσης.

3.4. Soft-Margin

Η δεύτερη λύση όσον αφορά τα μη γραμμικά διαχωρίσιμα δεδομένα είναι τα soft-margin SVM που δημοσιεύτηκαν από τους Corinna Cortes και Vladimir N. Vapnik το 1995 [21]-[22]. Τα soft-margin ή noisy SVM όπως αλλιώς ονομάζονται, έχουν την δυνατότητα κατηγοριοποίησης τέτοιων δεδομένων επιτρέποντας λάθη στην διαδικασία της εκπαίδευσης. Με την χρήση μιας υπερ-παραμέτρου «C» ή «nu» δίνεται ένα κόστος στα λάθη κάνοντας το πρόβλημα εύρεσης του ορίου απόφασης πιο ευέλικτο. Με την μέθοδο αυτή ο χρήστης έχει την δυνατότητα να διαλέξει κατά πόσο επιθυμεί ένα μεγάλο margin για θεωρητικά καλύτερη γενίκευση ή μια μεγάλη ακρίβεια εκπαίδευσης.



Σχήμα 3.3: Hard-margin και soft-margin SVM.

Στο σχήμα 3.3 φαίνεται η διαφορά των hard-margin από τα soft-margin SVM. Στην δεύτερη περίπτωση υπάρχουν κάποια δεδομένα, που συμβολίζονται ως διάτρητα, τα οποία χαλάνε την γραμμική διαχωριστικότητα. Αυτό δεν εμποδίζει το SVM από το να δουλέψει όμως αφού επιτρέπονται κάποια λάθη.

Πιο συγκεκριμένα, για κάθε σημείο ορίζεται μια τιμή ξ_i η οποία είναι η απόσταση του από το σωστό όριο του margin. Αυτή η τιμή είναι 0 όταν το σημείο βρίσκεται από την σωστή μεριά του margin, 1 όταν βρίσκεται ακριβώς πάνω στο όριο απόφασης (στην μέση του margin) και μεγαλύτερη του 1 όσο απομακρύνεται από αυτό. Το συνολικό επιτρεπτό άθροισμα των τιμών ξ_i για την εκπαίδευση ενός soft-margin SVM δίνεται από την υπερ-παράμετρο $C = \sum_{i=1}^{I} \xi_i$.

Η υλοποίηση ενός soft-margin SVM μπορεί να χρησιμοποιηθεί και για την κατασκευή ενός hard-margin SVM αφού για C = 0 στην ουσία δεν επιτρέπεται κανένα λάθος.

Όπως είναι προφανές, η υπερ-παράμετρος C δεν έχει άνω φράγμα, ούτε μπορεί να προκύψει κάποιο χρήσιμο συμπέρασμα από αυτήν πέραν του ότι όσο μεγαλύτερη είναι, τόσο μεγαλώνει το margin εις βάρος της ακρίβειας εκπαίδευσης. Το 2001 οι Chih-Chung Chang και Chih-Jen Lin [23] πρότειναν μια διαφορετική υπερ-παράμετρο που ονομάστηκε nu (ή αλλιώς ν) η οποία είναι φραγμένη στο διάστημα [0,1] και αποτελεί (α) άνω φράγμα στο ποσοστό των λαθών στην εκπαίδευση και (β) κάτω φράγμα στο ποσοστό των δειγμάτων εκπαίδευσης που θα αποτελέσουν support vectors.

Για παράδειγμα, με 100 δείγματα εκπαίδευσης και nu = 0.05 θα γίνουν το πολύ 5 λάθος κατηγοριοποιήσεις στα δεδομένα εκπαίδευσης και το SVM που θα προκύψει θα έχει τουλάχιστον 5 support vectors.

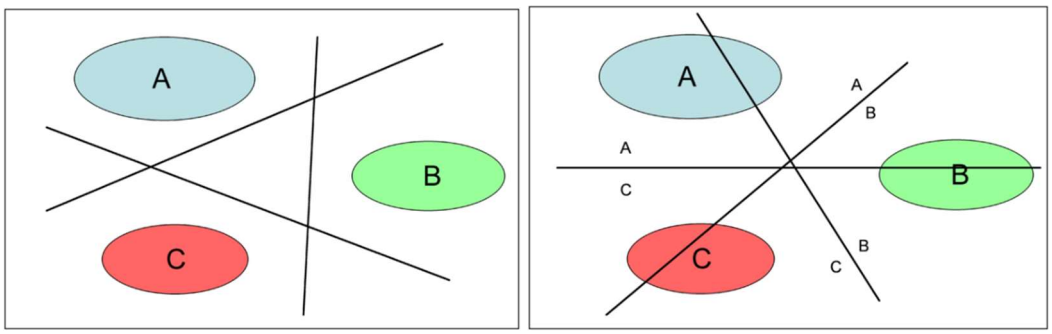
Σήμερα τα SVM που χρησιμοποιούνται συνήθως συνδυάζουν και τις δυο μεθόδους που αναφέρθηκαν, δηλαδή είναι soft-margin με μια υπερ-παράμετρο κόστους και κάνουν ταυτόχρονα χρήση του kernel trick, συνήθως με ένα radial basis function (RBF) kernel. Ο συνδυασμός αυτός σε πραγματικά δεδομένα συχνά οδηγεί στο καλύτερο αποτέλεσμα αφού η επιλογή ενός κατάλληλου kernel που βοηθάει στον διαχωρισμό των κλάσεων δίνει την δυνατότητα χρήσης υψηλής υπερ-παραμέτρου κόστους για ένα μεγάλο margin, αλλά ταυτόχρονα και με μεγάλη ακρίβεια εκπαίδευσης.

3.5. Multi-Class

Το άλλο πρόβλημα με τα SVM είναι πως εκ φύσεως προορίζονται για προβλήματα δυο κλάσεων. Για την κατηγοριοποίηση περισσότερων κλάσεων γίνεται χρήση one-versus-all (OvA) ή one-versus-one (OvO) τεχνικών.

Στην one-versus-all τεχνική εκπαιδεύεται ένα SVM ανά κλάση το οποίο διαχωρίζει αυτήν την κλάση από όλες τις υπόλοιπες. Για την πρόβλεψη, κάθε νέο δείγμα εκτιμάται από όλα τα SVM και κερδίζει αυτό με το μεγαλύτερο αποτέλεσμα. Για την one-versus-one τεχνική χρειάζεται ένα SVM ανά ζεύγος κλάσεων. Εδώ πάλι για την πρόβλεψη κάθε δείγμα εκτιμάται από όλα τα SVM και αυτό που θα κερδίσει είναι αυτό με τις περισσότερες νίκες σε επίπεδο ζευγών.

Όπως αναφέραμε νωρίτερα όμως, τα αποτελέσματα των SVM δεν είναι συγκρίσιμα μεταξύ τους καθώς δεν πρόκειται για πιθανότητες αλλά για απλές αποστάσεις σημείων. Επομένως θα αναρωτηθεί κανείς πώς γίνονται οι απαραίτητες συγκρίσεις στις τεχνικές αυτές. Η απάντηση είναι πως επειδή πρόκειται για two-class SVM μπορούν να χρησιμοποιηθούν μέθοδοι βαθμονόμησης των αποτελεσμάτων ώστε να γίνουν συγκρίσιμα. Τέτοιες μέθοδοι είναι π.χ. το Platt's Scaling [24]-[25] και το Isotonic Regression [26], για τις οποίες γίνεται λόγος στο κεφάλαιο 4.5.



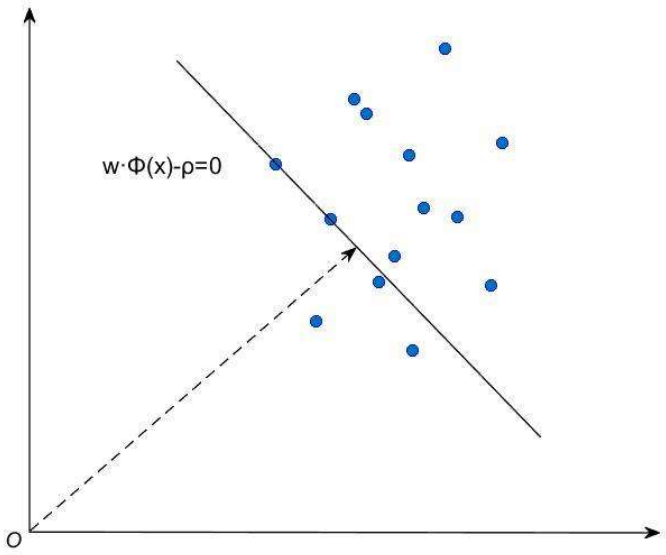
Σχήμα 3.4: Οι multi-class τεχνικές OvA (αριστερά) και OvO (δεξιά).

Η γνωστή βιβλιοθήκη LibSVM χρησιμοποιεί την τεχνική one-versus-one καθώς σύμφωνα με σχετική έρευνα [27] υπερτερεί στον χρόνο εκπαίδευσης, παρότι και οι δυο τεχνικές αποδίδουν το ίδιο.

3.6. One-Class

Σε κάποια πραγματικά προβλήματα είναι εύκολη η συλλογή θετικών δειγμάτων αλλά πάρα πολύ δύσκολη η συλλογή αρνητικών δειγμάτων, για παράδειγμα οι σπάνιες βλάβες σε βιομηχανικά συστήματα ή σε αεροπλάνα. Επομένως η δημιουργία ενός συστήματος για τον εντοπισμό ακραίων τιμών (outlier detection) είναι πολύ δύσκολη με τα προαναφερθέντα μοντέλα.

Για τις περιπτώσεις αυτές δημιουργήθηκαν τα one-class SVM. Αυτά εκπαιδεύονται μόνο με την χρήση θετικών δειγμάτων, δηλαδή αντικειμένων μιας κλάσης, και στην περίπτωση του linear kernel προσπαθούν να διαχωρίσουν τα σημεία στον διανυσματικό χώρο από το σημείο μηδέν. Σε περίπτωση που χρησιμοποιηθεί ένας RBF kernel το one-class SVM προσπαθεί να περικλύσει τα θετικά δείγματα με το decision boundary.



Σχήμα 3.5: Παράδειγμα linear kernel one-class SVM.

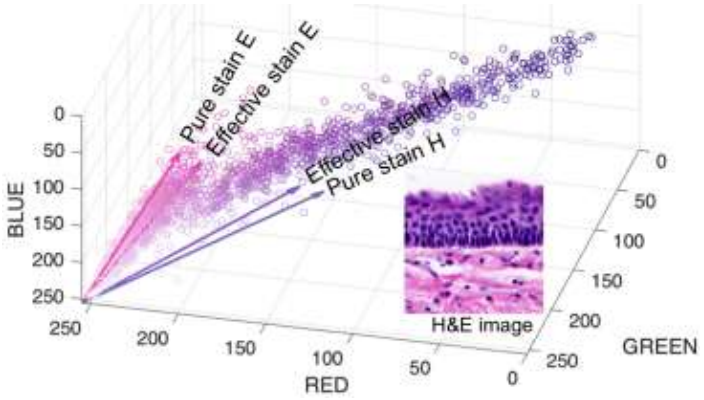
Τα one-class SVM μπορούν να χρησιμοποιηθούν και σε προβλήματα δύο κλάσεων αντί ενός soft-margin SVM, με την τεχνική one-versus-one. Αυτό είναι και το αντικείμενο της εργασίας αυτής.

Κεφαλαιο 4: Γενικοτερες Μεθοδοι και Έννοιες

4.1. Κανονικοποίηση Απόχρωσης για Stains

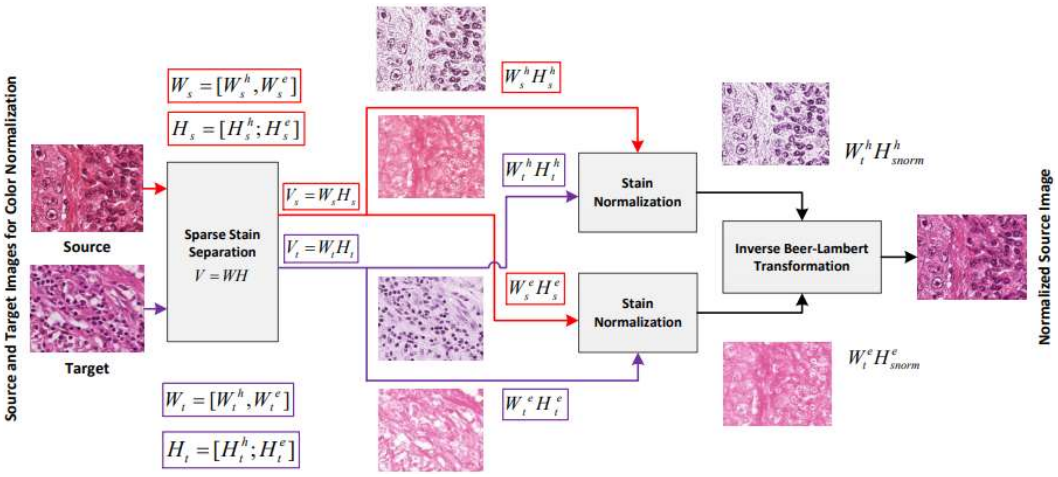
Τα δεδομένα που θα χρησιμοποιηθούν για την αξιολόγηση των μεθόδων της εργασίας είναι ιστοπαθολογικές εικόνες που ονομάζονται stains [2], επειδή στην ουσία είναι ιστός βαμμένος με αιματοξυλίνη και εωσίνη (H&E). Η αιματοξυλίνη βάφει σκούρο μπλε τους πυρήνες των κυττάρων και η εωσίνη βάφει ροζ το κυτόπλασμα και την εξωκυτταρική μήτρα, με άλλες κυτταρικές δομές να παίρνουν κάποια απόχρωση των δύο αυτών βασικών χρωμάτων. Αυτό κάνει πιο «ευανάγνωστη» την δομή του ιστού, στην συγκεκριμένη περίπτωση του μαστού, για να γίνει ευκολότερη η διάγνωση ασθενειών όπως η ανάπτυξη νεοπλασιών και ο εντοπισμός του καρκίνου.

Τέτοιες εικόνες παρουσιάζουν μεταξύ τους ανεπιθύμητες διαφορές στα χρώματα οι οποίες οφείλονται σε διαφορές στην κατασκευαστική μέθοδο των πρώτων υλών (των βαφών), στο μικροσκόπιο που τράβηξε την φωτογραφία και πιθανώς σε επιπλέον παράγοντες. Αυτές οι διαφορές μεταξύ των εικόνων δεν μας ενδιαφέρουν και θέλουμε να τις εξαλείψουμε. Στην έρευνα [5] αναφέρονται πιο διεξοδικά οι λόγοι για τους οποίους χρειάζεται το λεγόμενο stain color normalization καθώς επίσης και μια μέθοδος για να γίνει αυτό.



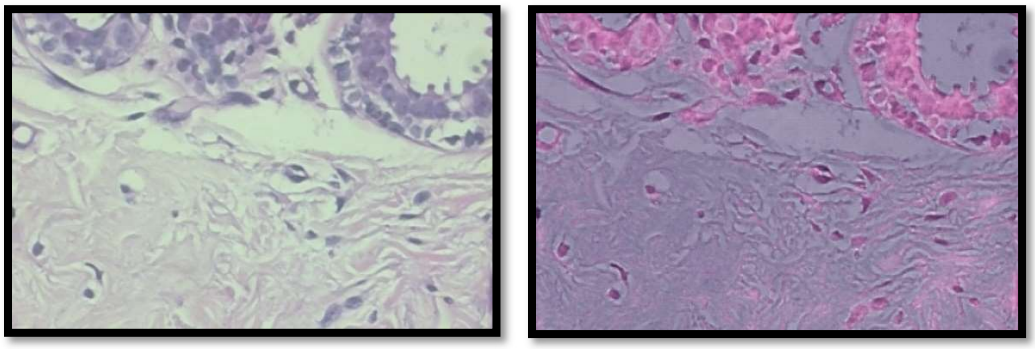
Σχήμα 4.1: Τα χρώματα ενός H&E stain σαν διανύσματα RGB.

Η μέθοδος αυτή δέχεται τις εικόνες προς διόρθωση και μια εικόνα σαν σημείο αναφοράς για τα χρώματα. Στην συνέχεια προσπαθεί όσο καλύτερα μπορεί να ξεχωρίσει τα χρώματα που έχουν ως βάση το μωβ από αυτά που έχουν ως βάση το ροζ, δημιουργώντας δύο εικόνες ανά αρχική εικόνα. Αυτές αλλάζουν απόχρωση ξεχωριστά ώστε να ταιριάξουν με τις δυο αποχρώσεις της εικόνας αναφοράς και στην συνέχεια συνδυάζονται πάλι σε μια εικόνα. Ένας τέτοιος αλγόριθμος πρέπει να παρεμβαίνει στα χρώματα χωρίς να αλλοιώνει την δομή των εικόνων, γι’ αυτό και αναπτύχθηκε η συγκεκριμένη μέθοδος.



Σχήμα 4.2: Η προτεινόμενη μέθοδος της δημοσίευσης [5].

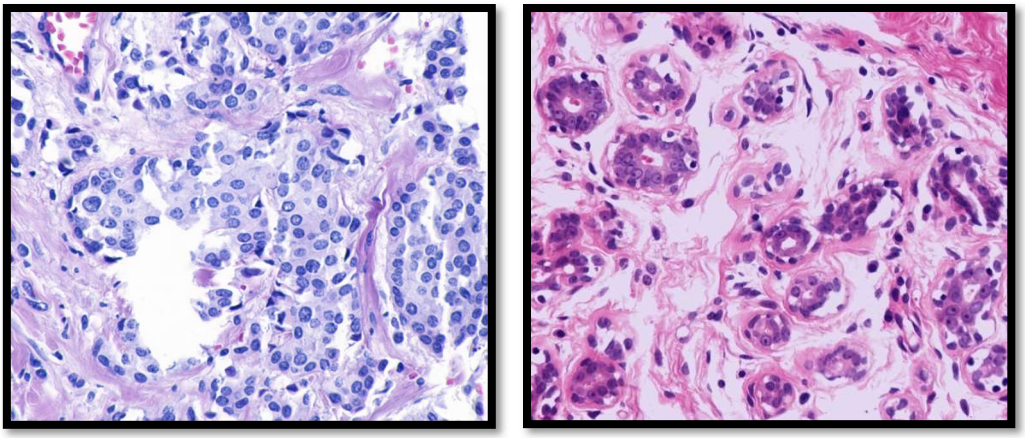
Λόγω εστίασης σε άλλα κομμάτια της εργασίας δεν ήταν εφικτό να αναπτυχθεί κώδικας για αυτήν την μέθοδο. Ο κώδικας που χρησιμοποιήσαμε στην παρούσα εργασία βρίσκεται στην ιστοσελίδα [28] και λειτουργεί με παρόμοιο τρόπο. Σε αυτό το σημείο πρέπει να σημειωθεί πως δεν είχαμε πάντα το επιθυμητό οπτικό αποτέλεσμα με την εφαρμογή του παραπάνω κώδικα. Γενικά παρατηρήθηκε η τάση το λευκό χρώμα να ροζιάζει. Ειδικότερα στο BreaKHis πολλές εικόνες είναι πάρα πολύ ξεθωριασμένες με αποτέλεσμα να μπερδεύεται η αιματοξυλίνη με την εωσίνη και να δημιουργείται ένα πάρα πολύ κακό αποτέλεσμα. Ακολουθούν παραδείγματα χρήσης και από τα δυο datasets.



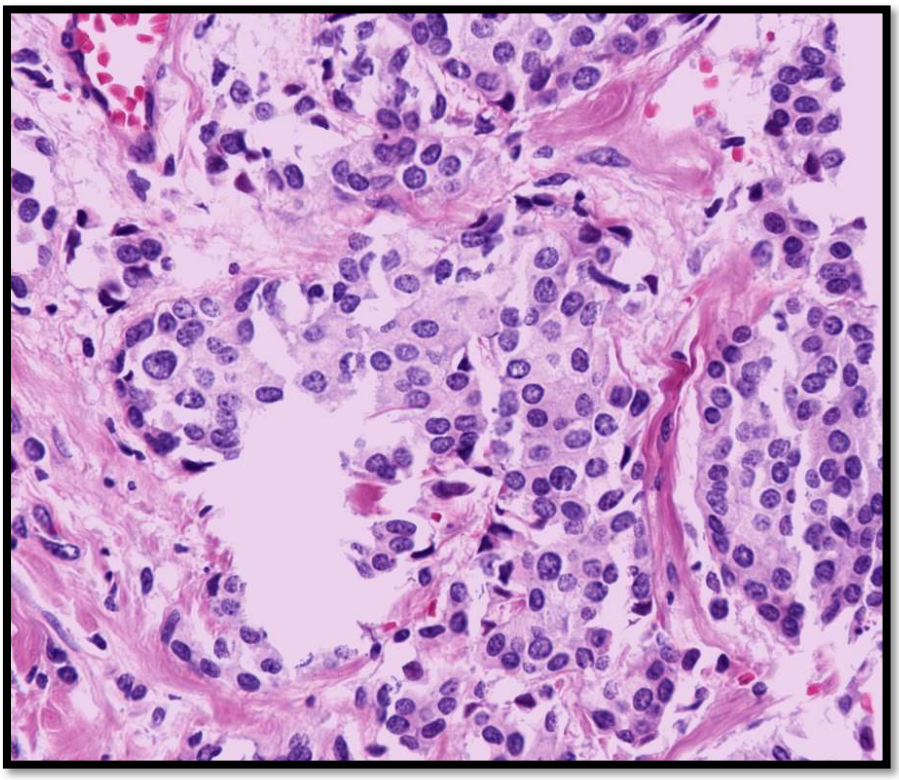
Σχήμα 4.3: Λάθος αναγνώριση της αιματοξυλίνης σαν εωσίνη στο BreaKHis.

