ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΔΠΜΣ: Εφαρμοσμένες Μαθηματικές Επιστήμες

Μέθοδοι Διανυσμάτων Υποστήριξης για την Ανάλυση Επιβίωσης

Διπλωματική Εργασία

Μούντζια Άννα

ΑΜ:09415014

Επιβλέπων Καθηγητής: Χ. Κουκουβίνος

Καθηγητής ΕΜΠ

Αθήνα, 2019

NATIONAL TECHNICAL UNIVERSITY OF ATHENS

SCHOOL OF APPLIED MATHEMATICAL AND PHYSICAL SCIENCES

MSc: Applied Mathematical Sciences

Support vector methods for survival analysis

Master's Thesis

Mountzia Anna

RN: 09415014

Supervisor: C. Koukouvinos

Professor NTUA

Athens, 2019

Περίληψη

Η Μηχανική Μάθηση έχει κάνει πλέον πολλά βήματα στη λύση στατιστικών προβλημάτων, γεγονός που δε θα μπορούσε να αφήσει ανεπηρέαστη και την Ανάλυση Επιβίωσης. Η διπλωματική αυτή εργασία ασχολείται με την αντιμετώπιση του προβλήματος της εκτίμησης του χρόνου επιβίωσης, μέσω των Μηχανών Διανυσμάτων Υποστήριξης (*Support Vector Machines*).

Το κύριο πρόβλημα που αντιμετωπίζουμε στην δημιουργία μοντέλων που εκτιμούν την επιβίωση είναι η παρουσία των αποκομμένων δεδομένων. Στην παρούσα εργασία παρουσιάζονται μοντέλα βασισμένα στην Ανάλυση Παλινδρόμησης αλλά και της Ταξινόμησης μέσω των Μηχανών Διανυσμάτων Υποστήριξης τα οποία λαμβάνουν υπόψη τους αυτή την ιδιαιτερότητα των δεδομένων.

Στο 1ο κεφάλαιο γίνεται μια εισαγωγή στις βασικές έννοιες της Ανάλυσης Επιβίωσης και την περιγραφή των αποκομμένων δεδομένων. Στο 2ο γίνεται μια εισαγωγή στις Μηχανές Διανυσμάτων Υποστήριξης και στις βασικές μεθόδους μηχανικής μάθησης. Στο 3ο παρουσιάζονται μοντέλα Ανάλυσης Παλινδρόμησης μέσω Μηχανών Διανυσμάτων Υποστήριξης για την Ανάλυση Επιβίωσης. Στο 4ο γίνεται ανάλυση μοντέλων Ταξινόμησης μέσω Μηχανών Διανυσμάτων Υποστήριξης. Στο 5ο γίνεται μια προσπάθεια εφαρμογής των μοντέλων αυτών μέσω της γλώσσας προγραμματισμού *R* σε πραγματικά δεδομένα. Τέλος, στο 6ο κεφάλαιο παρουσιάζονται κάποια γενικά συμπεράσματα.

Λέξεις Κλειδιά

Ανάλυση Επιβίωσης, Μηχανές Διανυσμάτων Υποστήριξης, Χρόνος Επιβίωσης, Αποκομμένα Δεδομένα, Ανάλυση Παλινδρόμησης με Μηχανές Διανυσμάτων Υποστήριξης, Ταξινόμηση με Μηχανές Διανυσμάτων Υποστήριξης

Abstract

Machine Learning has made a big progress in solving statistical problems, which could not leave the Survival Analysis unaffected. This diploma thesis deals with the problem of estimating survival time through Support Vector Machines.

The main problem we deal with in creating models that estimate survival time is the presence of censored data. In this paper, we present models based on Regression Analysis and Ranking through Support Vector Machines that take into account this particularity of the data.

Finally, an attempt is made to apply these models through the programming language R to real data.