Το λάθος στο παραπάνω σχήμα ήταν αρκετά συχνό ώστε να μας αποτρέψει από το να χρησιμοποιήσουμε την μέθοδο αυτή στο BreaKHis. Απ’ την

άλλη στο BCC η μέθοδος ήταν πετυχημένη και τα αποτελέσματα όπως θα φανεί και από τα πειράματα είναι ενθαρρυντικά.



Σχήμα 4.4: Αριστερά η αρχική εικόνα και δεξιά η εικόνα αναφοράς.



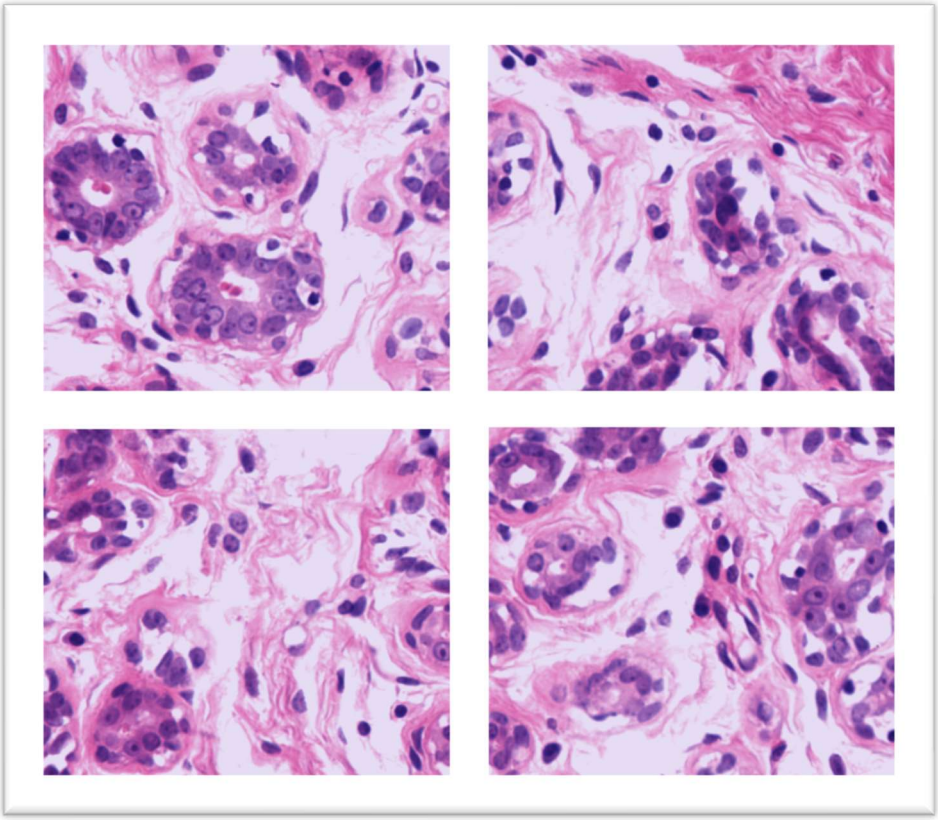
Σχήμα 4.5: Πετυχημένος μετασχηματισμός από το BCC. Παρατηρείται μια τάση το (όχι και τόσο καθαρό) λευκό να πλησιάζει προς το ροζ.

4.2. Κατακερματισμός Εικόνας – Patches

Κάθε πρόβλημα κατηγοριοποίησης εικόνων μπορεί να μετατραπεί σε MIL πρόβλημα με την δημιουργία πολλαπλών στιγμιότυπων (patches) από κάθε εικόνα, τα οποία κληρονομούν την κλάση της εικόνας-αντικειμένου (bag).

Η υλοποίηση της δημιουργίας στιγμιότυπων σε αυτήν την εργασία εξάγει στιγμιότυπα ίδιων διαστάσεων μεταξύ τους, πλήθους x-επί-x, δηλαδή x ανά διάσταση. Μπορεί κανείς να παρομοιάσει την διαδικασία αυτή με την κοπή μιας τετράγωνης πίτσας σε x-επί-x κομμάτια ίδιας επιφάνειας.

Προαιρετικά μπορεί να υπάρξει και επικάλυψη (overlap) μεταξύ των στιγμιότυπων, όπου κάθε μέρος της εικόνας (εκτός από τα άκρα) περιλαμβάνεται περισσότερες από μια φορές στα εξαχθέντα στιγμιότυπα τα οποία πολλαπλασιάζονται σε αριθμό. Αυτό μπορεί να θεωρηθεί σαν κάποιου είδους επαύξηση των δεδομένων (data augmentation). Το overlap μπορεί να φανεί ιδιαίτερα χρήσιμο σε συνδυασμό με την εφαρμογή απόρριψης στιγμιότυπων, δηλαδή ενός κανόνα που θα διαγράφει στιγμιότυπα σύμφωνα με έναν κανόνα (π.χ. αν περιέχουν πολύ λευκό).



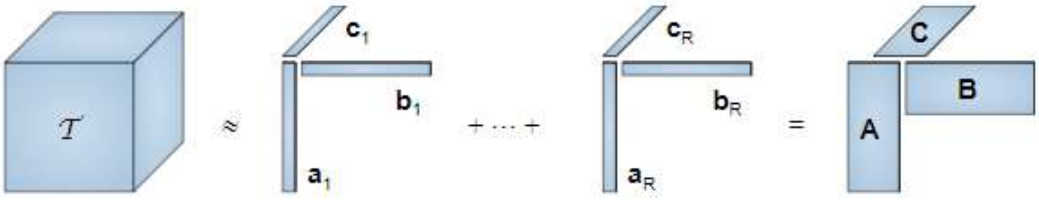
Σχήμα 4.6: Οπτικοποίηση 2-επί-2 κατακερματισμού μιας εικόνας σε 4 patches.

4.3. Εξαγωγή Χαρακτηριστικών – Non-Negative CPD

Για να μπορέσουμε να χρησιμοποιήσουμε αποτελεσματικά εικόνες σαν είσοδο σε ένα μοντέλο μηχανικής μάθησης όπως είναι τα SVM πρέπει να ακολουθήσουμε μια διαδικασία εξαγωγής χαρακτηριστικών (feature extraction) που θα αντικαταστήσουν τα pixels. Τα τελευταία αποτελούν το κατώτερο επίπεδο χαρακτηριστικών για μια εικόνα και η εξαγωγή χαρακτηριστικών υψηλότερου επιπέδου πέραν του ότι μειώνει την διαστατικότητα της εισόδου, μας επιτρέπει να δημιουργήσουμε έναν πιο έξυπνο κατηγοριοποιητή αφού δεν θα ασχολείται με τελείως low-level πληροφορία αλλά με πιο high-level αφηρημένες έννοιες.

Η συνηθισμένη μέθοδος feature extraction από εικόνες είναι η χρήση περιγραφέων (descriptors) οι οποίοι αφορούν το χρώμα, την υφή, το σχήμα, την κίνηση, την θέση, κ.τ.λ.. Μερικά παραδείγματα τέτοιων περιγραφέων είναι τα Parameter Free Threshold Adjacency Statistics (PFTAS) [6]-[7], Scale Invariant Feature Transform (SIFT) [8], Gray Level Co-occurrence Matrix (GLCM) [9] και Histogram of Oriented Gradients (HOG) [10]. Οι επιλογές είναι πάρα πολλές και καθιστούν το πρόβλημα της εξαγωγής χαρακτηριστικών application specific.

Για την εργασία αυτήν επιλέχθηκε μια μέθοδος που κάνει χρήση της Canonical Polyadic Decomposition (CPD). Η CPD ή αλλιώς PARAFAC ή CANDECOMP αποτελεί κάτι σαν γενίκευση της Singular Value Decomposition (SVD) για τανυστές και εφευρέθηκε πρώτη φορά από τον Frank L. Hitchcock το 1927 [11]. Αυτό που κάνει είναι η αποσύνθεση ενός τανυστή σε ένα άθροισμα απλών τανυστών τάξεως 1. Είναι προφανές πως αυτή η μέθοδος είναι τελείως application agnostic και μπορεί να εφαρμοστεί σε οποιοδήποτε πρόβλημα.



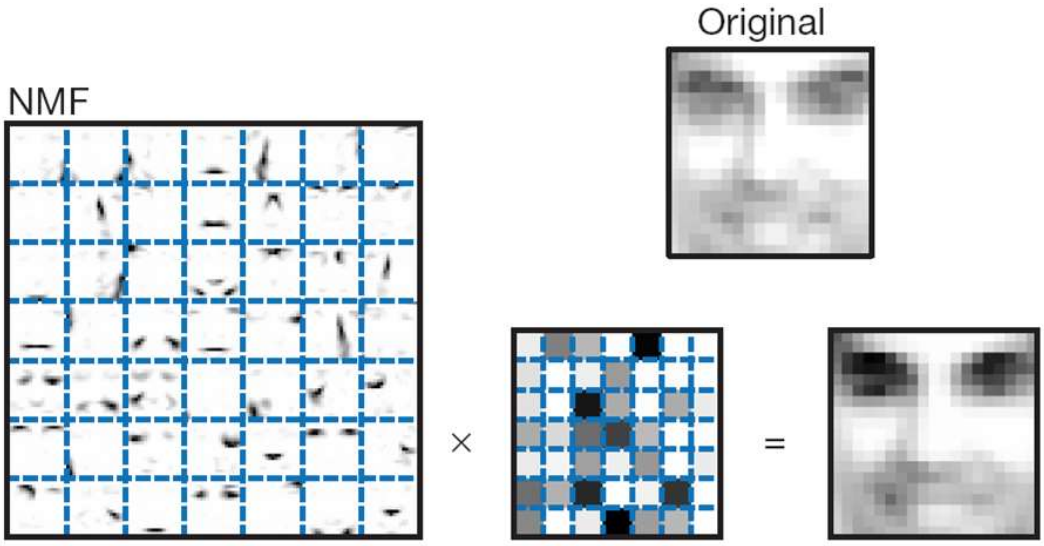
Σχήμα 4.7: Οπτικοποίηση των προϊόντων της CPD.

$$T \approx \sum_{r=1}^{R} a_r \otimes b_r \otimes c_r \qquad (4.1)$$

Όπως φαίνεται στο παραπάνω σχήμα, η CPD για έναν 3-διάστατο τανυστή T παράγει τρία μητρώα A, B, C. Αν υποθέσουμε πως ο T είναι διαστάσεων I x J x

K και πως R είναι η τάξη της CPD, τότε το μητρώο A θα έχει διαστάσεις I x R, το B διαστάσεις J x R και το C διαστάσεις K x R. Ο αρχικός τανυστής μπορεί να ανακατασκευαστεί (με κάποιες απώλειες) αθροίζοντας τα εξωτερικά γινόμενα των στηλών των A, B και C (εξίσωση 4.1).

Η μορφή της CPD που επιλέξαμε είναι η Non-Negative CPD (NN-CPD), η οποία επιβάλλει στα μητρώα A, B και C να είναι αυστηρά μη-αρνητικά. Αυτός ο περιορισμός βοηθάει στην ύπαρξη κάποιου είδους φυσικής σημασίας στα μητρώα της εξόδου, όπως ισχύει γενικά και με την Non-Negative Matrix Factorization (NMF) [29].

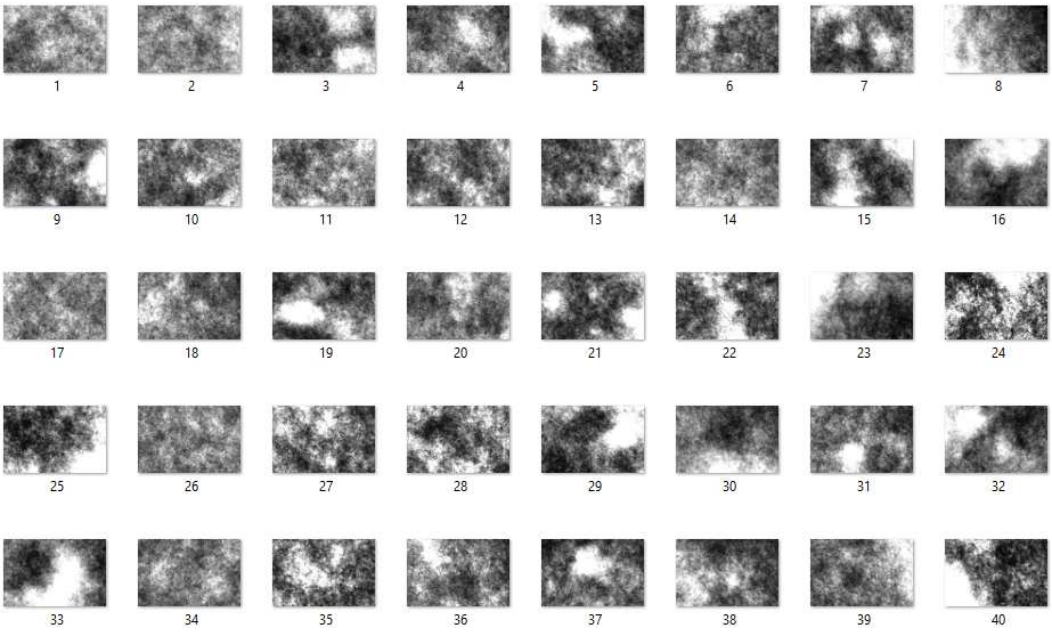


Σχήμα 4.8: Παράδειγμα NMF. Στο αριστερό μητρώο φαίνονται τα features, κομμάτια ενός ανθρώπινου προσώπου, με φυσική σημασία.

Στην MIL κατηγοριοποίηση εικόνων έχουμε ένα σύνολο από πολύχρωμες εικόνες, το οποίο μπορεί να αναπαρασταθεί από έναν 4-διάστατο τανυστή. Αλλάζοντας την διάταξη των pixels με τέτοιο τρόπο ώστε να είναι όλα το ένα δίπλα στο άλλο η διαστάσεις αυτού του τανυστή μπορούν να μειωθούν σε 3: εικόνα (πλήθος εικόνων), pixel (πλήθος pixel ανά εικόνα), RGB (τρία). Εφαρμόζοντας NN-CPD τάξης R σε έναν τέτοιο τανυστή θα παίρναμε τα ακόλουθα μητρώα: A [#εικόνων x R] , B [#pixels x R] , C [3 x R].

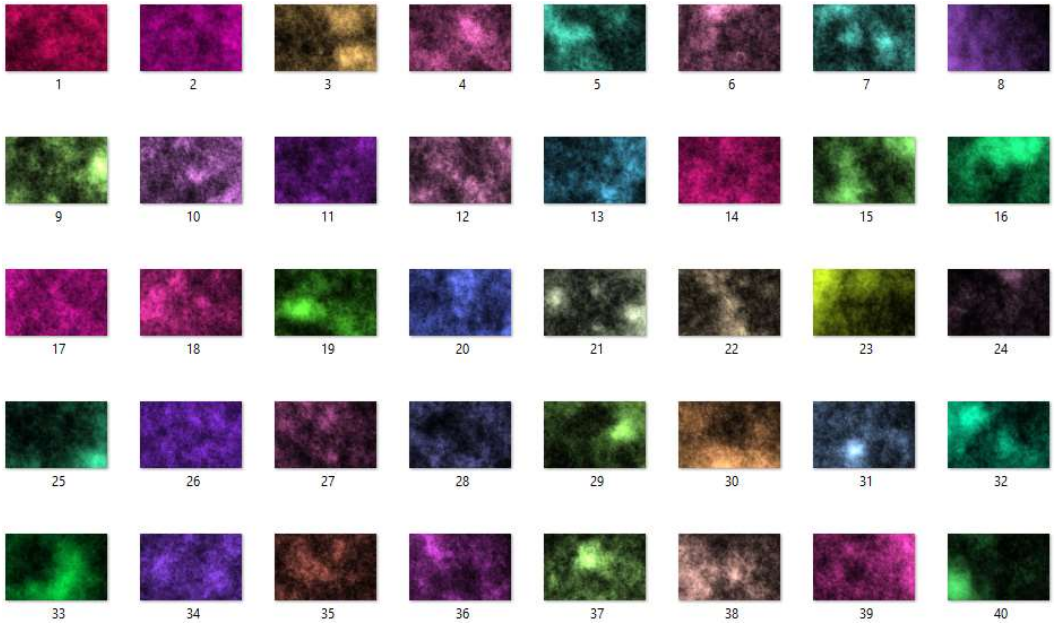
Αν επαναφέρει κάποιος την αρχική διάταξη των pixels στο μητρώο B εισάγοντας πάλι την διάσταση που εξαλείφθηκε νωρίτερα, εύκολα θα διαπιστώσει πως αυτό περιέχει R ασπρόμαυρες «φιγούρες» που ονομάζουμε χωρικά συστατικά (spatial components). Το τρίτο μητρώο χωρίς κάποια επεξεργασία περιέχει R διαφορετικά χρώματα σε μορφή RGB. Μελετώντας την εξίσωση 4.1 είναι φανερό πως κάθε φιγούρα από το B συνδέεται με ένα και μόνο χρώμα από το C και με μία στήλη από το A. Επομένως μπορούμε να θεωρήσουμε

πως κάθε στήλη του A συνδέεται με μια μονόχρωμη (όχι ασπρόμαυρη) φιγούρα. Τι είναι όμως το A; Ας θυμηθούμε πως οι διαστάσεις του είναι [#εικόνων x R]. Είναι φανερό πως κάθε μια από τις αρχικές εικόνες στον τανυστή T μπορεί να εκφραστεί ως ένας γραμμικός συνδυασμός των μονόχρωμων φιγούρων με τους συντελεστές αυτού του γραμμικού συνδυασμού να είναι τα στοιχεία της αντίστοιχης γραμμής του μητρώου A. Επομένως το A μπορεί να χρησιμοποιηθεί σαν ένα feature matrix για το πρόβλημα μας!



Σχήμα 4.9: Οι φιγούρες του μητρώου B από το BreaKHis (200x) για R=40.

Σχήμα 4.10: Τα χρώματα του μητρώου C από το BreaKHis (200x) για R=40.



Σχήμα 4.11: Οι έγχρωμες φιγούρες (εξωτερικά γινόμενα των στηλών των B, C).

Οι συντελεστές στο μητρώο A δεν μας λένε τίποτα εάν τα μητρώα B και C δεν είναι κοινά για ολόκληρο το σύνολο των δεδομένων. Αυτός είναι και ο λόγος που η CPD πρέπει να εφαρμοστεί μια φορά επάνω σε όλες τις εικόνες (και της εκπαίδευσης και της πρόβλεψης). Στην περίπτωση που ένα σύστημα σαν αυτό υλοποιηθεί για χρήση θα πρέπει να γίνεται εξαγωγή χαρακτηριστικών και από νέα δεδομένα που θα έρχονται στην διάθεση μας. Αυτό μπορεί να γίνει απλά με μία και μόνο επιπλέον επανάληψη της CPD στον ενημερωμένο τανυστή με τις πρόσθετες εικόνες.

Σε αυτό το σημείο πρέπει να αναφέρουμε πως η διαδικασία της CPD είναι μια αρκετά χρονοβόρα υπολογιστική διαδικασία με πολύ μεγάλες απαιτήσεις για μνήμη. Γι’ αυτόν τον λόγο στα πειράματα μας περιοριστήκαμε σε μικρές τάξεις R και στο ένα σύνολο δεδομένων αναγκαστήκαμε να μειώσουμε την ανάλυση των αρχικών εικόνων πιθανώς βλάπτοντας τα αποτελέσματα μας.

Σχήμα 4.12: Ανακατασκευή εικόνας μετά από NN-CPD με Rank 30, 40 και 120.

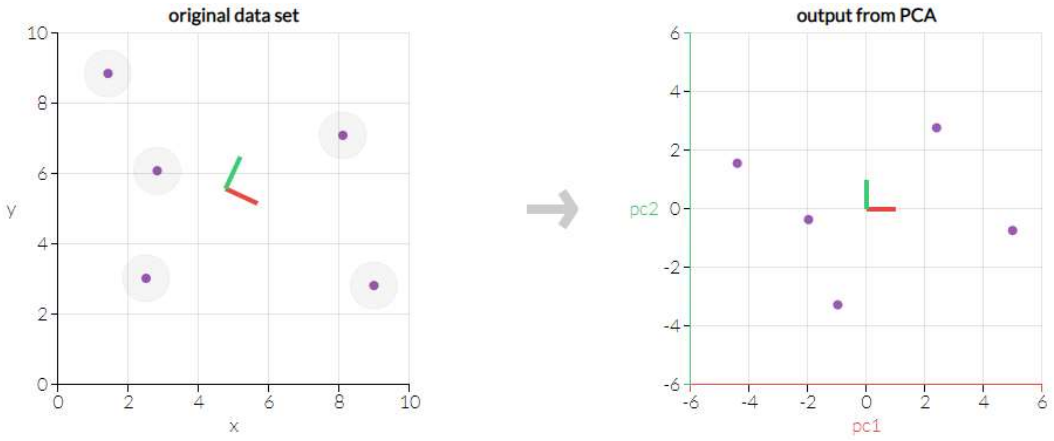
Στο σχήμα 4.12 φαίνεται η επίδραση της τάξης R της NN-CPD στην πληροφορία που χάνεται από μια εικόνα του BCC. Είναι φανερό πως η σχέση μεταξύ της τάξης R και της απώλειας στην ανακατασκευή (ή αλλιώς της πληροφορίας που μένει) δεν είναι γραμμική. Αυτό σημαίνει πως για να πετύχουμε κάτι σημαντικά καλύτερο από το αποτέλεσμα για R = 40 πρέπει να πάμε πιθανώς σε R = 400 όπου και οι υπολογιστικές απαιτήσεις αυξάνονται απαγορευτικά. Όσο το σύνολο δεδομένων μεγαλώνει τόσο το απαιτούμενο R μεγαλώνει επίσης μαζί του. Στο BreaKHis τα πράγματα είναι ακόμα χειρότερα για R = 40 απ' ό,τι φαίνεται στο σχήμα 4.12. Παρόλα αυτά όμως, όπως θα φανεί στο κεφάλαιο 6 από τα πειράματα ο αλγόριθμος μας τα πήγε εξαιρετικά καλά.

Επίσης όσον αφορά την επιλογή της τάξης R, έχει προταθεί μια μέθοδος που ονομάζεται CORCONDIA [30] για την εύρεση ενός βέλτιστου. Η μέθοδος αυτή όμως απαιτεί την επαναληπτική εκτέλεση της CPD πάνω στα συγκεκριμένα δεδομένα για κάθε τάξη R εντός ενός εύρους, κάτι που δεν ήταν εφικτό λόγω μνήμης ή και χρόνου στην περίπτωση μας.

Η υλοποίηση της NN-CPD που χρησιμοποιήσαμε στην παρούσα εργασία ανήκει στην βιβλιοθήκη TensorLy [31] και κάνει χρήση της μεθόδου των multiplicative updates [32] στο gradient descent.

4.4. Μείωση της Διαστατικότητας – PCA

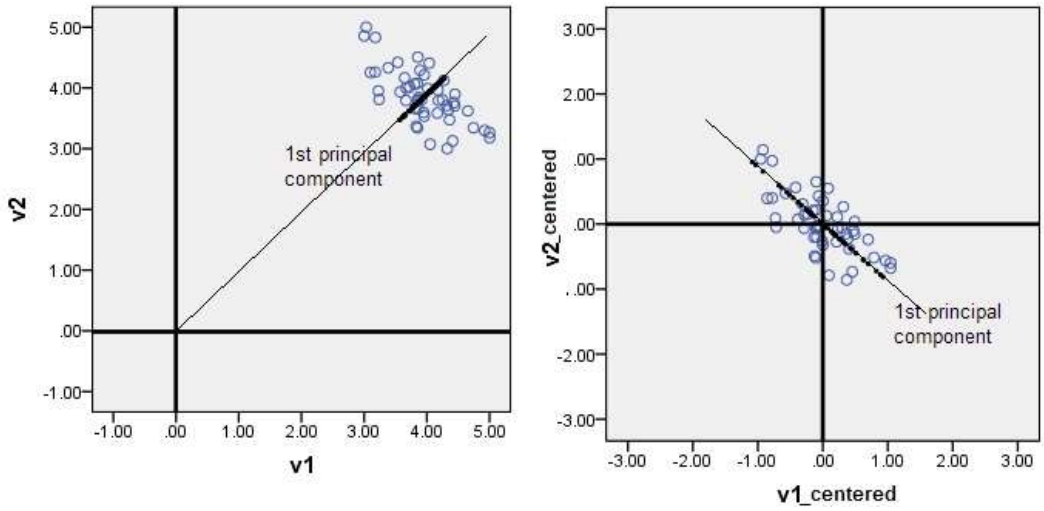
Τα SVM μπορούν και διαχειρίζονται μεγάλο αριθμό από features καλά [16] αρκεί να υπάρχει ένας εξίσου μεγάλος αριθμός από δεδομένα εκπαίδευσης. Αυτό όμως δεν σημαίνει πως αποδίδουν και (πάντα) καλύτερα με περισσότερα features. Στην εργασία μας τα features εξάγονται με την χρήση της NN-CPD και το πλήθος τους εξαρτάται από το rank του decomposition, επομένως μπορούν εύκολα να ξεπεράσουν τα 50. Ένας τρόπος να γίνει μείωση της διαστατικότητας (dimensionality reduction) πετώντας πιθανώς άχρηστα features είναι κάνοντας χρήση της Principal Component Analysis (PCA) στο feature matrix των δεδομένων της εκπαίδευσης και της πρόβλεψης μαζί, με την ίδια βάση.



Σχήμα 4.13: Παράδειγμα PCA σε 2-D δεδομένα. Δεξιά φαίνονται τα scores.

Η PCA είναι ένας ορθογώνιος μετασχηματισμός ο οποίος προβάλλει ένα σύνολο δεδομένων σε ένα νέο σύστημα συντεταγμένων. Στο αρχικό σύστημα συντεταγμένων του συνόλου δεδομένων οι άξονες είναι ένα σύνολο από πιθανώς συσχετιζόμενες μεταβλητές. Οι νέοι άξονες είναι μεταβλητές που ονομάζονται principal components και είναι γραμμικά ανεξάρτητες μεταξύ τους. Αποτελούν γραμμικούς συνδυασμούς των αρχικών μεταβλητών και δημιουργούνται με τέτοιο τρόπο ώστε η πρώτη να έχει την μεγαλύτερη δυνατή διακύμανση, η δεύτερη – με τον περιορισμό της ορθογωνιότητας ως προς την πρώτη – επίσης να έχει την μέγιστη δυνατή διακύμανση κ.ο.κ.. Το αποτέλεσμα της ανάλυσης αυτής είναι ένα μητρώο *component coefficients / loadings* και ένα δεύτερο μητρώο *component / factor scores*. Το πρώτο περιέχει τα διανύσματα της νέας ορθογώνιας βάσης (τα principal components) ενώ το δεύτερο τα μετασχηματισμένα δεδομένα.

Ένα σημείο παγίδα στην εφαρμογή της PCA είναι το data pretreatment στα αρχικά δεδομένα. Αρχικά κάθε μεταβλητή του συνόλου δεδομένων πρέπει να κεντραριστεί με την αφαίρεση της μέσης τιμής της. Αυτό το βήμα θεωρείται απαραίτητο για την ορθότητα της PCA αλλιώς το πρώτο principal component μπορεί να επιλεχθεί με λάθος τρόπο, όπως φαίνεται στο ακόλουθο σχήμα. Γι' αυτό και οι περισσότερες υλοποιήσεις της PCA εφαρμόζουν αυτομάτως centering στα αρχικά δεδομένα.



Σχήμα 4.14: Επιλογή του πρώτου principal component χωρίς και με centering.

Εκτός από το centering η PCA είναι πολύ ευαίσθητη στο scaling των αρχικών δεδομένων. Εδώ υπάρχουν άπειρες επιλογές και καμία δεν μπορεί να θεωρηθεί η μόνη ορθή. Σε περίπτωση που δεν εφαρμοστεί καθόλου scaling τότε έχουμε την λεγόμενη PCA βάσει συνδιασποράς, ενώ αν εφαρμοστεί standardization (διακύμανση κάθε μεταβλητής ίση με 1) τότε έχουμε PCA βάσει συσχέτισης. Όπως είπαμε οι επιλογές είναι άπειρες και καμία πιο ορθή από την άλλη. Μπορεί να εφαρμοστεί scaling κάθε μεταβλητής στο διάστημα [0,1], στο διάστημα [-1,1], λογαριθμικός μετασχηματισμός ή οτιδήποτε άλλο δουλεύει καλά για τα δεδομένα του κάθε προβλήματος.

Σε ένα πρόβλημα κατηγοριοποίησης συνηθίζεται η εφαρμογή της PCA πάνω στο feature matrix. Στο μητρώο scores του αποτελέσματος εμφανίζεται ένα σύνολο από νέα features ίδιου πλήθους τα οποία σε αντίθεση με τα αρχικά δεν έχουν φυσική σημασία (αφού αποτελούν γραμμικούς μετασχηματισμούς των αρχικών), όμως έχουν την εξής ιδιότητα: είναι ταξινομημένα σύμφωνα με το «variance explained» τους, με άλλα λόγια το κατά πόσο «εξηγεί» το κάθε ένα την διακύμανση στα αρχικά δεδομένα. Το άθροισμα των τιμών αυτών για όλα τα principal components ονομάζεται «variance retained» και αρχικά αθροίζεται στο

1. Μπορούμε αναλόγως τα δεδομένα να αγνοήσουμε ένα πλήθος από τις τελευταίες στήλες του μητρώου scores αν τα αντίστοιχα principal components έχουν πολύ μικρό «variance explained», χωρίς να χάσουμε σημαντική πληροφορία, δηλαδή διατηρώντας ένα υψηλό «variance retained».

Καλό είναι να σημειωθεί επίσης πως ένα feature με μεγάλη διακύμανση δεν είναι απαραίτητα και ένα καλό feature. Αυτό δουλεύει και αντίστροφα. Για παράδειγμα ένα feature μπορεί να έχει μεγάλη διακύμανση μόνο και μόνο από τα δεδομένα μιας κλάσης, χωρίς αυτό να σημαίνει κάτι για την διαχωριστικότητα των κλάσεων. Επίσης ένα feature μπορεί να έχει πολύ μικρή διακύμανση αλλά να είναι πολύ σημαντικό για την κατηγοριοποίηση. Επομένως το να μειώσουμε τις διαστάσεις με ένα πολύ μικρό «variance retained» δεν είναι κακό. Μάλιστα στα πειράματα μας είδαμε πάρα πολύ καλά αποτελέσματα και με «variance retained» της τάξης του 10%.

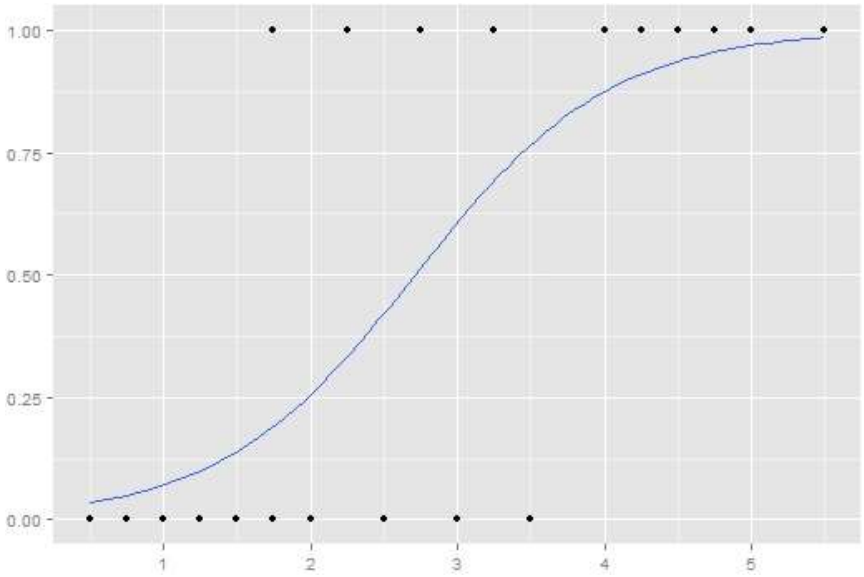
Η υλοποίηση της PCA που χρησιμοποιήσαμε στην εργασία μας είναι η εντολή «pca» που περιλαμβάνεται στην MATLAB R2018b.

4.5. Βαθμονόμηση Εξόδου για One-Class SVM

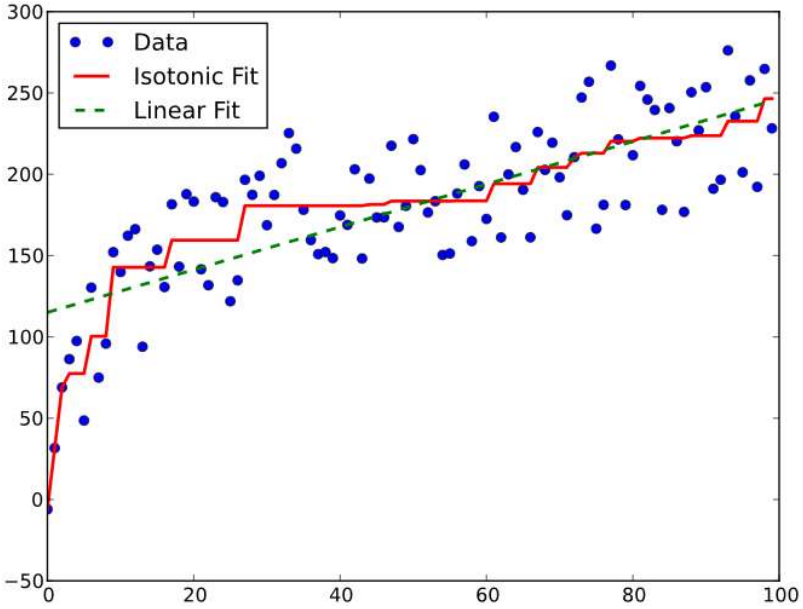
Όπως αναφέρθηκε στο κεφάλαιο 3 τα SVM είναι μη-πιθανοτικοί κατηγοριοποιητές. Το μόνο που μπορούν να μας πουν είναι σε ποια κλάση ανήκει ένα αντικείμενο βάσει του σε ποια μεριά του ορίου απόφασης «πέφτει» το διάνυσμα χαρακτηριστικών του, χωρίς κάποιον απόλυτο δείκτη σιγουριάς. Αυτό σημαίνει μεταξύ άλλων πως δεν υπάρχει σαφής και άμεσος τρόπος σύγκρισης των αποτελεσμάτων διαφορετικών SVM, πράγμα που θέλουμε να κάνουμε στην εργασία αυτή.

Στα αποτελέσματα της πρόβλεψης των SVM συμπεριλαμβάνονται οι τιμές απόφασης (decision values), που είναι στην ουσία η απόσταση κάθε διανύσματος από το όριο απόφασης. Οι τιμές αυτές δεν είναι φραγμένες και δεν μπορούν να χρησιμοποιηθούν σαν πιθανότητες, μπορούν όμως να μετατραπούν σε κάποιου είδους ψευδο-πιθανότητες με την χρήση ενός πιθανοτικού μοντέλου. Τέτοια μοντέλα συνήθως δημιουργούνται με βάση τις τιμές απόφασης που προκύπτουν από labeled δεδομένα, τα οποία χρησιμοποιούνται και για την εκπαίδευση των SVM.

Για την περίπτωση των two-class SVM έχει γίνει εκτενής έρευνα και υπάρχουν εδραιωμένες μέθοδοι εκτίμησης τέτοιων ψευδο-πιθανοτήτων, όπως το Platt’s scaling (logistic regression) [24]-[25] και το Isotonic Regression [26]. Στην περίπτωση των one-class SVM όμως κάτι τέτοιο είναι πιο δύσκολο και δεν έχει μελετηθεί εξίσου καλά.



Σχήμα 4.15: Παράδειγμα Logistic Regression.



Σχήμα 4.16: Παράδειγμα Isotonic Regression.

Παρ' ότι τα one-class SVM δημιουργήθηκαν και χρησιμοποιούνται για περιπτώσεις που υπάρχουν μόνο «θετικά» δείγματα, στην δική μας περίπτωση υπάρχουν δεδομένα από όλες τις κλάσεις. Επομένως το πρώτο πράγμα που δοκιμάσαμε ήταν να εφαρμόσουμε τις ίδιες τακτικές που υπάρχουν και για τα two-class SVM. Το αποτέλεσμα δεν ήταν καλό και μπορούμε να το εξηγήσουμε ως εξής. Τα two-class SVM προσπαθούν να διαχωρίσουν τα διανύσματα της μιας κλάσης από αυτά της άλλης στον διανυσματικό χώρο. Στα one-class SVM, στην περίπτωση του γραμμικού kernel, το όριο απόφασης είναι κάθετο στην ευθεία που περνάει από το κέντρο των δεδομένων και το κέντρο του διανυσματικού χώρου, αφού θεωρούν πως πιθανά αρνητικά διανύσματα βρίσκονται με μεγαλύτερη πιθανότητα αντιδιαμετρικά του κέντρου. Είναι προφανές λοιπόν ότι στα one-class SVM τα όρια απόφασης μπορούν να είναι πολύ διαφορετικά από αυτά των two-class SVM και ότι τα decision values που προκύπτουν από αρνητικά δεδομένα δεν είναι πάντα ενδεικτικά της απόστασης από την θετική κλάση, πράγμα βασικό για την εξαγωγή πιθανοτήτων.

Αφού απορρίψαμε τις τεχνικές των two-class SVM μελετήσαμε την σχετική βιβλιογραφία για τα one-class SVM και επιλέχθηκε η μέθοδος της δημοσίευσης των L. P. Jain, W. J. Scheirer και T. E. Boult [33] που ισχυρίζονται ότι είναι και η πρώτη πιθανοτική προσέγγιση για one-class SVM. Συμπληρωματικά επιλέχθηκε και η χρήση της standard logistic function που θα αναλυθεί παρακάτω.

4.5.1. Extreme Value Theory – Weibull

Η μέθοδος αυτή βασίζεται στην θεωρία των ακραίων τιμών (extreme value theory – EVT) η οποία έχει μελετηθεί σε μια σειρά από δημοσιεύσεις [33]-[34]-[35]-[36] πάνω σε προβλήματα κατηγοριοποίησης. Σύμφωνα με αυτήν μπορούμε να αντλήσουμε χρήσιμες πληροφορίες από ακραίες τιμές. Στην περίπτωση μας οι ακραίες τιμές είναι τα μικρά θετικά decision values, δηλαδή δεδομένων που βρίσκονται πολύ κοντά στο όριο απόφασης ενός one-class SVM, από την μεριά της θετικής κλάσης. Σε αυτό το στάδιο αναφερόμαστε αποκλειστικά σε δεδομένα εκπαίδευσης, τα οποία λόγω του σφάλματος εκπαίδευσης δεν παίρνουν όλα θετικά decision values.

Κάνοντας fit μια Weibull PDF σε ένα υποσύνολο (εξίσωση 4.3) [33] αυτών των τιμών παίρνουμε την αντίστοιχη CDF. Μέσω του θεωρήματος του Bayes, όπως εξηγείται στην έρευνα [33], αυτή η CDF μπορεί να μας δώσει για οποιοδήποτε decision value την πιθανότητα να προέκυψε από ένα θετικό αντικείμενο. Η πιθανότητα αυτή στην δημοσίευση αναφέρεται ως μη-κανονικοποιημένη καθώς εκεί πραγματεύονται προβλήματα ανοιχτού συνόλου κλάσεων και δεν μπορεί να υλοποιηθεί κανονικά το θεώρημα του Bayes (θέτουν τον παρονομαστή ίσο με 1). Δοκιμάζοντας να χρησιμοποιήσουμε ολόκληρο τον τύπο του Bayes (εξίσωση 4.2) μιας και τα προβλήματα μας είναι κλειστού τύπου δεν παρατηρήθηκε διαφορά στα αποτελέσματα σε σχέση με την «λειψή» εκδοχή του paper, παρά μόνο αυξημένη πολυπλοκότητα.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B|A)\,P(A) + P(B|\neg A)\,P(\neg A)} \qquad (4.2)$$

Όσον αφορά το υποσύνολο που αναφέρθηκε στην προηγούμενη παράγραφο, στο paper αναφέρουν σαν καλή επιλογή τα n μικρότερα θετικά decision values με το n να προκύπτει από την εξής σχέση:

$$n = 1.5 * |support\_vectors^{+}| \text{ , } n >= 3 \qquad (4.3)$$

Τα $support\_vectors^{+}$ είναι τα support vectors του SVM που έχουν θετικό decision value, και οι τιμές πρέπει να είναι τουλάχιστον τρείς για να μπορέσει να γίνει το fit της PDF. Η λογική πίσω από αυτήν την σχέση είναι πως ο αριθμός των support vectors ενός SVM μαρτυράει την πολυπλοκότητα του ορίου απόφασης και πως το πλήθος n πρέπει να είναι ανάλογο αυτής.

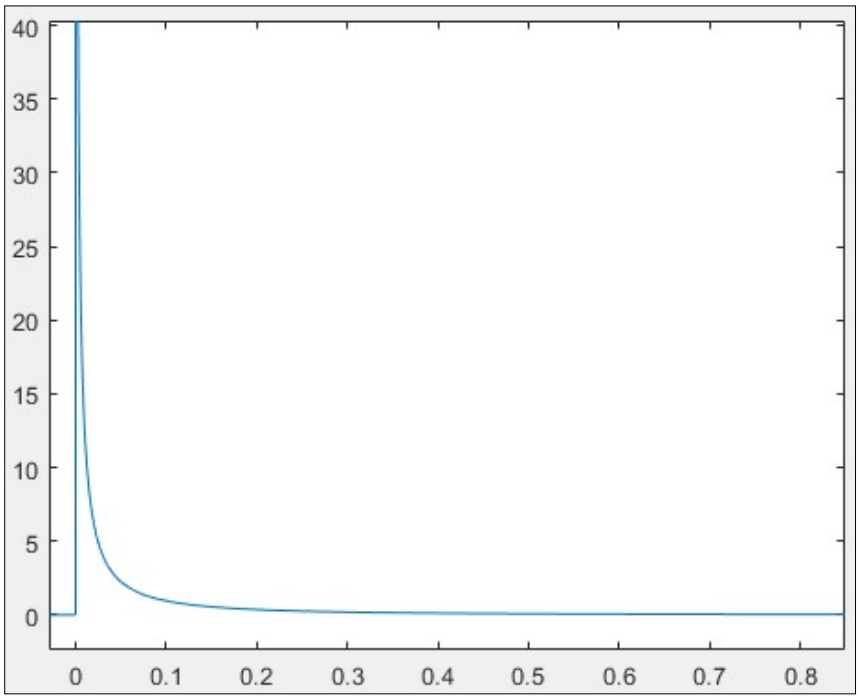
1. Υπολόγισε τα prior class probabilities από το training set.
2. Για κάθε one-class SVM ύστερα από την εκπαίδευση:
   α. Υπολόγισε τα decision values για τα δεδομένα εκπαίδευσης του.
   β. Κράτα μόνο τα θετικά.
   γ. Από αυτά κράτα μόνο τα n μικρότερα, n = 1.5 * |support_vectors^{+}|, n > 3
   δ. Κάνε fit μια Weibull PDF στα εναπομείναντα decision values.
3. Για κάθε one-class SVM κατά την πρόβλεψη:
   α. Πρόβαλε τα decision values στην αντίστοιχη Weibull CDF.
   β. Πολλαπλασίασε με το αντίστοιχο prior class probability.

Πίνακας 4.1: Η μέθοδος του EVT σε φυσική γλώσσα.