The 1st chapter introduces the basic tools we use in Survival Analysis and it describes censored data. The 2nd makes an introduction to Support Vector Machines and Machine Learning Methods. In the 3rd, Regression Analysis Models are presented through Support Vector Machines for Survival Analysis. In the 4th we describe Ranking Support Vector Machines models. In the 5th, an attempt is made to apply these models through the programming language R to real data. Finally, chapter 6 presents some general conclusions.

Key Words

Survival Analysis, Support Vector Machines, Survival Time, Censored Data, Support Vector Regression, Ranking Support Vector Machines

Ευχαριστίες

Θα ήθελα να εκφράσω θερμές ευχαριστίες στον επιβλέπων καθηγητή κ. Κουκουβίνο Χρήστο για την ευκαιρία που μου έδωσε να καταπιαστώ με ένα τόσο ενδιαφέρον και σύγχρονο θέμα για την διπλωματική μου εργασία, αλλά και για τις κατευθύνσεις που μου έδωσε πάνω σε αυτό. Θα ήθελα να εκφράσω πολλές ευχαριστίες, επίσης, στην υποψήφια διδάκτορα Λάππα Αγγελική για την πολύτιμη βοήθεια της με καίριες υποδείξεις, αλλά και την εξαιρετική συνεργασία της, κατά την εκπόνηση της εργασίας αυτής.

Θα ήθελα να ευχαριστήσω θερμά, επίσης, την οικογένεια μου, τους φίλους και τις φίλες για την στήριξή τους καθ' όλη την διάρκεια των σπουδών μου.

Περιεχόμενα

Κεφάλαιο 1- Εισαγωγή στην Ανάλυση Επιβίωσης

1.1 Γενικά περί Ανάλυσης Επιβίωσης.

Η ανάλυση επιβίωσης αναφέρεται στην μελέτη του χρονικού διαστήματος μέχρι την εμφάνιση ενός γεγονότος που ενδιαφέρει τον ερευνητή και έχει εφαρμογές σε πολλά επιστημονικά πεδία. Συνήθως αναφερόμαστε σε ανεπιθύμητα γεγονότα όπως είναι ο χρόνος θανάτου ενός ασθενή, η βλάβη ενός μηχανήματος, τα χρήματα που καταβάλλονται από μία ασφάλεια ζωής. Για την ανάλυση αυτών των δεδομένων έχει αναπτυχθεί εκτεταμένη στατιστική θεωρία, η οποία όταν αναφέρεται σε εφαρμογές της επιστήμης ή της οικονομίας ονομάζεται ανάλυση αξιοπιστίας (*reliability analysis*), και όταν αναφέρεται σε βιοϊατρικά δεδομένα ονομάζεται ανάλυση επιβίωσης (*survival analysis*).

Αυτό που μας ενδιαφέρει ιδιαίτερα είναι να μελετήσουμε το χρόνο επιβίωσης (*survival time*) μέχρι να συμβεί το γεγονός. Ο χρόνος επιβίωσης είναι μια θετική μεταβλητή η οποία μετράει το χρόνο κατά τον οποίο ξεκινήσαμε να παρακολουθούμε ένα άτομο κι αυτό συμβαίνει συνήθως με την έναρξη κάποιων γεγονότων όπως η πρώτη διάγνωση μίας ασθένειας ή η έναρξη θεραπείας σε μια κλινική μελέτη. Στις επιδημιολογικές έρευνες, το σημείο που ξεκινάμε την παρακολούθηση είναι η έναρξη της έκθεσης του ατόμου σε ένα παράγοντα κινδύνου. Το σημείο που σταματάμε την παρατήρηση είναι όταν συμβεί το γεγονός όπως πχ η αποβίωση ενός ασθενούς, η ίαση, η μετάσταση κτλ.

Ένα σημαντικό χαρακτηριστικό το οποίο πρέπει να λαμβάνουμε υπόψη μας στα δεδομένα διάρκειας ζωής είναι ότι συχνά υφίστανται κάποιου είδους αποκοπή και τα ονομάζουμε αποκομμένα (*censored*). Αυτό συνήθως συμβαίνει επειδή τα άτομα μπορεί να εισέρχονται στη μελέτη σε διαφορετικούς χρόνους, με συνέπεια ο χρόνος παρακολούθησης μερικών ατόμων να μην είναι επαρκής ώστε να καταγραφεί ο χρόνος μέχρι την πραγματοποίηση του υπό μελέτη γεγονότος.