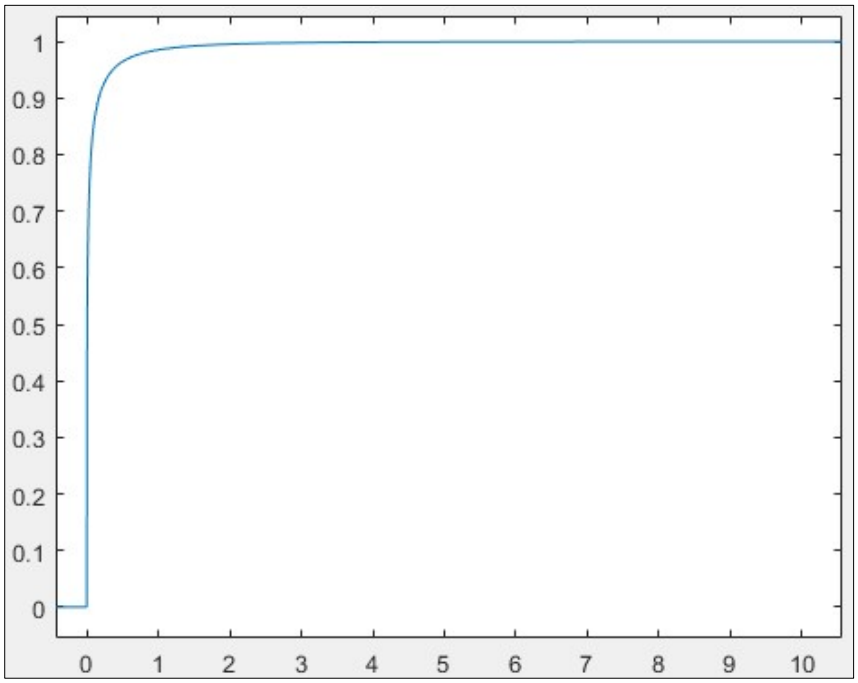
Οι μαθηματικοί τύποι της Weibull κατανομής είναι οι ακόλουθοι:

$$PDF_{Weibull}(x|a,b) = \frac{b}{a}(\frac{x}{a})^{b-1}e^{-(\frac{x}{a})^{b}} \quad (4.4)$$

$$CDF_{Weibull}(x|a,b) = 1 - e^{-(\frac{x}{a})^{b}} \quad (4.5)$$



Σχήμα 4.17: Η PDF της Weibull για a≈0.0143 και b≈0.3389.



Σχήμα 4.18: Η CDF της Weibull για a≈0.0143 και b≈0.3389.

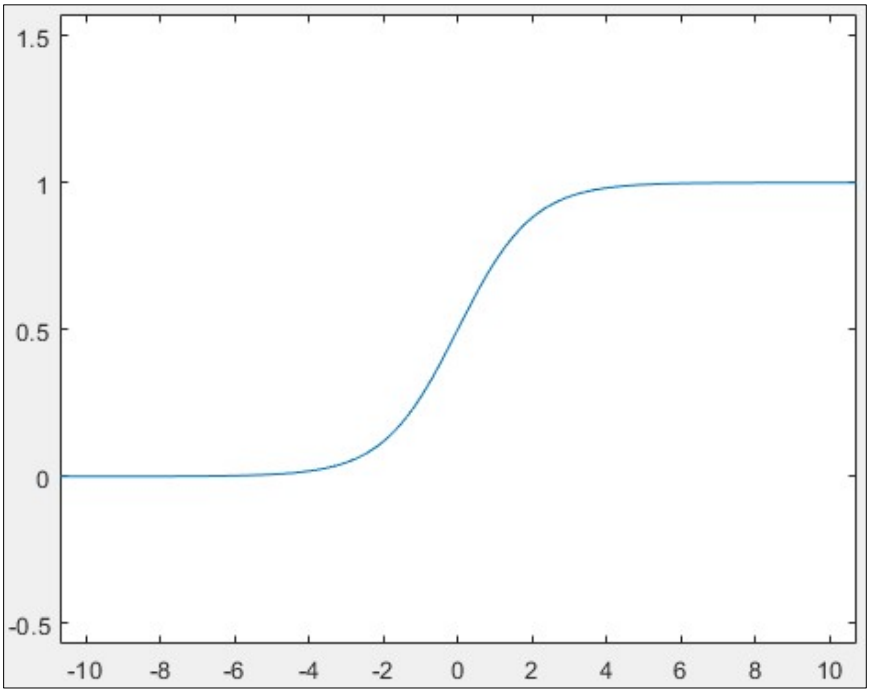
Όπως διαπιστώθηκε από τα πειράματα, η μέθοδος αυτή δείχνει την δύναμη της όταν τα one-class SVM εκπαιδεύονται με διαφορετικές υπερ-παραμέτρους, εκεί που η επόμενη μέθοδος υστερεί.

Για το fitting των Weibull PDF στα decision values έγινε χρήση της συνάρτησης «wblfit» που περιλαμβάνεται στην MATLAB R2018b.

4.5.2. Standard Logistic Function

Η δεύτερη μέθοδος για την βαθμονόμηση των decision values που υλοποιήθηκε στην εργασία είναι η προβολή στην standard logistic function η οποία είναι μια απλή σιγμοειδής συνάρτηση χωρίς παραμέτρους. Η μέθοδος αυτή μας επιτρέπει να φράξουμε τις τιμές μας στο διάστημα [0,1] χωρίς όμως οι νέες τιμές να αντιπροσωπεύουν κάποιου είδους πιθανότητες. Ο μαθηματικός τύπος της standard logistic function είναι:

$$f(x) = \frac{1}{1+e^{-x}} \qquad (4.6)$$



Σχήμα 4.19: Οπτικοποίηση της standard logistic function.

Αυτή η μέθοδος μπορεί να δείχνει «χαζή» αλλά όπως θα φανεί και από τα πειράματα είναι πάρα πολύ αποτελεσματική όταν τα one-class SVM εκπαιδεύονται με τις ίδιες υπερ-παραμέτρους.

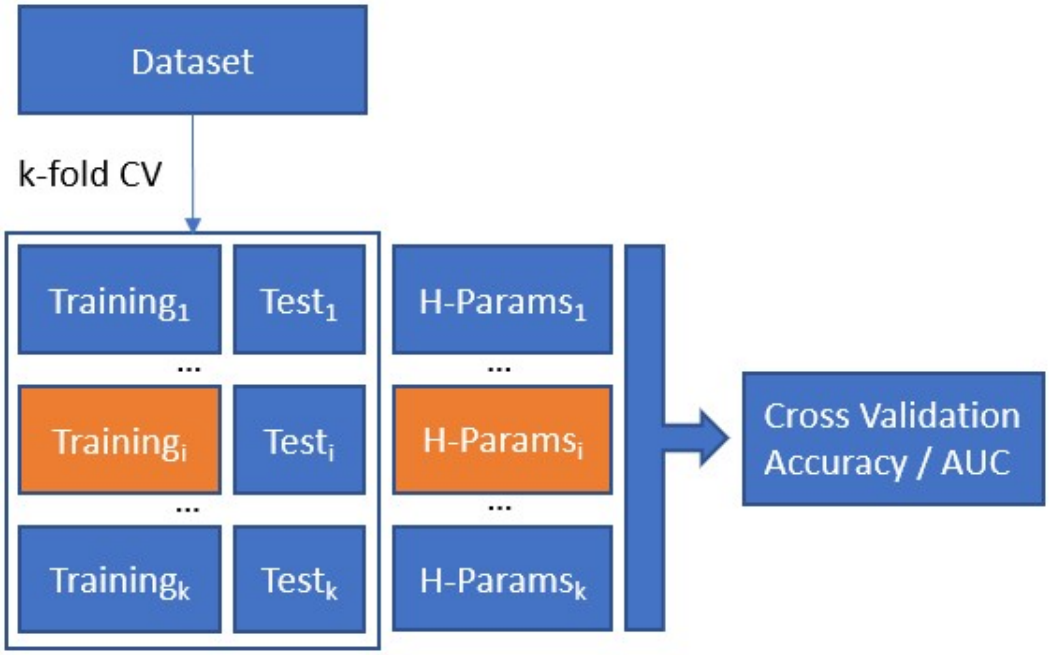
4.6. Εμφωλευμένο Cross Validation – Βελτιστοποίηση Υπερ-Παραμέτρων

Για την αξιολόγηση του προτεινόμενου αλγορίθμου επιλέχθηκε να γίνει stratified cross-validation, αφού είναι η πλέον προτιμώμενη μέθοδος και είναι ενδεικτική των αναμενόμενων επιδόσεων σε πραγματικά σενάρια χρήσης. Η λέξη stratified σημαίνει πως φροντίζουμε κάθε fold να περιέχει περίπου την ίδια κατανομή κλάσεων στα training και στα test sets.

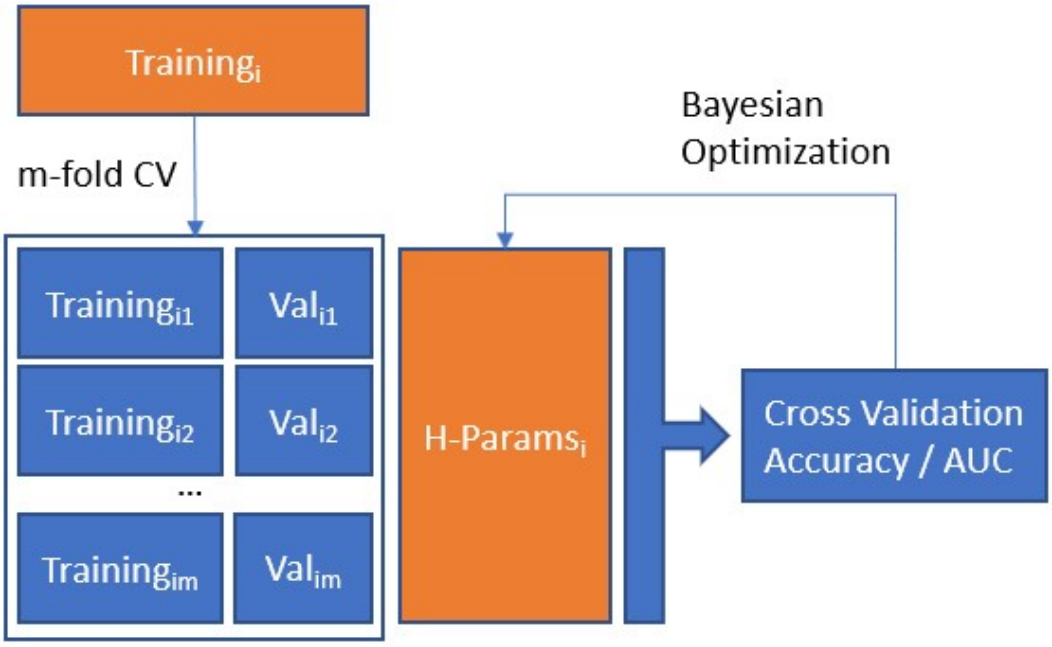
Ο αλγόριθμος που αναπτύχθηκε χρειάζεται κάποιες υπερ-παραμέτρους για να δουλέψει. Καλούμαστε επομένως να βρούμε έναν τρόπο βελτιστοποίησης αυτών των υπερ-παραμέτρων (hyper-parameter tuning) ώστε να πετύχουμε τις βέλτιστες επιδόσεις για κάθε fold του cross-validation.

Ένας απαράβατος κανόνας για το hyper-parameter tuning είναι πως δεν πρέπει να χρησιμοποιούνται τα δεδομένα που προορίζονται για την πρόβλεψη και αξιολόγηση, καθώς έτσι δεν μπορεί να αξιολογηθεί η ικανότητα γενίκευσης του αλγορίθμου αφού πολύ εύκολα μπορεί να γίνει over-fitting.

Σύμφωνα με τα παραπάνω, η μέθοδος που επιλέχθηκε είναι η ακόλουθη. Για κάθε fold του cross-validation, γίνεται επαναληπτικά ένα ξεχωριστό stratified cross-validation στο training set του με σκοπό το hyper-parameter tuning. Οι υπερ-παράμετροι που αποδίδουν κατά μέσο όρο καλύτερα στα «εσωτερικά» folds είναι και αυτές που θα χρησιμοποιηθούν στο «εξωτερικό» cross-validation.



Σχήμα 4.20: Το «εξωτερικό» cross validation.



Σχήμα 4.21: Το «εσωτερικό» cross validation.

Σαν μέθοδος βελτιστοποίησης δεν επιλέχθηκε το συνηθισμένο grid search αλλά ο Bayesian optimizer [15] της MATLAB R2018b (εντολή «bayesopt») σαν λιγότερο άπληστη προσέγγιση. Σαν loss function για τον Bayesian optimizer χρησιμοποιήσαμε την παρακάτω (εξίσωση 4.7), όπου $\overrightarrow{acc}$ το διάνυσμα με τα test accuracies των «εσωτερικών» folds.

$$f(\overrightarrow{acc}) = 1 - mean(\overrightarrow{acc}) \qquad (4.7)$$

Μετά το hyper-parameter tuning ο αλγόριθμος τρέχει για κάθε «εξωτερικό» fold και προκύπτουν οι δυο μετρικές που σημειώνουμε, ο μέσος όρος του accuracy και ο μέσος όρος του AUC (area under ROC curve) εφόσον πρόκειται για δυαδική κατηγοριοποίηση.

4.7. Μετρικές Accuracy και AUC

Οι μετρικές αξιολόγησης που χρησιμοποιήθηκαν είναι δυο, το απλό accuracy και το area under the ROC (receiver operating characteristic) curve [37]. Ο τύπος για τον υπολογισμό του accuracy είναι ο ακόλουθος:

$$Accuracy = \frac{Σωστές\ Προβλέψεις}{Πλήθος\ Προβλέψεων} \qquad (4.8)$$

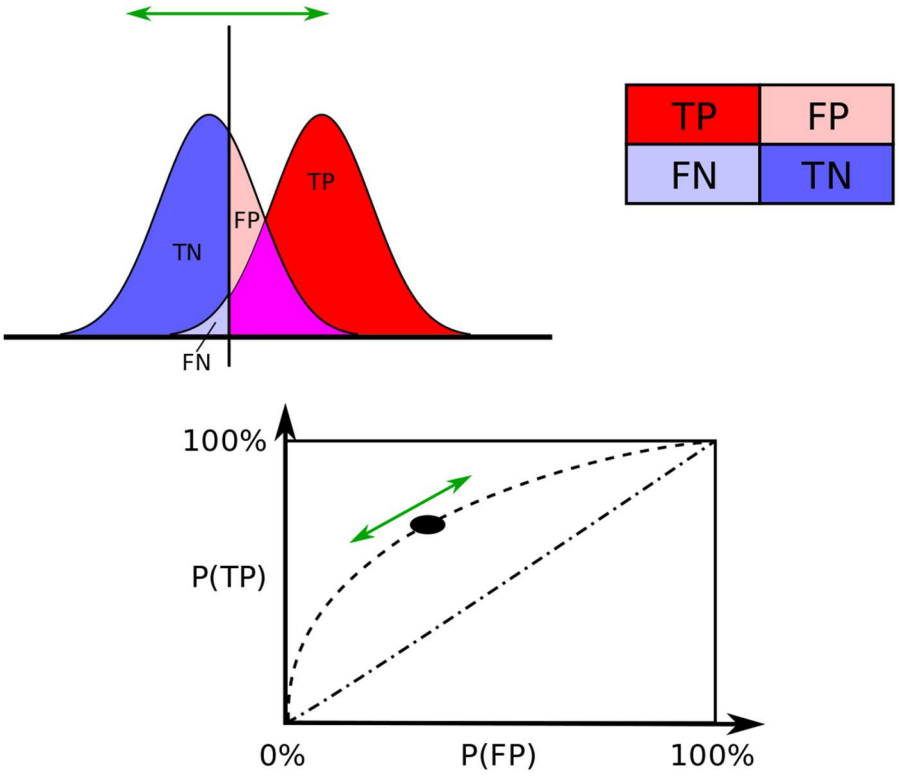
Είναι προφανές πως μια τέτοια μετρική που δεν λαμβάνει υπόψιν την κατανομή των κλάσεων στο σύνολο δεδομένων μπορεί να οδηγήσει σε παραπλανητικά αποτελέσματα. Παρά το γεγονός αυτό, αποτελεί την πιο συνηθισμένη μετρική στην σχετική βιβλιογραφία και είναι απολύτως λογικό να την θέλουμε στα πειράματα μας.

Η επόμενη μετρική είναι αρκετά πιο πολύπλοκη και χρήσιμη ειδικά για έναν αλγόριθμο που αφορά την διάγνωση κάποιας ασθένειας. Το AUC είναι το εμβαδόν κάτω από την καμπύλη ROC [37] ενός δυαδικού κατηγοριοποιητή και δείχνει τι σχέση υπάρχει ανάμεσα στα true positives και στα false positives που προκύπτουν από αυτόν για ένα συγκεκριμένο σύνολο δεδομένων. Ορίζεται μόνο για δυαδικά προβλήματα.

Ένας δυαδικός κατηγοριοποιητής συνήθως παίρνει απόφαση για ένα αντικείμενο σύμφωνα με την τιμή κάποιας συνεχούς μεταβλητής, έστω X, η οποία μπορεί να θεωρηθεί σαν κάποιου είδους «σκορ» για το αντικείμενο. Στην περίπτωση μας, όπως θα γίνει κατανοητό και στο κεφάλαιο 5, αυτή η μεταβλητή είναι για κάθε bag η διαφορά της bag-level μετρικής του πρώτου SVM από αυτήν του δεύτερου. Η μεταβλητή X ακολουθεί μια κατανομή πυκνότητας πιθανότητας $f_1(X)$ σε περίπτωση που το bag είναι όντως θετικό και μια άλλη κατανομή $f_0(X)$ σε περίπτωση που το bag είναι όντως αρνητικό. Ένα παράδειγμα δυο τέτοιων κατανομών φαίνεται στο σχήμα 4.22. Δεδομένης κάθε φορά μιας τιμής T (threshold, κάθετη γραμμή στο σχήμα) ένα bag χαρακτηρίζεται θετικό αν η μεταβλητή X που προκύπτει από αυτό είναι μεγαλύτερη από το T και αρνητικό αν η X είναι μικρότερη από το T. Έτσι λοιπόν για κάθε T ορίζεται ένα ποσοστό true positives (TP), false positives (FP), false negatives (FN) και true negatives (TN), με τα TP και FP να δίνονται από τους ακόλουθους τύπους:

$$TP(T) = \int_T^\infty f_1(x)dx \qquad FP(T) = \int_T^\infty f_0(x)dx \qquad (4.9)$$

Η καμπύλη ROC δείχνει το ποσοστό των true positives για κάθε ποσοστό των false positives. Αυτό που κάνει ένας αλγόριθμος για να την δημιουργήσει είναι να μεταβάλλει την τιμή T ώστε να βρίσκει διαφορετικά ζεύγη TP-FP.



Σχήμα 4.22: Οπτικοποίηση των κατανομών της εξόδου ενός δυαδικού κατηγοριοποιητή για κάποια δεδομένα και της αντίστοιχης καμπύλης ROC.

Η μετρική του AUC δεν είναι κάτι άλλο παρά το εμβαδόν κάτω από αυτήν την καμπύλη. Είναι προφανές πως ένα μεγάλο AUC υποδηλώνει λιγότερη επικάλυψη μεταξύ των κατανομών της εξόδου του κατηγοριοποιητή για τα θετικά και τα αρνητικά δείγματα. Πιο απλά, ένα μεγαλύτερο AUC είναι δείγμα καλύτερου διαχωρισμού των κλάσεων από τον δεδομένο κατηγοριοποιητή για ένα συγκεκριμένο σύνολο δεδομένων.

Για τον υπολογισμό της μετρικής του AUC έγινε χρήση της συνάρτησης «perfcurve» που περιλαμβάνεται στην MATLAB R2018b.

Κεφαλαιο 5: Κατηγοριοποιητης MIL με One-Class SVM

5.1. Είσοδος και Έξοδος του Αλγορίθμου

Ήρθε η ώρα να μιλήσουμε για τον αλγόριθμο που αναπτύχθηκε στα πλαίσια αυτής της εργασίας. Πρόκειται για έναν αλγόριθμο κατηγοριοποίησης MIL ο οποίος κάνει χρήση πολλαπλών one-class SVM με τον RBF kernel που αναφέρθηκε στην ενότητα 3.3. Αρχικά πρέπει να περιγράψουμε την είσοδο και την έξοδο του αλγορίθμου μας.

Η συνηθισμένη αναπαράσταση του dataset όταν πρόκειται για MIL προβλήματα είναι η εξής: (α) ένας πίνακας X που ονομάζεται feature matrix διαστάσεων [#_instances x #_features] που περιέχει το feature vector κάθε στιγμιότυπου, (β) ένα διάνυσμα Y μήκους [#_instances] που περιέχει το class label κάθε στιγμιότυπου και (γ) ένα διάνυσμα ID μήκους [#_instances] που περιέχει το bag ID που αντιστοιχεί σε κάθε στιγμιότυπο.

Για την δημιουργία αυτού του dataset περνάμε ένα σύνολο από εικόνες από τις διαδικασίες του stain color normalization (ενότητα 4.1), του κατακερματισμού (ενότητα 4.2) και του feature extraction (ενότητα 4.3).

Για λόγους ευκολίας η υλοποίηση μας παίρνει το dataset ενιαίο στην είσοδο μαζί με μια δομή data_sets που περιέχει πληροφορία (λίστες με bag ID) για το πως αυτό θα διασπαστεί σε training set και test set, διατηρώντας την ακεραιότητα των bags.

Η υπόλοιπη είσοδος αποτελείται από υπερ-παραμέτρους. Το διάνυσμα h_params περιέχει τις υπερ-παραμέτρους εκπαίδευσης για τα one-class SVM και το variance retained που επιθυμούμε από την PCA στην αρχή του αλγορίθμου. Τα cal_mode και dec_mode προσδιορίζουν με ποιον τρόπο ο αλγόριθμος μας θα κάνει την βαθμονόμηση των decision values και ποια bag-level μετρική θα χρησιμοποιήσει για την σύντηξη τους.

Αυτό που παίρνουμε σαν έξοδο από τον αλγόριθμο μας είναι η μετρική accuracy (ποσοστό σωστών προβλέψεων) και στην περίπτωση της δυαδικής κατηγοριοποίησης το AUC (area under ROC curve). Η μετρική του AUC ορίζεται μόνο για προβλήματα δυο κλάσεων αφού προκύπτει από την σχέση των true positives με τα false positives όπως εξηγήσαμε στην ενότητα 4.7.

5.2. Bag-Level Μετρικές

Ο αλγόριθμος μας είναι ένας instance space MIL κατηγοριοποιητής. Αυτό σημαίνει πως αρχικά παράγει κάποιες τιμές σε επίπεδο στιγμιότυπων οι οποίες στην συνέχεια πρέπει να μετατραπούν σε τιμές για τα αντικείμενα.

Πιο συγκεκριμένα, τα διανύσματα χαρακτηριστικών κάθε αντικειμένου στο test set δοκιμάζονται από κάθε SVM του αλγορίθμου μας και προκύπτει ένα σύνολο από instance-level decision values από κάθε SVM για κάθε αντικείμενο. Αυτά τα decision values πρέπει να μετατραπούν σε bag-level decision values ώστε να μπορέσει στην συνέχεια να γίνει μια πρόβλεψη για το label του κάθε αντικειμένου. Οι μετρικές που χρησιμοποιήσαμε για αυτόν τον σκοπό είναι τρεις: (α) entropy, (β) mean και (γ) mean & informative windows.

- Entropy:

Μετράμε την εντροπία στα instance-level decision values ενός αντικειμένου ξεχωριστά για κάθε SVM.

- Mean:

Μετράμε τον αριθμητικό μέσο των instance-level decision values ενός αντικειμένου ξεχωριστά για κάθε SVM.

- Mean & Informative Windows:

Για αυτήν την μετρική ορίσαμε τα informative windows. Αυτά είναι τα στιγμιότυπα για τα οποία συμφωνούν όλα τα SVM, για τα οποία δηλαδή μόνο ένα SVM παρήγαγε θετικό decision value. Η ιδέα πίσω από τα informative windows είναι πως λόγω της συμφωνίας των SVM η πιθανότητα αυτά να περιέχουν «κακή» πληροφορία φθίνει. Αφού βρεθούν τα informative windows, γίνεται ο υπολογισμός της μετρικής ως εξής. Αν ένα αντικείμενο περιέχει τέτοια στιγμιότυπα τότε μετράμε τον αριθμητικό μέσο των instance-level decision values μόνο αυτών των στιγμιότυπων του αντικειμένου ξεχωριστά για κάθε SVM. Για τα αντικείμενα που δεν έχουν informative windows εφαρμόζεται η προηγούμενη μετρική του απλού mean.

5.3. Μέθοδος

Ο αλγόριθμος αρχικά εκτελεί μια PCA στο feature matrix X όπως περιγράφεται στην ενότητα 4.4 χρησιμοποιώντας το variance retained που περιέχεται στο διάνυσμα h_params στην είσοδο. Ύστερα χωρίζει το data set σε training set και test set σύμφωνα με τη δομή data_sets. Από το training set υπολογίζονται τα prior class probabilities.

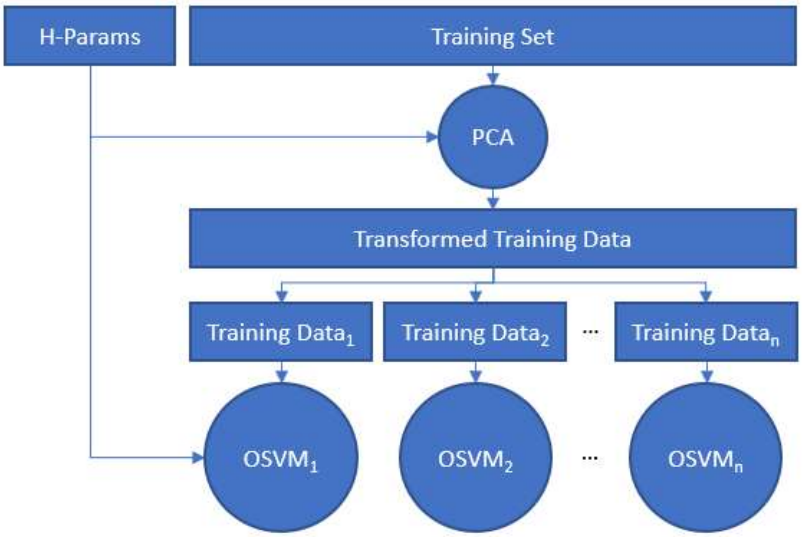
Το επόμενο βήμα είναι η εκπαίδευση των one-class SVM που είναι τόσα όσες είναι και οι κλάσεις του προβλήματος. Το κάθε ένα από αυτά εκπαιδεύεται μόνο με τα «δικά του» instances, αυτά δηλαδή που έχουν το αντίστοιχο class label. Τα SVM που χρησιμοποιούνται είναι soft-margin one-class nu-SVM που κάνουν χρήση του RBF kernel (ενότητα 3.3). Επομένως οι υπερ-παράμετροι για την εκπαίδευση τους είναι δυο, το nu που ρυθμίζει πόσο μικρό ή μεγάλο margin είμαστε διατεθειμένοι να έχουμε, και το gamma που όσο πιο μικρό είναι τόσο πιο «απλό» decision boundary θα έχουμε. Οι υπερ-παράμετροι gamma και nu δίνονται στο διάνυσμα h_params της εισόδου το οποίο μπορεί να περιέχει ένα μόνο σετ gamma και nu για όλα τα one-class SVM ή και ξεχωριστά σετ για το καθένα.

Στο πέρας της εκπαίδευσης των one-class SVM ο αλγόριθμος παίρνει τα τελικά decision values για τα δεδομένα εκπαίδευσης του κάθε one-class SVM και κάνει fit τις Weibull PDF όπως εξηγήσαμε στην υποενότητα 4.5.1. Σε αυτό το σημείο έχουν δημιουργηθεί τόσες Weibull CDF όσα είναι και τα one-class SVM και εδώ τελειώνει η φάση της εκπαίδευσης.

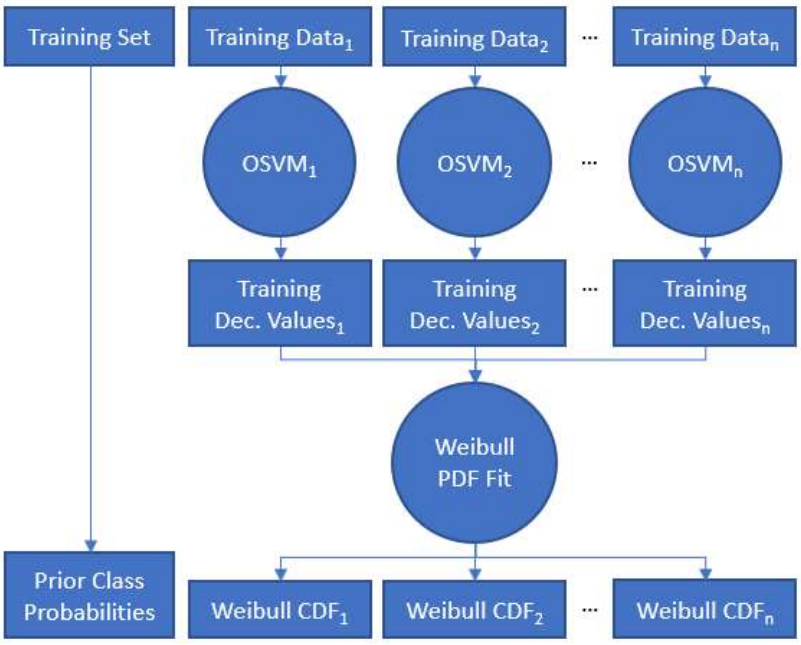
Ξεκινώντας την φάση της πρόβλεψης, ολόκληρο το test set περνάει από κάθε one-class SVM ξεχωριστά. Ανάλογα με την τιμή της μεταβλητής cal_mode γίνεται η βαθμονόμηση των decision values είτε με την μέθοδο της Weibull (υποενότητα 4.5.1) είτε με την standard logistic function (υποενότητα 4.5.2).

Σε αυτό το σημείο έχουμε για κάθε one-class SVM, για κάθε bag του test set, ένα σύνολο από calibrated decision values για τα instances του. Σκοπός του αλγορίθμου είναι να συντήξει αυτά τα decision values σε μια τιμή για κάθε bag, από κάθε one-class SVM. Εδώ ο χρήστης με την μεταβλητή dec_mode μπορεί να επιλέξει ανάμεσα στις τρεις μετρικές entropy, mean και mean & informative windows που περιγράψαμε προηγουμένως στην ενότητα 5.2.

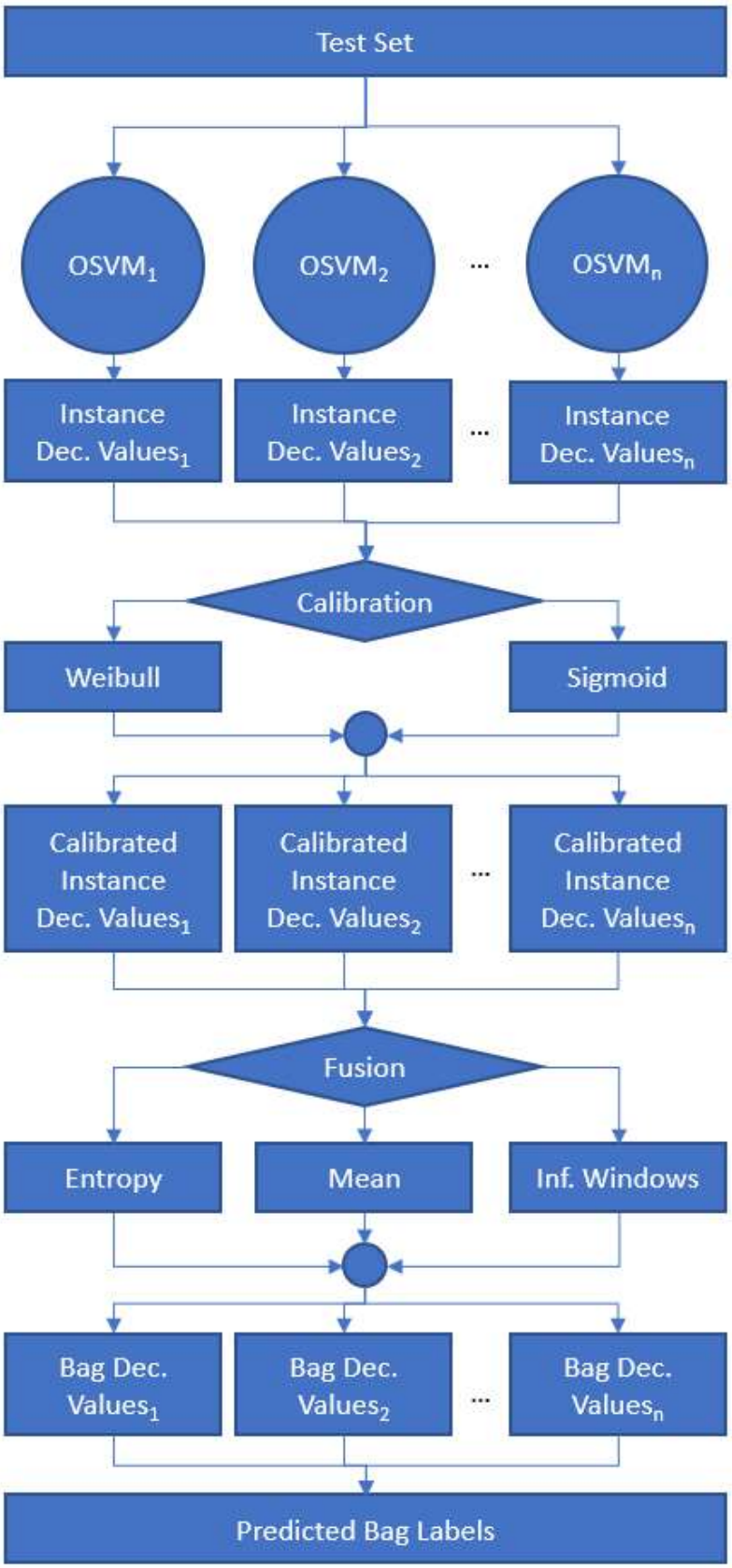
Τέλος, σε κάθε bag ανατίθεται η ετικέτα του one-class SVM με την μεγαλύτερη μετρική, υπολογίζεται το ποσοστό των bags που εκτιμήθηκαν σωστά και σε περίπτωση που το πρόβλημα είναι δυαδικό υπολογίζεται το AUC από τις διαφορές των μετρικών μεταξύ των δυο one-class SVM.



Σχήμα 5.1: Το πρώτο στάδιο της εκπαίδευσης.



Σχήμα 5.2: Το δεύτερο στάδιο της εκπαίδευσης.



Σχήμα 5.3: Το στάδιο της πρόβλεψης.

INPUT:

Το dataset (X, Y, ID), η δομή data_sets που περιέχει τα bag IDs του training και του test set, το διάνυσμα h_params που περιέχει τις υπερ-παραμέτρους gamma, nu και variance retained, οι ακέραιοι cal_mode και dec_mode που προσδιορίζουν τον τρόπο της βαθμονόμησης και της σύντηξης των decision values.

OUTPUT:

Accuracy και AUC εάν αυτό ορίζεται.

ALGORITHM:

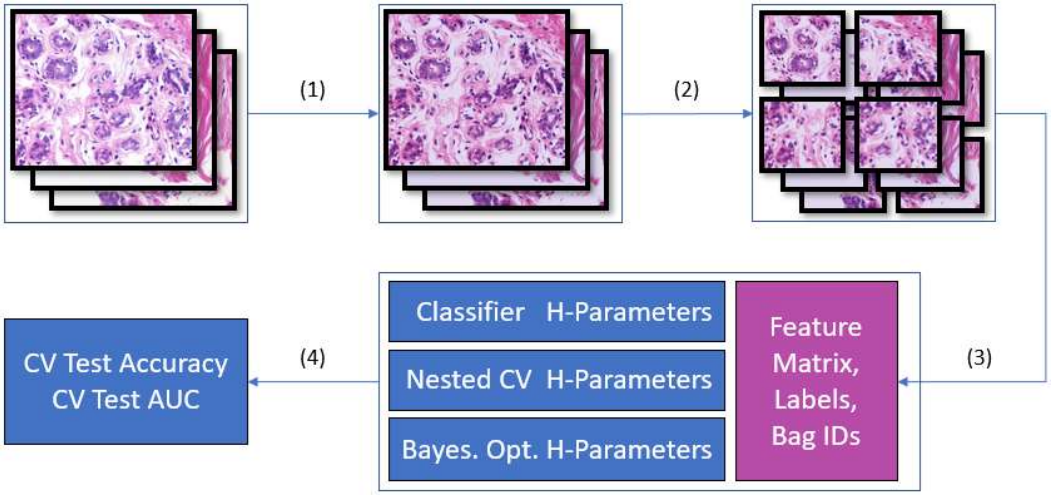
1. Εφάρμοσε PCA στο X με το variance retained από το διάνυσμα h_params.
2. Από το dataset (X, Y, ID) δημιούργησε ένα training set και ένα test set σύμφωνα με τα bag IDs από την δομή data_sets.
3. Από το training set υπολόγισε τα prior class probabilities.
4. Εκπαίδευσε τόσα soft-margin one-class nu-SVM με RBF kernel, όσες είναι και οι κλάσεις του προβλήματος με τα αντίστοιχα δεδομένα από το training set και τις υπερ-παραμέτρους gamma και nu από το διάνυσμα h_params.
5. Για κάθε SVM πάρε τα decision values των δεδομένων εκπαίδευσης και κάνε fit μια Weibull PDF (υποενότητα 4.5.1).
6. Τρέξε το test set με κάθε SVM και βαθμονόμησε τα decision values είτε με τις Weibull CDF (υποενότητα 4.5.1) είτε με την απλή σιγμοειδή (υποενότητα 4.5.2), αναλόγως με την τιμή του cal_mode.
7. Κάνε σύντηξη των calibrated decision values για κάθε SVM με χρήση ενός εκ των τριών bag-level μετρικών entropy, mean ή mean & informative windows, αναλόγως με την τιμή του dec_mode.
8. Σε κάθε bag του test set ανέθεσε την ετικέτα που αντιστοιχεί στο SVM με την μεγαλύτερη bag-level μετρική για αυτό.
9. Υπολόγισε το test accuracy και το AUC εάν αυτό ορίζεται.

Πίνακας 5.1: Ψευδοκώδικας για τον MIL κατηγοριοποιητή μας.

Κεφαλαιο 6: Πειραματα

6.1. Πειραματική Διαδικασία

Στο κεφάλαιο αυτό αρχικά θα μιλήσουμε για την πειραματική διαδικασία και τα μέρη της, εξηγώντας διάφορες επιλογές που έγιναν και μερικά πράγματα για τα πειράματα που θα παρουσιαστούν αργότερα.



Σχήμα 6.1: Το pipeline της πειραματικής διαδικασίας.

Στο παραπάνω σχήμα τα αριθμημένα βέλη αντιπροσωπεύουν τις εξής διαδικασίες: (1) stain color normalization (προαιρετικό) & resolution scaling (προαιρετικό), (2) image segmentation (προαιρετικό), (3) NN-CPD (rank R) & feature scaling (προαιρετικό), (4) nested cross-validation.

Ξεκινώντας από ένα αρχικό dataset με labeled εικόνες το πρώτο βήμα είναι να γίνει το stain color normalization - αν το επιθυμούμε – και επιπλέον να μειωθεί η ανάλυση των εικόνων αν αυτό κρίνεται απαραίτητο για λόγους χωρητικότητας μνήμης αργότερα στην NN-CPD. Στην συνέχεια αναλόγως αν το dataset είναι ήδη multiple instance (BreaKHis) ή όχι (BCC) επιλέγουμε αν θα κάνουμε image segmentation και με τι υπερ-παραμέτρους. Μετά εφαρμόζουμε NN-CPD τάξης R (υπερ-παράμετρος) και προαιρετικά κάποιου είδους scaling στα features που προκύπτουν από αυτή. Ύστερα ορίζουμε τι calibration mode (EVT, SIGMF) και fusion mode (ENTR, MEAN, INFW) θα χρησιμοποιήσει ο classifier, αν τα SVM θα έχουν ξεχωριστές υπερ-παραμέτρους εκπαίδευσης ή όχι, πόσα «εξωτερικά» και «εσωτερικά» folds θέλουμε στο εμφωλευμένο cross validation και τις υπερ-παραμέτρους για τον Bayesian optimizer (τα εύρη τιμών για τις μεταβλητές και τις επαναλήψεις). Τέλος, μένει να τρέξουμε το εμφωλευμένο

cross validation με όλα τα προαναφερθέντα δεδομένα και να πάρουμε τα αποτελέσματα μας.

Τα αποτελέσματα των πειραμάτων μας είναι οι μετρικές CV accuracy και CV AUC, που αποτελούν τον μέσο όρο των αντίστοιχων μετρικών που προκύπτουν από τα «εξωτερικά» folds του nested cross validation, συνοδευόμενες από τις τυπικές αποκλίσεις τους.

Οι σημαντικότερες υπερ-παράμετροι κάθε πειράματος είναι τρείς και είναι αυτές που θα χρησιμοποιηθούν για την ονομασία των αποτελεσμάτων. Η πρώτη είναι ο αριθμός των υπερ-παραμέτρων για την εκπαίδευση των SVM και την PCA (3 ή 5 στην δυαδική κατηγοριοποίηση) και οι άλλες δυο είναι οι μέθοδοι του fusion (INFW, MEAN, ENTR) και του calibration (SIGMF, EVT) των decision values. Για παράδειγμα, η ονομασία «3HP–INFW–SIGMF» υποδηλώνει πως όλα τα SVM χρησιμοποιούν τις ίδιες υπερ-παραμέτρους (ένα ζεύγος gamma και nu), ότι τα decision values βαθμονομήθηκαν με την μέθοδο της standard logistic function και ότι έγινε χρήση της bag-level μετρικής των informative windows.

Εκτός από τις ονομασίες των πειραμάτων πρέπει να εξηγήσουμε και τις ονομασίες των δεδομένων. Στην προκειμένη περίπτωση είναι πιο εύκολο να δώσουμε δυο παραδείγματα:

BCC SCN 5-1-0 R40 [0,1]: Σύνολο δεδομένων BCC με Stain Color Normalization, 5-1-0 segmentation (κατακερματισμός 5-επί-5 με overlap factor 1 και κανόνα απόρριψης 0, δηλαδή χωρίς επικαλύψεις και χωρίς πέταγμα patches), NN-CPD τάξης 40 και scaling του κάθε feature στο διάστημα [0,1].