1.2 Αποκοπή Δεδομένων

Όπως είπαμε και παραπάνω τα δεδομένα διάρκειας ζωής μπορεί να περιέχουν αποκομμένες (*censored*) παρατηρήσεις. Το γεγονός αυτό συμβαίνει λόγω του περιορισμένου χρονικού διαστήματος παρατήρησης ή επειδή πολλές φορές χάνονται τα ίχνη των παρατηρούμενων ασθενών από άλλα μη ενδιαφέροντα γεγονότα (*Klein και Moeschberger*, 2005).

Για ένα πρόβλημα επιβίωσης, ο χρόνος (T) μέχρι το γεγονός ενδιαφέροντος είναι γνωστός με ακρίβεια, μόνο για εκείνες τις περιπτώσεις που συνέβη το γεγονός κατά τη διάρκεια της περιόδου μελέτης. Στις υπόλοιπες περιπτώσεις στις οποίες μπορεί να χαθεί το ίχνος τους κατά τη διάρκεια του χρόνου παρατήρησης ή ο χρόνος του γεγονότος να είναι μεγαλύτερος από τον χρόνο παρατήρησης, μπορούμε μόνο να έχουμε αποκοπή του χρόνου (C), όπου ο χρόνος αυτός μπορεί να είναι ο χρόνος που σταμάτησε η έρευνα, που βγήκε απ' την έρευνα ή που τελείωσε η παρατήρηση. Γενικά, στην Ανάλυση Επιβίωσης μπορούμε να παρατηρήσουμε είτε το χρόνο επιβίωσης (T_i) είτε τον χρόνο αποκοπής (C_i), αλλά όχι και τα δύο, για οποιαδήποτε δεδομένη περίπτωση i. Ενδιαφέρον παρουσιάζει το είδος της αποκοπής των παρατηρήσεων καθώς και ο τρόπος μελέτης τους. Έχουμε, λοιπόν, τις εξής κατηγορίες:

Δεξιά αποκοπή

Ονομάζεται δεξιά αποκοπή, επειδή το γεγονός που δεν καταγράφηκε βρίσκεται στα δεξιά του χρόνου αποκοπής, δηλαδή γνωρίζουμε ότι το γεγονός δεν έχει συμβεί στο τέλος της έρευνας.

***α)* Αποκοπή τύπου I**

Αυτού του τύπου δεξιά αποκοπή χρησιμοποιείται κυρίως όταν χρειαζόμαστε αποτελέσματα άμεσα γιατί, είτε λόγω χρόνου είτε λόγω κόστους ο ερευνητής σταματάει την έρευνα και βλέπει τα αποτελέσματα προτού συμβεί σε όλους το γεγονός που παρατηρούμε. Για παράδειγμα, στη μελέτη μιας ασθένειας δεν μπορούμε να περιμένουμε μέχρι όλα τα άτομα να αποβιώσουν. Η αποκοπή αυτή γίνεται ως εξής:

Έστω t_c ο προκαθορισμένος χρόνος που θα μπορούσε να τελειώσει η μελέτη. Εμείς εδώ αντί να παρατηρούμε τους χρόνους επιβίωσης T_1, T_2,…,T_n, παρατηρούμε τους χρόνους: Y_1, Y_2 , …, Y_n όπου:

$$Y_i = \begin{matrix} T_i \ , T_i \leq t_c \\ T_c \ , \ T_i > t_c \end{matrix} \qquad (1.2.1)$$