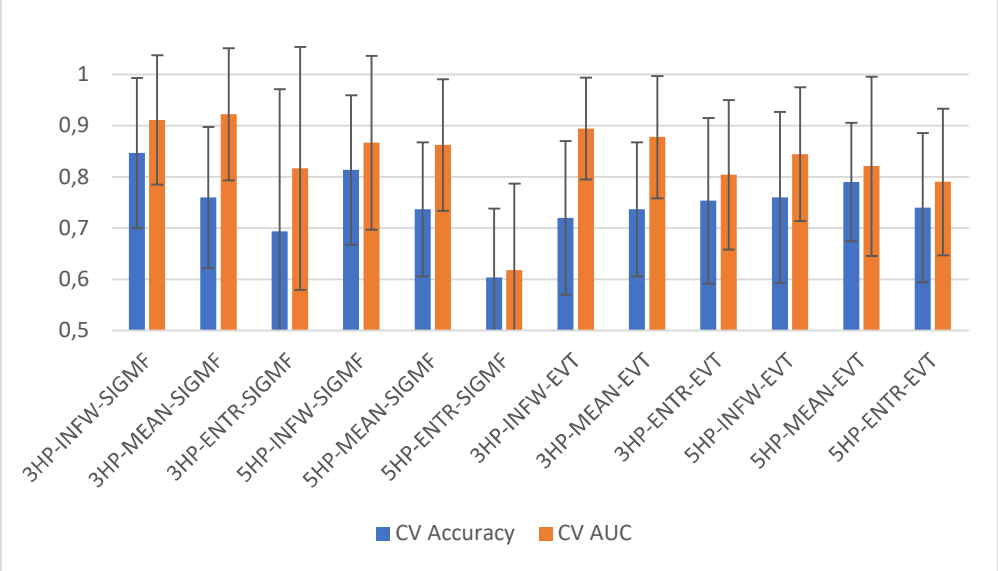
BKH200 33% R40 [0,1]: Σύνολο δεδομένων BreaKHis (εικόνες 200x μεγέθυνσης), χωρίς stain color normalization, με μειωμένη ανάλυση στο 33% της κάθε διάστασης, χωρίς segmentation, με NN-CPD τάξης 40 και scaling του κάθε feature στο διάστημα [0,1].

Πίνακας 6.1: Παραδείγματα ονομασιών δεδομένων.

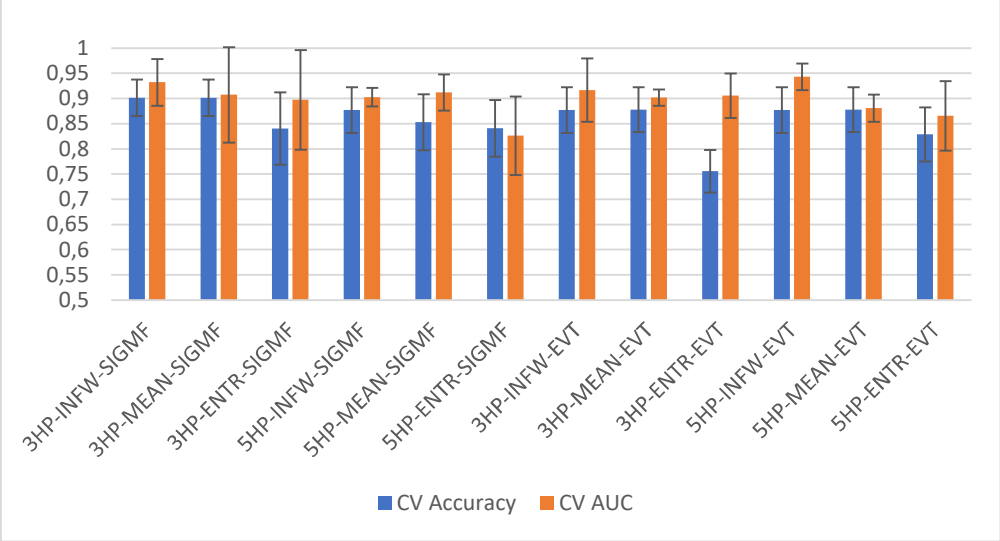
Σε αυτό το σημείο πρέπει να έχει γίνει κατανοητό πως η πειραματική μας διαδικασία είναι παραμετροποιήσιμη σε πάρα πολλά σημεία. Είναι επομένως επιτακτική η ανάγκη να εστιάσουμε σε ένα μικρό υποσύνολο αυτών των παραμέτρων ώστε να μειωθεί ο αριθμός των πειραμάτων και να γίνει εφικτή η αξιολόγηση του αλγορίθμου μας.

6.1.1. Σύγκριση Μεθόδων

Αρχικά θα πρέπει να επιλέξουμε τους καλύτερους από τους δώδεκα στο σύνολο συνδυασμούς υπερ-παραμέτρων (ή αλλιώς μεθόδων) που θα χρησιμοποιήσουμε για να τρέξουμε κάθε (προετοιμασμένο) σύνολο δεδομένων. Για τον σκοπό αυτό δημιουργήθηκαν τα ακόλουθα δυο διαγράμματα από τα καλύτερα δεδομένα του BCC και του BreaKHis.



Σχήμα 6.2: Σύγκριση μεθόδων για τα δεδομένα «BCC SCN 5-1-0 R40 [0,1]».



Σχήμα 6.3: Σύγκριση μεθόδων για τα δεδομένα «BKH200 33% R40 [0,1]».

Κοιτώντας τις πορτοκαλί μπάρες του AUC και στα δυο σχήματα μπορεί κανείς να καταλάβει αμέσως πως σε κάθε περίπτωση το καλύτερο AUC προκύπτει από τις μεθόδους INFW, με τις μεθόδους MEAN να αποδίδουν μερικές φορές σχεδόν το ίδιο (σε περιθώρια λάθους) αλλά ταυτόχρονα με μικρότερο accuracy. Οι μέθοδοι ENTR είναι οι χειρότερες σε κάθε συνδυασμό.

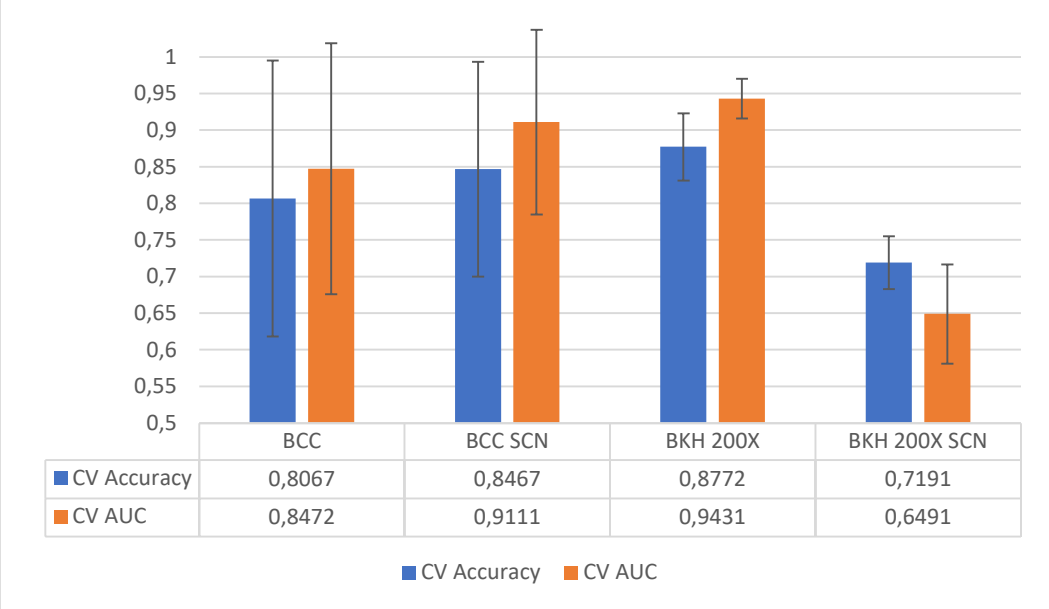
Σε δεύτερο στάδιο, κοιτώντας τις μπλε μπάρες του accuracy ξανά προκύπτει το ίδιο συμπέρασμα. Οι καλύτερες μέθοδοι είναι οι INFW με τις μεθόδους MEAN να αποδίδουν πάλι κοντά σε κάποιες περιπτώσεις αλλά με χειρότερο AUC. Οι μέθοδοι ENTR και πάλι είναι οι χειρότερες σε κάθε συνδυασμό.

Έχοντας καταλήξει στην INFW ως την ανώτερη bag-level μετρική, απομένουν τέσσερις μέθοδοι: «3HP-INF-SIGMF», «5HP-INF-SIGMF», «3HP-INF-EVT» και «5HP-INF-EVT». Οι διαφορές μεταξύ τους είναι μικρές αλλά καταλήξαμε στην πρώτη και την τελευταία συνδυάζοντας τα παραπάνω αποτελέσματα με την εξής λογική: η μέθοδος βαθμονόμησης SIGMF δεν παραμετροποιείται. Αυτό φαίνεται να χαλάει τα αποτελέσματα όταν τα SVM εκπαιδεύονται με διαφορετικές υπερ-παραμέτρους. Αντιθέτως η μέθοδος EVT με την Weibull παραμετροποιείται ξεχωριστά για κάθε SVM, επιτρέποντας στον αλγόριθμο μας να εκμεταλλευτεί τους παραπάνω βαθμούς ελευθερίας χωρίς να χαλάει την βαθμονόμηση.

Καταλήξαμε επομένως στην χρήση δύο μεθόδων: «3HP-INF-SIGMF» και «5HP-INF-EVT». Η πρώτη τα πηγαίνει καλά και στα δυο σύνολα δεδομένων ενώ η δεύτερη δεν εντυπωσίασε στο BCC αλλά μόνο στο BreaKHis. Παρακάτω θα δούμε επιπλέον συγκριτικά όσον αφορά την διαδικασία προετοιμασίας των δεδομένων πριν το cross validation.

6.1.2. Stain Color Normalization

Σε ό,τι αφορά την κανονικοποίηση της απόχρωσης των stains, όπως αναφέρθηκε στην ενότητα 4.1 ο χρόνος δεν μας επέτρεψε να υλοποιήσουμε την δική μας μέθοδο βασισμένη σε έρευνες της επιλογής μας. Ο κώδικας που χρησιμοποιήσαμε δούλεψε καλά με το σύνολο δεδομένων BCC αλλά στο σύνολο δεδομένων BreaKHis λόγω κακής ποιότητας των αρχικών εικόνων απέτυχε σε μεγάλο βαθμό. Αναγκαστήκαμε επομένως να περιορίσουμε την χρήση του μόνο στο πρώτο dataset.



	BCC	BCC SCN	BKH 200X	BKH 200X SCN
CV Accuracy	0,8067	0,8467	0,8772	0,7191
CV AUC	0,8472	0,9111	0,9431	0,6491

Σχήμα 6.4: Σύγκριση αποτελεσμάτων με και χωρίς stain color normalization.

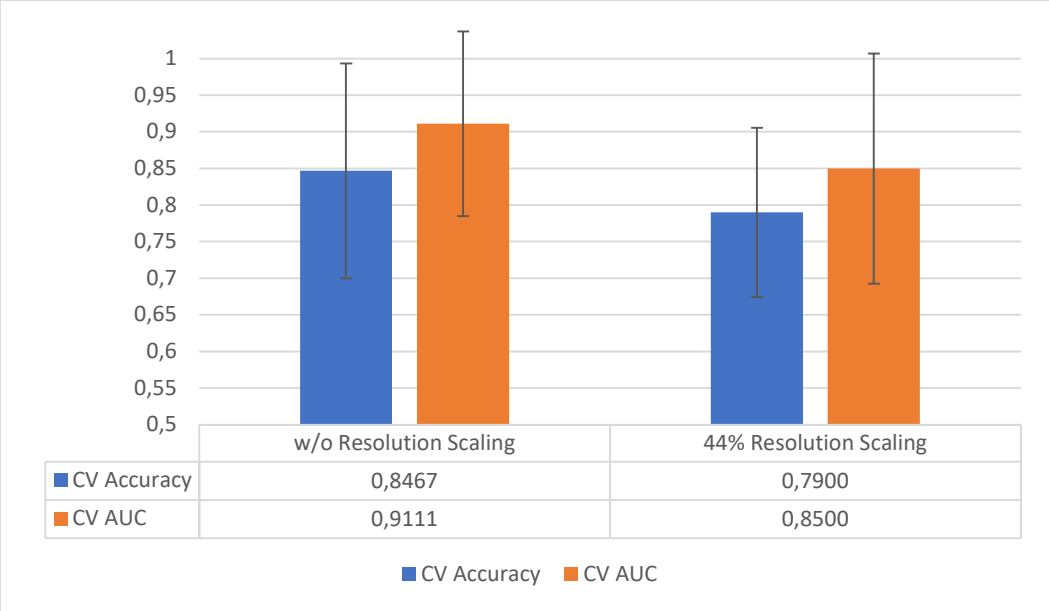
Στο παραπάνω διάγραμμα εμφανίζονται ενδεικτικά τα καλύτερα αποτελέσματα που πήραμε από την μέθοδο «3HP-INFW-SIGMF» για τα δεδομένα «BCC 5-1-0 R40 [0,1]» και από την μέθοδο «5HP-INFW-EVT» για τα δεδομένα «BKH200 33% R40 [0,1]» με και χωρίς stain color normalization. Στο BCC που η κανονικοποίηση της απόχρωσης δούλεψε σωστά είναι εμφανής η βελτίωση και στις δυο μετρικές, τόσο στον μέσο όρο όσο και στην τυπική τους απόκλιση. Από την άλλη στο BreaKHis που ο κώδικα δεν δούλεψε όπως θα έπρεπε τα αποτελέσματα ήταν καταστροφικά.

Ανάλογες βελτιώσεις παρατηρήθηκαν και σε επιπλέον πειράματα που έγιναν στο BCC. Το συμπέρασμα είναι πως το stain color normalization, άμα μπορεί να εφαρμοστεί σωστά στο σύνολο δεδομένων, είναι ικανό να βελτιώσει τα αποτελέσματα σημαντικά.

6.1.3. Resolution Scaling

Λόγω περιορισμένης χωρητικότητας της κύριας μνήμης στα υπολογιστικά μας συστήματα αναγκαστήκαμε να μειώσουμε την ανάλυση των εικόνων για το σύνολο δεδομένων BreaKHis στο 33%. Είναι σημαντικό να πάρουμε μια εκτίμηση για το αν τα αποτελέσματα μας χειροτέρεψαν και σε τι βαθμό.

Για τον σκοπό αυτό δοκιμάσαμε να κάνουμε το ίδιο στο σύνολο δεδομένων BCC το οποίο μπορούμε να τρέξουμε και στην αρχική του ανάλυση και να συγκρίνουμε τα αποτελέσματα. Ενδεικτικά θα δείξουμε τα αποτελέσματα για τα δεδομένα «BCC SCN 5-1-0 R40 [0,1]» με resolution scaling 44% και χωρίς, με χρήση της μεθόδου «3HP-INFW-SIGMF».



Σχήμα 6.5: Σύγκριση αποτελεσμάτων με και χωρίς resolution scaling.

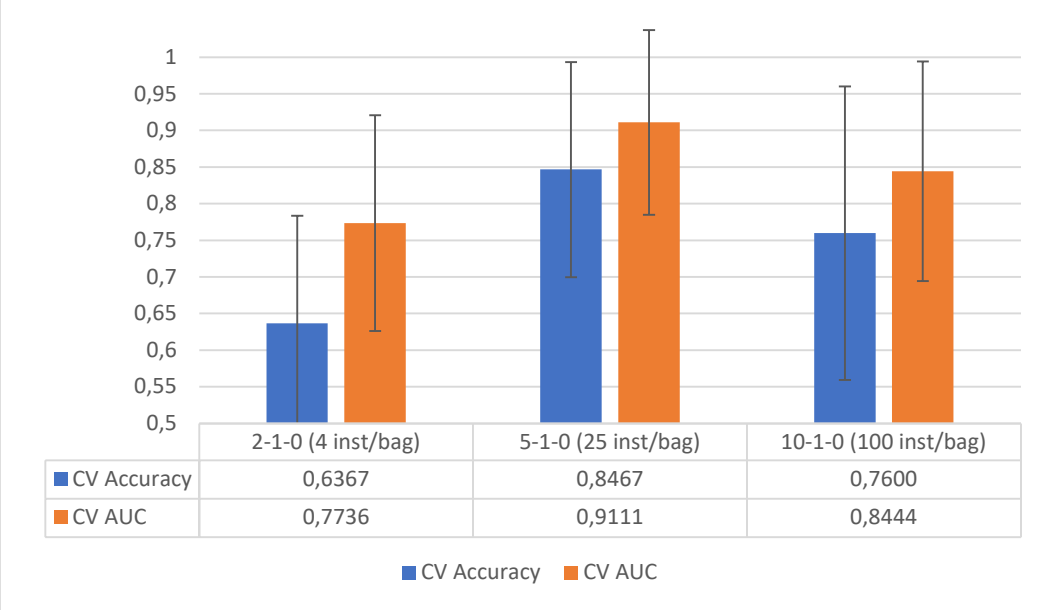
Όπως φαίνεται η επίδραση της μειωμένης ανάλυσης στα αποτελέσματα μπορεί να είναι μεγάλη. Αυτό σημαίνει πως τα αποτελέσματα για το BreaKHis που θα παρουσιάσουμε παρακάτω θα μπορούσαν να είναι (σημαντικά) καλύτερα αν είχαμε στην διάθεση μας περισσότερη μνήμη για την εφαρμογή της NN-CPD χωρίς resolution scaling.

6.1.4. Image Segmentation

Ίσως η πιο δύσκολη επιλογή στην πειραματική μας διαδικασία είναι οι υπερ-παράμετροι που θα χρησιμοποιηθούν για τον κατακερματισμό των εικόνων, αν αυτός πραγματοποιηθεί.

Πρόκειται για μια τεχνική η οποία μπορεί να δημιουργήσει τρομακτική αύξηση στο πλήθος των στιγμιότυπων και να πολλαπλασιάσει τον χρόνο εκτέλεσης των πειραμάτων. Ο σημαντικότερος παράγοντας για τον χρόνο εκτέλεσης στα πειράματα μας είναι ο αριθμός των στιγμιότυπων που υπάρχουν στο training set, καθώς ο χρόνος εκπαίδευσης των SVM είναι πολυωνυμικός ως προς αυτόν.

Εκτός από τον χρόνο εκτέλεσης όμως ο τρόπος που θα γίνει το image segmentation μπορεί να επηρεάσει δραματικά την επίδοση του αλγορίθμου.



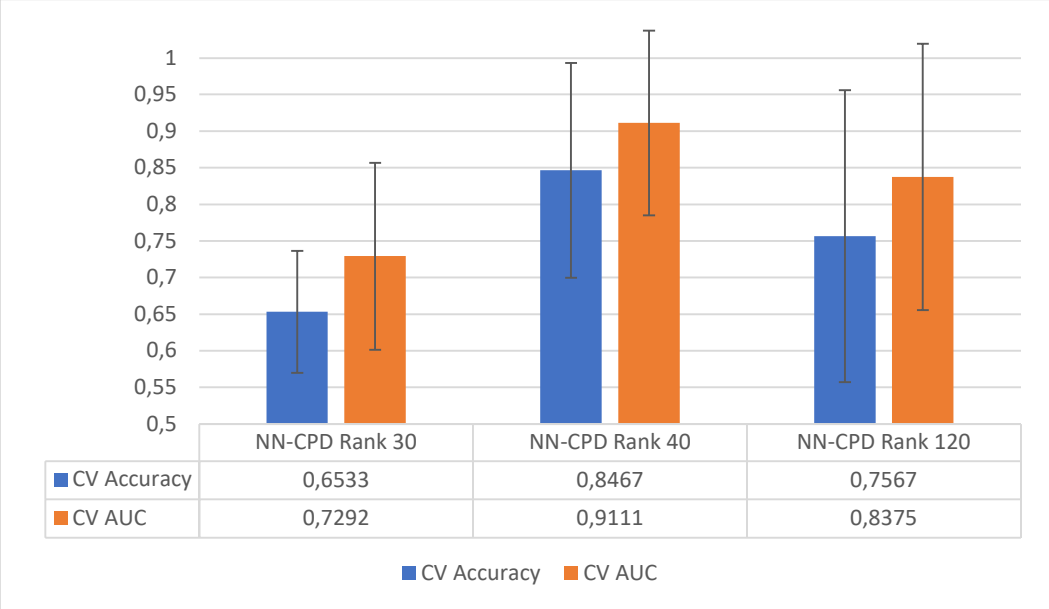
Σχήμα 6.6: Σύγκριση αποτελεσμάτων για διαφορετικά πλήθη στιγμιότυπων.

Είναι φανερό πως δεν υπάρχει γραμμική σχέση εδώ πέρα. Με πολύ μικρό πλήθος στιγμιότυπων δεν αφήνουμε περιθώρια στην MIL αρχιτεκτονική να λάμψει, όμως με έναν μεγάλο αριθμό από στιγμιότυπα αρχίζει να υπερέχει πιθανώς η άχρηστη πληροφορία στις bag-level μετρικές με τον ίδιο τρόπο που συμβαίνει και στην SIL (single instance learning) κατηγοριοποίηση στο στάδιο του feature extraction.

6.1.5. NN-CPD Rank

Η τάξη της NN-CPD αποσύνθεσης είναι πολύ σημαντική υπερ-παράμετρος η οποία όμως είναι αδύνατον να προσδιοριστεί με βέλτιστο τρόπο χωρίς την διεξαγωγή πολλών δοκιμών. Η διαδικασία της αποσύνθεσης επίσης είναι και η πιο χρονοβόρα διαδικασία της παρούσας εργασίας.

Για το σύνολο δεδομένων BreaKHis περιοριστήκαμε στην χρήση της τάξης R = 40 αφού μετά και από 33% resolution scaling η χωρητικότητα της κύριας μνήμης που είχαμε στην διάθεση μας δεν μας επέτρεπε να πάμε παραπάνω. Πρέπει να δούμε λοιπόν, χρησιμοποιώντας ξανά το σύνολο δεδομένων BCC που είναι μικρό, πως η τάξη της NN-CPD επηρεάζει τα αποτελέσματα. Οι δοκιμές έγιναν για τάξεις 30, 40 και 120 πάνω στα δεδομένα «BCC SCN 5-1-0 [1,0]» με χρήση της μεθόδου «3HP-INFW-SIGMF».