***β)* Αποκοπή τύπου II**

Πειράματα που περιέχουν αυτού του είδους την αποκοπή συναντιούνται συχνά σε τεχνικά συστήματα. Σε αυτή την περίπτωση όλα τα παρατηρούμενα αντικείμενα έχουν μπει στην έρευνα την ίδια ακριβώς στιγμή και το τεστ σταματά όταν σε r από αυτά (n συνολικά) συμβεί το γεγονός που μελετάμε. Αυτού του είδους η αποκοπή χρησιμοποιείται για την εξοικονόμηση χρόνου και χρήματος γιατί μπορεί να χρειαστεί πολύς χρόνος προκειμένου σε όλα τα αντικείμενα να συμβεί το γεγονός που μελετάμε. Η αποκοπή αυτή γίνεται ως εξής:

Έστω $r < n$ προκαθορισμένο. Παρατηρούμε μέχρι το r γεγονός. Εμείς εδώ, αντί να παρατηρούμε τους χρόνους επιβίωσης T_1, T_2 , …, T_n, παρατηρούμε τους χρόνους:

$$\begin{aligned} Y_{(1)} &= T_{(1)} \\ Y_{(2)} &= T_{(2)} \\ &\dots \\ Y_{(r)} &= T_{(r)} \\ Y_{(r+1)} &= T_{(r)} \\ &\dots \\ Y_{(n)} &= T_{(r)} \end{aligned} \qquad (1.2.2)$$

όπου $T_{(1)}, T_{(2)}$, …, $T_{(n)}$ να είναι το διατεταγμένο δείγμα των T_1, T_2, …, T_n.

***γ)* Τυχαία αποκοπή**

Ο κάθε ασθενής μπορεί να αποκοπεί τυχαία σε διαφορετικούς χρόνους λόγω αιτιών όπως:

- Απώλειες από την επανεξέταση (*loss to follow up*)

- Διακοπή (*drop out*)

- τέλος της μελέτης.

Έστω T_1, T_2, …, T_n οι χρόνοι επιβίωσης και C_1, C_2, …, C_n οι χρόνοι αποκοπής. Εδώ αντί για τα T_1, T_2 , …, T_n παρατηρούμε τα ζεύγη (Y_i , Δ_i) όπου:

$$Y_i = min(T_i, C_i), i = 1, 2, \dots, n \qquad (1.2.3)$$

$$\Delta_i = \begin{cases} 1, & T_i \leq c_\iota \quad (\text{μη αποκομμένη}) \\ 0, & T_i > c_i \quad\quad (\text{αποκομμένη}) \end{cases} \qquad (1.2.4)$$

Η Δ_i είναι δείκτρια που μας δείχνει αν η παρατήρηση είναι «αποκομμένη» ή όχι. Η δεξιά αποκοπή είναι η πιο διαδεδομένη μορφή αποκοπής.

Αριστερή αποκοπή

Αυτού του είδους η αποκοπή χρησιμοποιείται γιατί το γεγονός που μας ενδιαφέρει μπορεί να έχει συμβεί πριν αρχίσουμε να παρατηρούμε και συνήθως ο χρόνος, ο οποίος συνέβη αυτό, είναι άγνωστος.

Έστω T_1, T_2, …, T_n οι χρόνοι επιβίωσης και C_1, C_2, …, C_n οι χρόνοι αποκοπής. Εδώ αντί για τα T_1, T_2, …, T_n παρατηρούμε τα ζεύγη (Y_i, Δ_i) όπου:

$$Y_i = \max(T_{i,} C_i), \; i = 1, 2, \dots, n \qquad (1.2.5)$$

$$\Delta_i = \begin{cases} 1, & T_i \geq c_i \quad (\text{μη αποκομμένη}) \\ 0, & T_i < c_i \quad\quad (\text{αποκομμένη}) \end{cases} \qquad (1.2.6)$$

Η Δ_i είναι δείκτρια που μας δείχνει αν η παρατήρηση είναι αποκομμένη ή όχι.

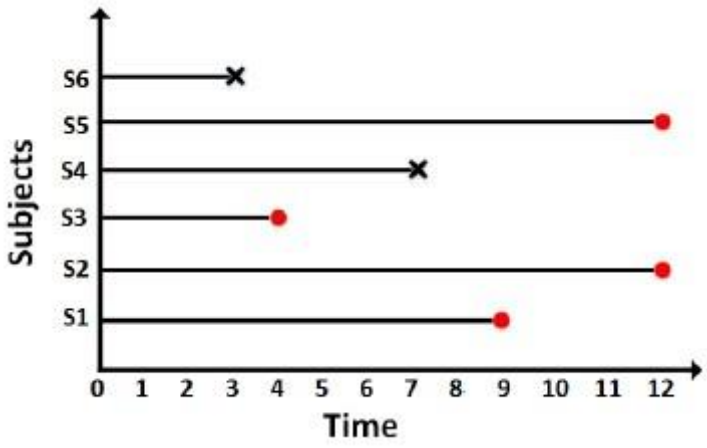
Αποκοπή σε διάστημα

Μερικές φορές έχουμε αποκοπή και από δεξιά και από αριστερά. Για παράδειγμα, σε μια μελέτη που παρατηρούμε το χρόνο που τα παιδιά μαθαίνουν να διαβάζουν, μπορεί όταν ξεκινήσουμε την έρευνα κάποια παιδιά να έχουν μάθει ήδη να διαβάζουν και κάποια άλλα να μην έχουν μάθει έως το τέλος της μελέτης μας.

Έστω T_1, T_2, …, T_n οι χρόνοι επιβίωσης και L_1, L_2, …, L_n οι χρόνοι αποκοπής από αριστερά και R_1, R_2, …, R_n οι χρόνοι αποκοπής από δεξιά. Δηλαδή έχουμε:

$$L_i \leq T_i \leq R_i.$$

Στο παρακάτω σχήμα φαίνεται ένα οπτικοποιημένο παράδειγμα για την καλύτερη κατανόηση της αποκοπής και της δομής των δεδομένων επιβίωσης.



Σχήμα 1.1 Μια απεικόνιση του προβλήματος της ανάλυσης επιβίωσης

Στο σχήμα αυτό παρουσιάζεται η παρατήρηση 6 περιπτώσεων σε μια διαχρονική μελέτη για 12 μήνες, δίνονται οι πληροφορίες για την εμφάνιση των γεγονότων κατά τη διάρκεια της μελέτης και καταγράφεται η χρονική περίοδος εμφάνισής τους. Μπορούμε, λοιπόν, να διαπιστώσουμε ότι μόνο τα αντικείμενα (*subjects*) $S4$ και $S6$ έχουν εκδηλώσει το γεγονός (τα οποία είναι σημειωμένα στο σχήμα με $'X'$) κατά τη διάρκεια της παρακολούθησης και ο χρόνος παρατήρησης γι' αυτούς είναι ο χρόνος εκδήλωσης του γεγονότος. Από την άλλη τα αντικείμενα $S1$, $S2$, $S3$ και $S5$ (τα οποίο είναι σημειωμένα με κόκκινο χρώμα στο σχήμα) είναι αυτά τα οποία δεν πραγματοποιήθηκε το γεγονός εντός της περιόδου των 12 μηνών. Ειδικότερα, τα αντικείμενα $S2$ και $S5$ είναι αποκομμένα, καθώς δεν πραγματοποιήθηκε κανένα γεγονός κατά τη διάρκεια της περιόδου μελέτης της έρευνα, ενώ τα αντικείμενα $S1$ και $S3$ είναι αποκομμένα λόγω του ότι βγήκαν απ' την έρευνα ή σταμάτησε η παρατήρησή τους εντός της χρονικής περιόδου της έρευνας.

Κολοβά δεδομένα

Ένα κολοβό δείγμα δεδομένων (*truncated data*) μπορεί να το σκεφτεί κανείς σαν ένα κομμάτι ενός μεγαλύτερου δείγματος, στο οποίο όλες οι παρατηρήσεις, που είναι έξω από κάποια όρια, αποκλείονται από το δείγμα (οι παρατηρήσεις αυτές χάνονται εντελώς από το δείγμα). Υπάρχουν δεδομένα που είναι κολοβά από δεξιά, από αριστερά και σε διάστημα.