	NN-CPD Rank 30	NN-CPD Rank 40	NN-CPD Rank 120
CV Accuracy	0,6533	0,8467	0,7567
CV AUC	0,7292	0,9111	0,8375

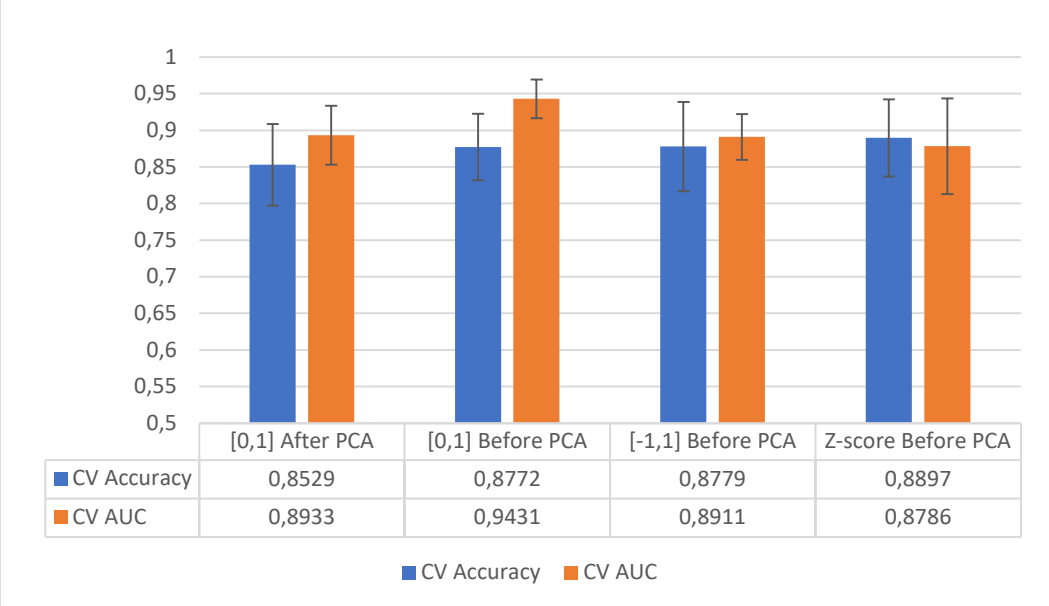
Σχήμα 6.7: Σύγκριση αποτελεσμάτων για διαφορετικής τάξης NN-CPD's.

Όπως είναι εμφανές η επιλογή της τάξης είναι δύσκολη. Πολύ μικρή τάξη μπορεί να οδηγήσει σε μεγάλες απώλειες πληροφορίας και σε πολύ κακά αποτελέσματα. Από την άλλη μεριά, η επιλογή αρκετά μεγάλης τάξης δεν μας εγγυάται καλύτερα αποτελέσματα από μια μέση λύση αφού δημιουργούνται σημαντικά περισσότερα features τα οποία δυσκολεύουν τον αλγόριθμο μας (ακόμα και με την χρήση PCA όπως φαίνεται).

6.1.6. Feature Scaling

Το τελευταίο βήμα στην προετοιμασία των δεδομένων πριν την εκτέλεση του εμφωλευμένου cross validation είναι προαιρετικά ένα feature scaling. Εδώ να πούμε πως σύμφωνα με την βιβλιογραφία [16]-[38] τα δεδομένα στην είσοδο των SVM πρέπει να είναι φραγμένα ή έστω κανονικοποιημένα σε κάποιον βαθμό. Όπως αναφέραμε στην ενότητα 4.4 οι περισσότερες υλοποιήσεις της PCA (συμπεριλαμβανομένης αυτής της MATLAB που χρησιμοποιούμε) κάνουν αυτομάτως centering στα δεδομένα. Ο μόνος σκοπός του feature scaling λοιπόν είναι να αλλάξει την τυπική απόκλιση των δεδομένων ή και να τα περιορίσει σε ένα κλειστό διάστημα.

Δοκιμάσαμε διάφορους τρόπους feature scaling στα δεδομένα «BKH200 33% R40» χρησιμοποιώντας την μέθοδο «5HP-INFW-EVT». Τα αποτελέσματα εμφανίζονται στην συνέχεια.



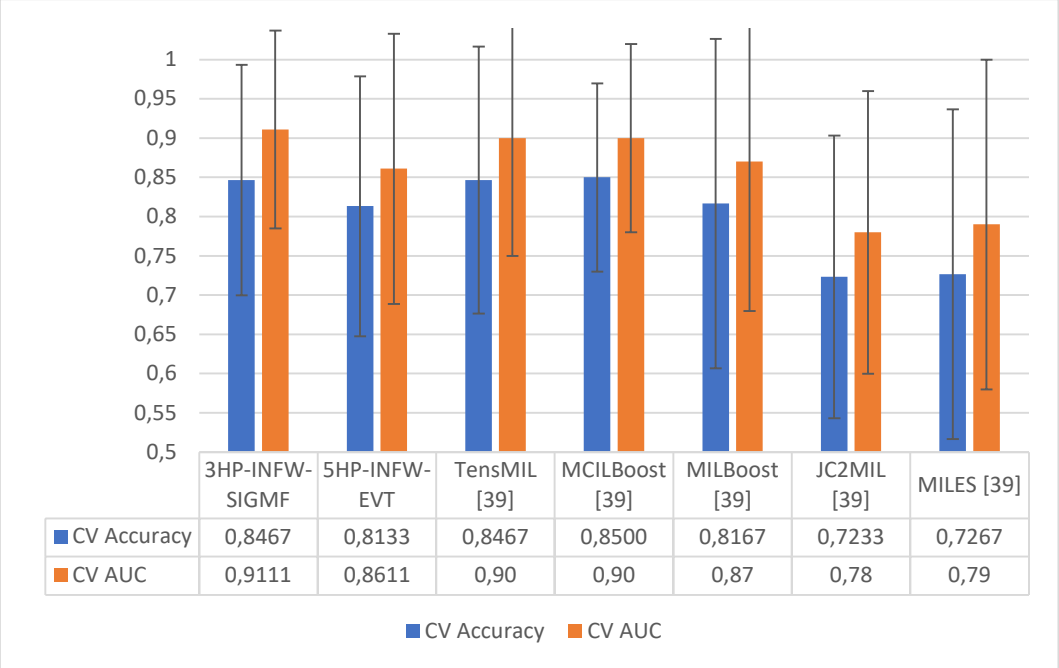
	[0,1] After PCA	[0,1] Before PCA	[-1,1] Before PCA	Z-score Before PCA
CV Accuracy	0,8529	0,8772	0,8779	0,8897
CV AUC	0,8933	0,9431	0,8911	0,8786

Σχήμα 6.8: Σύγκριση αποτελεσμάτων για διαφορετικά feature scaling.

Ενώ οι περισσότεροι τρόποι οδήγησαν σε παρόμοια αποτελέσματα ένας ξεχωρίζει. Αυτός είναι το scaling κάθε feature ξεχωριστά στο κλειστό διάστημα [0,1] και είναι το feature scaling που επιλέξαμε να κάνουμε σε όλα μας τα πειράματα.

6.2. Αποτελέσματα BCC

Το Breast Cancer Cell του UCSB αποτελεί ένα δύσκολο σύνολο δεδομένων καθώς είναι πάρα πολύ μικρό. Αυτή είναι και η αιτία για τις σχετικά μεγάλες τυπικές αποκλίσεις. Τα καλύτερα αποτελέσματα με τον αλγόριθμο μας τα πήραμε χρησιμοποιώντας τα δεδομένα «BCC SCN 5-1-0 R40 [0,1]». Οι λόγοι που μας οδήγησαν σε αυτά τα δεδομένα αναλύθηκαν στην προηγούμενη ενότητα. Στο σχήμα 6.9 και στον πίνακα 6.2 παρουσιάζουμε τα καλύτερα μας αποτελέσματα χρησιμοποιώντας 10-fold cross validation και παραθέτουμε για σύγκριση μια σειρά από αποτελέσματα από την δημοσίευση [39].



Σχήμα 6.9: Συγκριτικά αποτελέσματα 10-fold cross validation από το BCC.

BCC	Acc	AUC
3HP-INFW-SIGMF	0.8467 (0.15)	0.9111 (0.13)
5HP-INFW-EVT	0.8133 (0.17)	0.8611 (0.17)
TensMIL [39]	0.8467 (0.17)	0.90 (0.15)
MCILBoost [39]	0.8500 (0.12)	0.90 (0.12)
MILBoost [39]	0.8167 (0.21)	0.87 (0.19)
JC2MIL [39]	0.7233 (0.18)	0.78 (0.18)
MILES [39]	0.7267 (0.21)	0.79 (0.21)

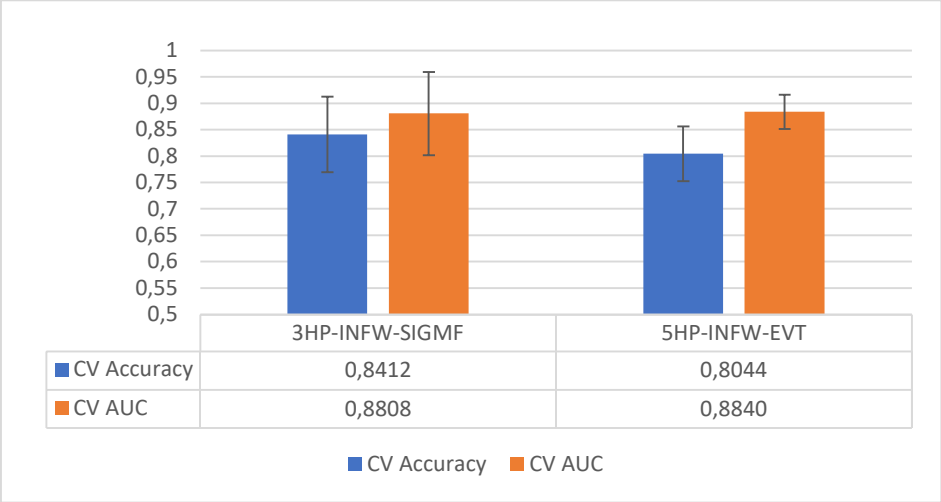
Πίνακας 6.2: Συγκριτικά αποτελέσματα 10-fold cross validation από το BCC.

Στην δημοσίευση [39] τα δεδομένα έχουν κατακερματιστεί με τον ίδιο τρόπο (10-1-0) και έχει γίνει εξαγωγή χαρακτηριστικών με χρήση ALS CPD (όχι non-negative) τάξης R = 120. Στην συνέχεια εφαρμόστηκαν 5 διαφορετικοί αλγόριθμοι πάνω στα εξαχθέντα features. Λόγω της ομοιότητας στο data pretreatment μεταξύ των δυο μεθόδους θεωρούμε πως η σύγκριση αυτή είναι ουσιαστικής σημασίας και δείχνει πως ο αλγόριθμος μας είναι ανταγωνιστικός.

6.3. Αποτελέσματα BreaKHis

Προχωρώντας στο επόμενο σύνολο δεδομένων το οποίο είναι πολύ μεγαλύτερο τα αποτελέσματα είναι σαφώς βελτιωμένα σε σύγκριση με πριν. Όπως εξηγήσαμε στην υποενότητα 1.2.2 το BreaKHis περιέχει εικόνες σε τέσσερα διαφορετικά επίπεδα μεγέθυνσης του μικροσκοπίου. Τα αποτελέσματα μας είναι ξεχωριστά για κάθε επίπεδο μεγέθυνσης ενώ όλα τα δεδομένα έχουν υποστεί 33% resolution scaling, NN-CPD τάξης 40 και feature scaling στο κλειστό διάστημα [0,1].

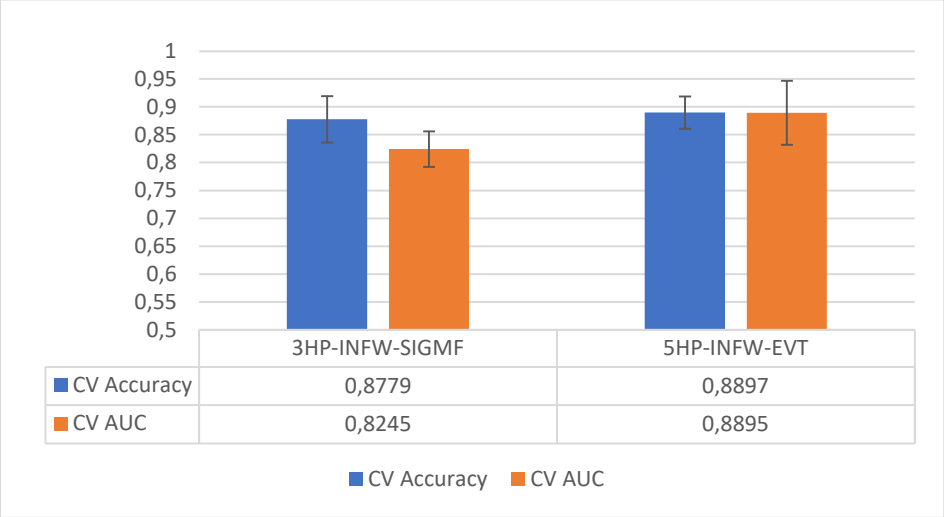
Το resolution scaling και η τάξη της αποσύνθεσης ήταν αναγκαστικές επιλογές λόγω περιορισμών από τα υπολογιστικά μας συστήματα. Δεδομένου του μεγάλου μεγέθους του dataset επίσης μειώσαμε το cross validation από 10-fold σε 5-fold θεωρώντας το αρκετό για την σωστή αξιολόγηση του αλγορίθμου μας. Ακολουθούν τα καλύτερα μας αποτελέσματα για κάθε επίπεδο μεγέθυνσης ξεχωριστά.



Σχήμα 6.10: Αποτελέσματα 5-fold cross validation από το BKH 40X.

BKH 40X	Acc	AUC
3HP-INFW-SIGMF	**0.8412 (0.072)**	0.8808 (0.079)
5HP-INFW-EVT	0.8044 (0.052)	**0.8840 (0.033)**

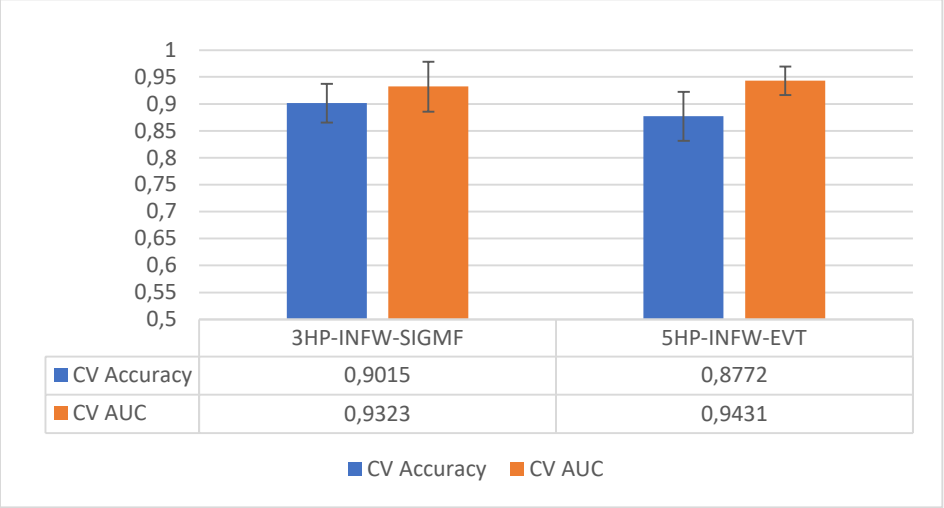
Πίνακας 6.3: Αποτελέσματα 5-fold cross validation από το BKH 40X.



Σχήμα 6.11: Αποτελέσματα 5-fold cross validation από το BKH 100X.

BKH 100X	Acc	AUC
3HP-INFW-SIGMF	0.8779 (0.042)	0.8245 (0.032)
5HP-INFW-EVT	**0.8897 (0.029)**	**0.8895 (0.058)**

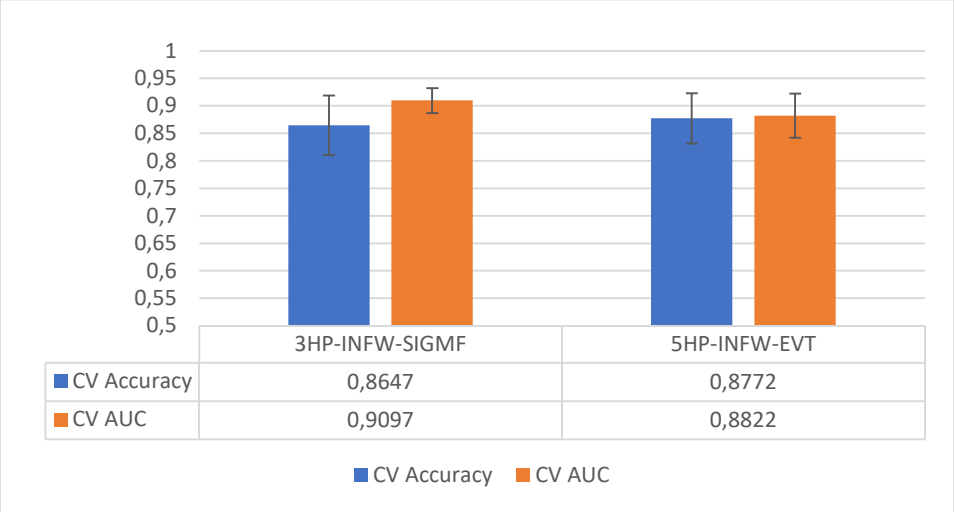
Πίνακας 6.4: Αποτελέσματα 5-fold cross validation από το BKH 100X.



Σχήμα 6.12: Αποτελέσματα 5-fold cross validation από το BKH 200X.

BKH 200X	Acc	AUC
3HP-INFW-SIGMF	**0.9015 (0.036)**	0.9323 (0.046)
5HP-INFW-EVT	0.8772 (0.046)	**0.9431 (0.027)**

Πίνακας 6.5: Αποτελέσματα 5-fold cross validation από το BKH 200X.



Σχήμα 6.13: Αποτελέσματα 5-fold cross validation από το BKH 400X.

BKH 400X	Acc	AUC
3HP-INFW-SIGMF	0.8647 (0.054)	**0.9097 (0.023)**
5HP-INFW-EVT	**0.8772 (0.046)**	0.8822 (0.040)

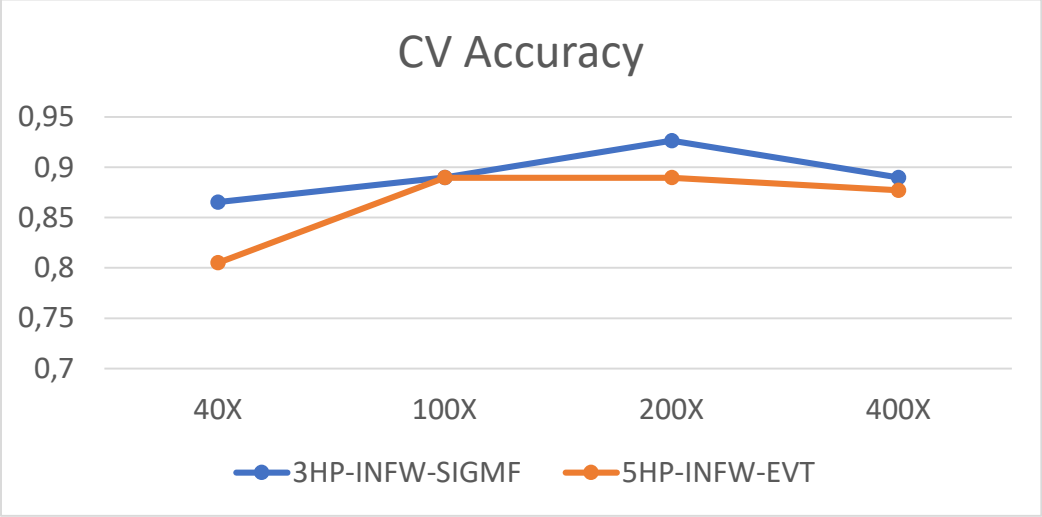
Πίνακας 6.6: Αποτελέσματα 5-fold cross validation από το BKH 400X.

Όλα τα παραπάνω αποτελέσματα προέκυψαν ύστερα από έναν αριθμό επαναλήψεων της πειραματικής διαδικασίας, και επιλέχθηκαν σύμφωνα με τον τύπο:

$$score_{balanced} = acc - std_{acc} + auc - std_{auc} \qquad (6.1)$$

Ο σκοπός αυτής της επιλογής ήταν να φανούν οι επιδόσεις του αλγορίθμου σε ένα πιο ισορροπημένο πλαίσιο. Στην σχετική βιβλιογραφία με το σύνολο δεδομένων BreaKHis πάρα πολύ σπάνια συναντάται η μετρική του AUC όμως. Θα ήταν άδικο επομένως να συγκρίνουμε τα προηγούμενα αποτελέσματα με την βιβλιογραφία αφού μπορούμε να τα βελτιώσουμε ως προς την μετρική του accuracy. Για αυτόν τον λόγο χρησιμοποιήθηκαν και οι επόμενοι τύποι αξιολόγησης ώστε να επιλέξουμε τα καλύτερα αποτελέσματα μόνο με βάση το accuracy ή το AUC.