Γενικά, αν Y ο χρόνος που περικόπτει τα δεδομένα από αριστερά τότε παρατηρούνται μόνο άτομα X για τα οποία , $X \geq Y$. Στη δεξιά περικοπή έχουμε ότι παρατηρούνται μόνο άτομα X για τα οποία, $X \leq Y$. Στην περικοπή σε διάστημα έχουμε ότι παρατηρούμε μόνο άτομα X για τα οποία, $L \leq X \leq R$, όπου το L είναι ένα κατώτατο όριο χρόνου και το R είναι ένα ανώτατο όριο αντίστοιχα. Η διαφορά των κολοβών με τα αποκομμένα δεδομένα είναι ότι στα κολοβά δεδομένα εμείς γνωρίζουμε πόσες παρατηρήσεις απέτυχαν να μπουν στο διάστημα αποκοπής μας.

1.3 Βασικές Συναρτήσεις του Χρόνου Επιβίωσης

Έστω $T > 0$ τυχαία μεταβλητή που παριστάνει το χρόνο επιβίωσης μιας υπό μελέτη ομάδας.

Αυτός ο χρόνος μπορεί να είναι ο χρόνος λειτουργίας ενός μηχανήματος, ο χρόνος μέχρι το θάνατο, η επανεμφάνιση μιας ασθένειας ή και η ίασή της. *Τέσσερις είναι οι βασικές συναρτήσεις* που χαρακτηρίζουν την κατανομή του T και αυτές είναι:

-η Συνάρτηση Επιβίωσης *(survival function)*

-η Συνάρτηση Διακινδύνευσης *(hazard function ή risk function)*

-η Συνάρτηση πυκνότητας-πιθανότητας *(probability density function)*

-η Μέση Υπολειπόμενη Διάρκεια Ζωής *(mean residual life)*

1.3.1 Η Συνάρτηση Επιβίωσης

Η συνάρτηση επιβίωσης (*survival function*) συμβολίζεται με $S(t)$, η οποία είναι η πιθανότητα ένα άτομο να επιβιώσει μέχρι το χρόνο t.

Ορίζεται ως:

$$S(t) = 1 - F(t) = P[T > t] \tag{1.3.1}$$

Έχουμε ότι:

$$S(0) = 1 \tag{1.3.2}$$

και

$$S(\infty) = 0. \tag{1.3.3}$$

Η συνάρτηση επιβίωσης είναι μια μη αρνητική φθίνουσα συνάρτηση του t. Όταν η T είναι συνεχής τυχαία μεταβλητή, τότε η $S(t)$ υπολογίζεται ως εξής:

$$S(t) = \int_t^{\infty} f(u)du. \tag{1.3.4}$$

Όταν η T είναι διακριτή τυχαία μεταβλητή, τότε η $S(t)$ υπολογίζεται ως εξής:

$$S(t) = \sum_{u \geq t} f(u). \tag{1.3.5}$$

Η γραφική παράσταση της $S(t)$ συναρτήσει του T είναι γνωστή ως καμπύλη επιβίωσης και είναι πολύ σημαντική στην ανάλυση δεδομένων χρόνου επιβίωσης.

Η συνάρτηση πυκνότητας-πιθανότητας της τυχαίας μεταβλητής T υπολογίζεται ως εξής:

$$f(t) = \frac{d}{dt}F(t) = -\frac{d}{dt}S(t). \tag{1.3.6}$$

Ο μέσος χρόνος επιβίωσης υπολογίζεται ως εξής:

$$\mu(t) = E(t) = \int_0^\infty t\, f(t)dt = -\int_0^\infty t\frac{dS(t)}{dt}dt$$
$$= -[tS(t)]_0^\infty + \int_0^\infty S(t)\, dt$$
$$= \int_0^\infty S(t)\, dt \qquad (1.3.7)$$

1.3.2 Η Συνάρτηση Διακινδύνευσης

Η συνάρτηση διακινδύνευσης (*hazard function/ risk function*) συμβολίζεται με $h(t)$ και εκφράζει την τάση προς διακοπή ενός αντικειμένου στο χρονικό διάστημα $(t, t + dt]$ με δεδομένη την επιβίωσή του μέχρι τη χρονική στιγμή t.