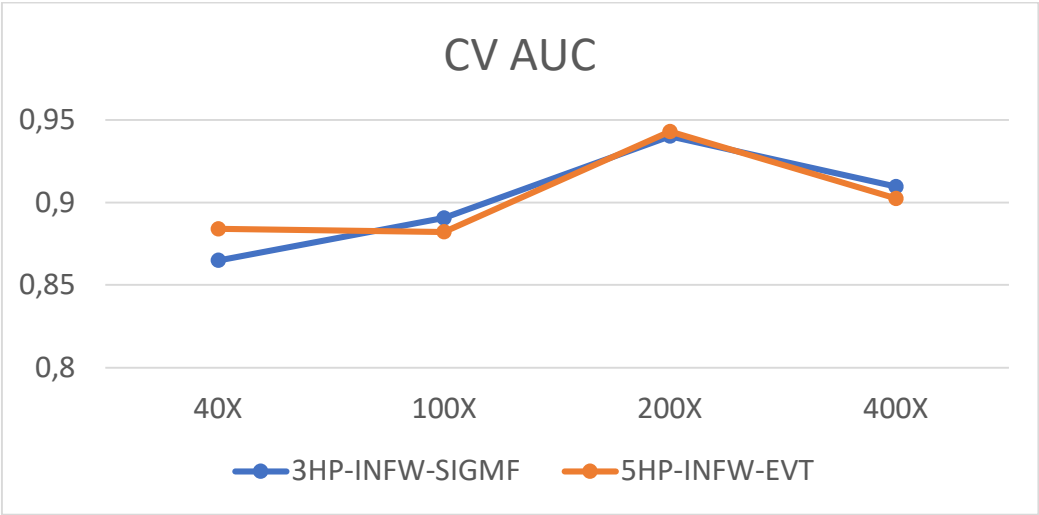
$$score_{acc} = acc - std_{acc} \quad , \quad score_{auc} = auc - std_{auc} \qquad (6.2)$$



Σχήμα 6.14: Αποτελέσματα 5-fold cross validation από το BreaKHis (accuracy).

BreaKHis	40X	100X	200X	400X
3HP-INFW-SIGMF	**0.8654 (0.053)**	**0.8897 (0.029)**	**0.9265 (0.029)**	**0.8897 (0.029)**
5HP-INFW-EVT	0.8051 (0.023)	**0.8897 (0.029)**	0.8897 (0.029)	0.8772 (0.046)

Πίνακας 6.7: Αποτελέσματα 5-fold cross validation από το BreaKHis (accuracy).



Σχήμα 6.15: Αποτελέσματα 5-fold cross validation από το BreaKHis (AUC).

BreaKHis	40X	100X	200X	400X
3HP-INFW-SIGMF	0.8650 (0.043)	0.8907 (0.068)	0.9402 (0.045)	**0.9097 (0.023)**
5HP-INFW-EVT	**0.8840 (0.033)**	**0.8822 (0.029)**	**0.9431 (0.027)**	0.9025 (0.027)

Πίνακας 6.8: Αποτελέσματα 5-fold cross validation από το BreaKHis (AUC).

BreaKHis	40X	100X	200X	400X
3HP-INFW-SIGMF	0.865 (0.053)	**0.890 (0.029)**	**0.927 (0.029)**	**0.890 (0.029)**
5HP-INFW-EVT	0.805 (0.023)	**0.890 (0.029)**	0.890 (0.029)	0.877 (0.046)
MILCNN [40]	0.869 (0.054)	0.857 (0.048)	0.859 (0.039)	0.834 (0.053)
Non-parametric [40]	**0.921** (**0.059**)	**0.891** (**0.052**)	0.872 (0.043)	0.827 (0.030)
MI-SVM Linear [40]	0.856 (0.056)	0.821 (0.059)	0.846 (0.048)	0.809 (0.049)
MI-SVM Poly [40]	0.848 (0.027)	0.825 (0.046)	0.839 (0.042)	0.813 (0.042)
MI-SVM RBF [40]	0.790 (0.021)	0.719 (0.029)	0.762 (0.019)	0.730 (0.035)
mi-SVM Linear [40]	0.795 (0.043)	0.834 (0.046)	0.836 (0.047)	0.810 (0.052)
mi-SVM Poly [40]	0.752 (0.061)	0.798 (0.048)	0.765 (0.039)	0.685 (0.051)
mi-SVM RBF [40]	0.778 (0.016)	0.754 (0.015)	0.738 (0.023)	0.729 (0.034)
Citation-kNN [40]	0.737 (0.046)	0.728 (0.054)	0.757 (0.031)	0.772 (0.036)
EM-DD [40]	0.783 (0.056)	0.806 (0.052)	0.771 (0.063)	0.787 (0.057)
DD [40]	0.705 (0.061)	0.645 (0.043)	0.683 (0.036)	0.712 (0.033)
Iterated-discrim. APR [40]	0.738 (0.038)	0.665 (0.041)	0.842 (0.049)	0.680 (0.056)

Πίνακας 6.9: Συγκριτικά αποτελέσματα από το BreaKHis (accuracy).

Στον πίνακα 6.9 παρουσιάζουμε τα αποτελέσματα μας σε σύγκριση με τα αντίστοιχα αποτελέσματα του πίνακα 2 της δημοσίευσης [40]. Σε αυτήν την δημοσίευση για κάθε εικόνα στο training set πήραν τυχαία 1000 patches ανάλυσης 64x64 ενώ από το training set πήραν 100. Από αυτά τα patches δημιούργησαν feature vectors μήκους 162 με χρήση PFTAS. Τα συγκεκριμένα αποτελέσματα είναι «patient as bag», δηλαδή όλα αυτά τα feature vectors αντιστοιχήθηκαν με τους ασθενείς και όχι με τις αρχικές εικόνες.

Συνοψίζοντας τον πίνακα 6.9, ενώ στο επίπεδο μεγέθυνσης 40X χάνουμε, ο αλγόριθμος μας είναι πολύ ανταγωνιστικός στο 100X και στα επίπεδα 200X και 400X κερδίζει.

ΚΕΦΑΛΑΙΟ 7: ΣΥΖΗΤΗΣΗ

7.1. Συμπεράσματα

Ο στόχος αυτής της διπλωματικής εργασίας ήταν η εφαρμογή ενός MIL κατηγοριοποιητή που κάνει χρήση πολλαπλών one-class SVM στο πρόβλημα της κατηγοριοποίησης ιστοπαθολογικών εικόνων νεοπλασιών του μαστού. Στα πλαίσια αυτής της εργασίας περιοριστήκαμε στην δυαδική κατηγοριοποίηση (καλοήθειες – κακοήθειες).

Για την σύντηξη των instance-level decision values των one-class SVM σε bag-level decision values μελετήθηκαν οι τρείς διαφορετικές μετρικές που περιγράφονται στην ενότητα 5.2. Από αυτές σύμφωνα με τα πειράματα η καλύτερη και ως προς την ακρίβεια αλλά και ως προς το AUC φαίνεται να είναι αυτή με τα informative windows, με τον αριθμητικό μέσο να ακολουθεί σχετικά κοντά και την εντροπία να είναι σαφώς χειρότερη.

Εκτός από την σύντηξη των instance-level decision values έπρεπε να μελετηθεί και η βαθμονόμηση τους καθώς αυτά προέρχονται από διαφορετικά one-class SVM. Εδώ εφαρμόστηκαν δυο διαφορετικές μέθοδοι τις οποίες εξηγήσαμε στην ενότητα 4.5. Η μέθοδος της Weibull φαίνεται να συμπαθεί την περίπτωση που τα SVM εκπαιδεύονται με διαφορετικές υπερ-παραμέτρους (πράγμα που εξηγείται από το γεγονός ότι σε αυτήν την μέθοδο εμπεριέχεται fitting) ενώ η μέθοδος της απλής σιγμοειδούς συμπαθεί την περίπτωση που τα SVM εκπαιδεύονται με τις ίδιες υπερ-παραμέτρους. Πάραυτα, από τα πειράματα που τρέξαμε δεν φάνηκε να υπερέχει κάποια από τις δυο μεθόδους στην μεταξύ τους σύγκριση (στο σύνολο δεδομένων BreaKHis που είναι και πιο μεγάλο-αντιπροσωπευτικό). Αυτό οφείλεται στο γεγονός πως ενώ η μέθοδος της Weibull μας επιτρέπει να εκπαιδεύσουμε καλύτερα τα SVM με διαφορετικές υπερ-παραμέτρους πάνω στο training set, η επιρροή του σφάλματος της γενίκευσης μεγαλώνει καθώς ο αλγόριθμος μας είναι ευάλωτος σε αυτό σε δυο περισσότερα σημεία (gamma και nu του δεύτερου one-class SVM).

Άλλο ένα κομμάτι του κατηγοριοποιητή μας είναι η PCA ανάλυση του συνόλου δεδομένων μετά από την οποία κάποια ή και κανένα από τα principal components με το μικρότερο variance πετάγονται. Το variance retained (κεφάλαιο 4.4) είναι μια από τις υπερ-παραμέτρους που αφήνουμε τον Bayesian optimizer να βελτιστοποιήσει για μας (μαζί με τις υπερ-παραμέτρους εκπαίδευσης των SVM). Από τα πειράματα διαπιστώσαμε πως αυτό δεν μένει ποτέ στο 1, δηλαδή ο Bayesian optimizer βρίσκει συνεχώς πως είναι καλύτερα να πετάξει έναν αριθμό από features.

Κατά την διάρκεια εκπόνησης της εργασίας διαπιστώθηκε πως ο κατηγοριοποιητής αυτός είναι πάρα πολύ ευαίσθητος σε αλλαγές στις υπερ-παραμέτρους εκπαίδευσης των SVM και του variance retained της PCA. Γι’ αυτόν

τον λόγο τα αποτελέσματα που παρουσιάσαμε στο κεφάλαιο 6 είναι τα καλύτερα από μια σειρά 50-100 δοκιμών. Η χρήση πολλών «εσωτερικών» folds στο nested cross validation μας (π.χ. 5-10) φαίνεται να μας δίνει μια παραπάνω σταθερότητα σε αντίθεση με το optimization με χρήση 2 μόνο folds. Το εύρος τιμών για την υπερ-παράμετρο gamma του RBF kernel που χρησιμοποιήσαμε σε όλα μας τα πειράματα ήταν το [1,10] αφού φάνηκε να αποδίδει βέλτιστα, ενώ τα nu και variance retained είχαν το πλήρες εύρος [0,1].

Συνεχίζοντας με την προετοιμασία των δεδομένων πριν την είσοδο τους στον κατηγοριοποιητή μας, ασχοληθήκαμε με το stain color normalization, τον κατακερματισμό των εικόνων σε patches και την εξαγωγή χαρακτηριστικών με χρήση non-negative CPD.

Το stain color normalization διαπιστώσαμε πως βοηθάει και πρέπει να εφαρμόζεται. Δυστυχώς ο κώδικας που είχαμε στα χέρια μας δεν δούλεψε για το BreaKHis αλλά μόνο για το BCC.

Τον κατακερματισμό εικόνας τον χρησιμοποιήσαμε μόνο στο BCC όπου κάναμε image-level κατηγοριοποίηση. Στο BreaKHis κάναμε patient-level κατηγοριοποίηση και υπάρχουν ήδη 24 εικόνες κατά μέσο όρο ανά ασθενή οι οποίες μπορούν να θεωρηθούν σαν patches. Στο BCC που εφαρμόστηκε δεν φάνηκε να υπάρχει κάποιο βέλτιστο «κόψιμο» που να μπορεί να προβλεφθεί.

Η non-negative CPD μπορούμε να πούμε, κρίνοντας από τα αποτελέσματα, πως μας έδωσε αρκετά καλά features. Ξανά εδώ δεν φάνηκε να υπάρχει κάποια βέλτιστη τάξη που να μπορεί να προβλεφθεί. Ένα πράγμα που διαπιστώσαμε πως μπορεί να επηρεάσει την ποιότητα των features της CPD είναι η μειωμένη ανάλυση των αρχικών εικόνων, πράγμα που αναγκαστήκαμε να εφαρμόσουμε με το BreaKHis αφού δεν είχαμε αρκετή διαθέσιμη μνήμη για την αποσύνθεση.

Το τελευταίο πράγμα με το οποίο πειραματιστήκαμε ήταν το scaling των features πριν την είσοδο τους στον κατηγοριοποιητή μας. Καθώς αυτός περιέχει SVM και πριν από αυτά μια PCA η θεωρία μας λέει ότι πρέπει να εφαρμοστεί κάποιου είδους scaling κοντά στην μονάδα. Όπως διαπιστώσαμε μέσα από πειράματα η καλύτερη μέθοδος για να γίνει αυτό (για τα συγκεκριμένα δεδομένα) είναι το mapping του κάθε feature ξεχωριστά στο κλειστό διάστημα [0,1].

Τέλος, τα αποτελέσματα που πήραμε και στα δυο σύνολα δεδομένων μπορούμε να πούμε πως ανταγωνίζονται σε αρκετά ικανοποιητικό βαθμό αντίστοιχα αποτελέσματα από την σχετική βιβλιογραφία, ξεπερνώντας μάλιστα και ένα συνελικτικό δίκτυο βαθιάς εκμάθησης όπως φαίνεται στον πίνακα 6.9.

7.2. Μελλοντική Έρευνα

Όπως αναφέραμε σε αρκετά σημεία της εργασίας υπάρχουν πράγματα που δεν μπορέσαμε να δοκιμάσουμε λόγω της περιορισμένης μνήμης στα υπολογιστικά μας συστήματα αλλά και λόγω χρόνου.

Μιλώντας για το σύνολο δεδομένων BreaKHis υπάρχουν προοπτικές για ακόμα καλύτερα αποτελέσματα με την χρήση της μεθόδου που αναπτύξαμε. Αυτά θα μπορούσαν να επιτευχθούν αρχικά με την εφαρμογή του stain color normalization έπειτα από την δημιουργία ενός καλύτερου κώδικα για αυτό, καθώς επίσης και με την χρήση περισσότερης μνήμης RAM για την NN-CPD ώστε αυτή να εφαρμοστεί στις εικόνες χωρίς μειωμένη ανάλυση.

Σχετικά με το BreaKHis θα μπορούσαν επίσης να δοκιμαστούν μερικές διαφορετικές τάξεις για την NN-CPD (στην εργασία δοκιμάσαμε μόνο την τάξη 40 που ήταν και η μεγαλύτερη που μπορούσαμε) όπως επίσης και η τεχνική του κατακερματισμού εικόνας για την δημιουργία περισσότερων στιγμιότυπων.

Ένα σημείο ενδιαφέροντος για μελλοντική έρευνα πάνω στην βελτίωση της μεθόδου μας αφορά την δημιουργία και την επιλογή των patches. Εδώ θα μπορούσε να δοκιμαστεί το overlap που περιγράψαμε στην ενότητα 4.2, σε συνδυασμό με το πέταγμα ορισμένων patches. Μια δυσκολία σε αυτό το κομμάτι είναι το ελλιπές ιατρικό υπόβαθρο σχετικά με το ποιο θεωρείται καλό και ποιο κακό patch. Μία άλλη πιθανή μέθοδος εδώ θα μπορούσε να είναι η τυχαία επιλογή ενός προκαθορισμένου αριθμού από patches συγκεκριμένου μεγέθους (64x64 ενδεχομένως που δουλεύει καλά στα CNN δίκτυα) συνοδευόμενη πάλι με κάποιο κριτήριο απόρριψης ή και όχι.

Ένα άλλο σημείο έρευνας θα μπορούσε να είναι η ταυτόχρονη χρήση πολλαπλών feature matrices που προκύπτουν από τις ίδιες εικόνες αλλά με διαφορετικά patches και εκτελέσεις της NN-CPD.

Όσον αφορά το σύνολο δεδομένων BreaKHis, άλλη μια ιδέα θα ήταν να γίνει χρήση όλων των επιπέδων μεγέθυνσης ταυτόχρονα για την κατηγοριοποίηση του κάθε ασθενή. Κάτι τέτοιο δεν συναντήθηκε στην βιβλιογραφία και επομένως τα συγκριτικά πειράματα θα ήταν μια δύσκολη υπόθεση.

Ένα ακόμα πράγμα που δεν δοκιμάσαμε στα πλαίσια αυτής της εργασίας ήταν η multi-class κατηγοριοποίηση. Το BreaKHis διαθέτει συνολικά 8 κλάσεις και θα μπορούσε να χρησιμοποιηθεί για μια τέτοια έρευνα. Θα ήταν πολύ ενδιαφέρουσα η μελέτη των τεχνικών βαθμονόμησης και σύντηξης που αναπτύξαμε σε περιπτώσεις με περισσότερα από δυο one-class SVM.

Βιβλιογραφία

[1] Stewart, B. W. & Wild, C. World cancer report 2014. international agency for research on cancer. *World Health Organization* 505 (2014)

[2] Chan, J. K. (2014). The wonderful colors of the hematoxylin–eosin stain in diagnostic surgical pathology. *International journal of surgical pathology*, *22*(1), 12-32.

[3] Gelasca, E. D., Byun, J., Obara, B., & Manjunath, B. S. (2008, October). Evaluation and benchmark for biological image segmentation. In *2008 15th IEEE International Conference on Image Processing* (pp. 1816-1819). IEEE.

[4] Spanhol, F. A., Oliveira, L. S., Petitjean, C., & Heutte, L. (2015). A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, *63*(7), 1455-1462.

[5] Vahadane, A., Peng, T., Sethi, A., Albarqouni, S., Wang, L., Baust, M., ... & Navab, N. (2016). Structure-preserving color normalization and sparse stain separation for histological images. *IEEE transactions on medical imaging*, *35*(8), 1962-1971.

[6] Hamilton, N. A., Pantelic, R. S., Hanson, K., & Teasdale, R. D. (2007). Fast automated cell phenotype image classification. *BMC bioinformatics*, *8*(1), 110.

[7] Coelho, L. P., Ahmed, A., Arnold, A., Kangas, J., Sheikh, A. S., Xing, E. P., ... & Murphy, R. F. (2010). Structured literature image finder: extracting information from text and images in biomedical literature. In *Linking Literature, Information, and Knowledge for Biology* (pp. 23-32). Springer, Berlin, Heidelberg.

[8] Lowe, D. G. (1999, September). Object recognition from local scale-invariant features. In *iccv* (Vol. 99, No. 2, pp. 1150-1157).

[9] Haralick, R. M., Shanmugam, K., & Dinstein, I. H. (1973). Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6), 610-621.

[10] Dalal, N., & Triggs, B. (2005, June). Histograms of oriented gradients for human detection.

[11] Hitchcock, F. L. (1927). The expression of a tensor or a polyadic as a sum of products. *Journal of Mathematics and Physics*, *6*(1-4), 164-189.

[12] Amores, J. (2013). Multiple instance classification: Review, taxonomy and comparative study. *Artificial intelligence*, *201*, 81-105.

[13] Schölkopf, B., Williamson, R. C., Smola, A. J., Shawe-Taylor, J., & Platt, J. C. (2000). Support vector method for novelty detection. In *Advances in neural information processing systems* (pp. 582-588).

[14] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, *24*(6), 417.

[15] Mockus, J. (2012). *Bayesian approach to global optimization: theory and applications* (Vol. 37). Springer Science & Business Media.

[16] Support Vector Machines — scikit-learn 0.21.3 documentation. Accessed in: 2019. Available: https://scikit-learn.org/stable/modules/svm.html

[17] Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[18] Vapnik, V. (1963). Pattern recognition using generalized portrait method. *Automation and remote control*, *24*, 774-780.

[19] Vapnik, V., & Chervonenkis, A. (1964). A note on class of perceptron. *Automation and Remote Control*, *24*.

[20] Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992, July). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory* (pp. 144-152). ACM.

[21] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.

[22] Vapnik, V. N. (1995). The nature of statistical learning. *Theory*.

[23] Chang, C. C., & Lin, C. J. (2001). Training v-support vector classifiers: theory and algorithms. *Neural computation*, *13*(9), 2119-2147.

[24] Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, *10*(3), 61-74.

[25] Lin, H. T., Lin, C. J., & Weng, R. C. (2007). A note on Platt's probabilistic outputs for support vector machines. *Machine learning*, *68*(3), 267-276.

[26] Clifton, L., Clifton, D. A., Zhang, Y., Watkinson, P., Tarassenko, L., & Yin, H. (2014). Probabilistic novelty detection with support vector machines. *IEEE Transactions on Reliability*, *63*(2), 455-467.

[27] Hsu, C. W., & Lin, C. J. (2002). A comparison of methods for multiclass support vector machines. *IEEE transactions on Neural Networks*, *13*(2), 415-425.

[28] Stain Normalization | Kaggle. Accessed in 2019. Available: https://www.kaggle.com/robotdreams/stain-normalization

[29] Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788.

[30] Bro, R., & Kiers, H. A. (2003). A new efficient method for determining the number of components in PARAFAC models. *Journal of Chemometrics: A Journal of the Chemometrics Society*, *17*(5), 274-286.

[31] Jean Kossaifi, Yannis Panagakis, Anima Anandkumar and Maja Pantic, TensorLy: Tensor Learning in Python, https://arxiv.org/abs/1610.09555.

[32] Shashua, A., & Hazan, T. (2005, August). Non-negative tensor factorization with applications to statistics and computer vision. In *Proceedings of the 22nd international conference on Machine learning* (pp. 792-799). ACM.

[33] Jain, L. P., Scheirer, W. J., & Boult, T. E. (2014, September). Multi-class open set recognition using probability of inclusion. In *European Conference on Computer Vision* (pp. 393-409). Springer, Cham.

[34] Scheirer, W., Rocha, A., Micheals, R., & Boult, T. (2010, September). Robust fusion: Extreme value theory for recognition score normalization. In *European Conference on Computer Vision* (pp. 481-495). Springer, Berlin, Heidelberg.

[35] Scheirer, W. J., Rocha, A., Micheals, R. J., & Boult, T. E. (2011). Meta-recognition: The theory and practice of recognition score analysis. *IEEE transactions on pattern analysis and machine intelligence*, *33*(8), 1689-1695.

[36] Scheirer, W. J., Kumar, N., Belhumeur, P. N., & Boult, T. E. (2012, June). Multi-attribute spaces: Calibration for attribute fusion and similarity search. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2933-2940). IEEE.

[37] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, *27*(8), 861-874.

[38] Using Support Vector Machines Effectively | Neeraj Kumar, Accessed in: 2019. Available: https://neerajkumar.org/writings/svm/

[39] Papastergiou, T., Zacharaki, E. I., & Megalooikonomou, V. (2018). Tensor Decomposition for Multiple-Instance Classification of High-Order Medical Data. *Complexity*, *2018*.

[40] Sudharshan, P. J., Petitjean, C., Spanhol, F., Oliveira, L. E., Heutte, L., & Honeine, P. (2019). Multiple instance learning for histopathological breast cancer image classification. *Expert Systems with Applications*, *117*, 103-111.