Έχουμε ότι

$$P[t < T \le t + dt | T > t] = \frac{P[t < T \le t + dt]}{P[T > t]}$$
$$= \frac{S(t) - S(t + dt)}{S(t)} \approx \frac{f(t)dt}{S(t)} \qquad (1.3.8)$$

και ορίζουμε τη συνάρτηση διακινδύνευσης ως:

$$h(t) = \lim_{\Delta t \to 0} \frac{P[t < T \le t + dt | T \ge t]}{\Delta t} = \frac{f(t)}{S(t)}, \qquad (1.3.9)$$

όπου T είναι συνεχής τυχαία μεταβλητή.

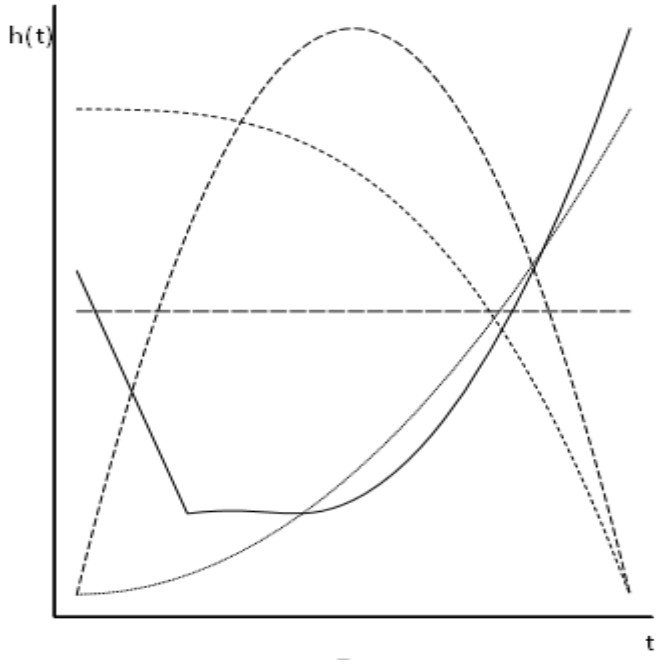
Μία σχετική ποσότητα που χρησιμοποιείται επίσης είναι η Σωρευτική συνάρτηση διακινδύνευσης (*cumulative hazard function*) που συμβολίζεται με $H(t)$ και δίνεται από τη σχέση**:**

$$H(t) = \int_0^t h(x)dx \qquad (1.3.10)$$

και υπολογίζεται ως εξής:

$$H(t) = \int_0^t h(x)dx = \int_0^t \frac{f(x)}{1-F(x)}dx = \int_0^t -\frac{d}{dx}\ln[1-F(x)]\,dx =$$
$$= -\ln[1-F(x)]\Big|_0^t = -\ln[1-F(t))] = -\ln(S(t)) \Longrightarrow \quad (1.3.11)$$
$$S(t) = e^{-H(t)}$$

Παρακάτω βλέπουμε ένα διάγραμμα από διάφορες συναρτήσεις της πάντα θετικής συνάρτησης διακινδύνευσης και παρατηρούμε ότι μπορεί να είναι αύξουσες, φθίνουσες ή και στο σχήμα τύπου *bathtub* (όπως είναι για παράδειγμα το γράφημα της συνεχόμενης γραμμής).



Σχήμα 1.2 Γραφικές παραστάσεις συναρτήσεων h(t) σε συνάρτηση με το χρόνο.

Η συνάρτηση διακινδύνευσης, συνήθως, μας δίνει περισσότερες πληροφορίες από ότι η συνάρτηση επιβίωσης και για το λόγο αυτό πολλές φορές ο υπολογισμός της $h(t)$ είναι η κύρια μέθοδος για την ανάλυση δεδομένων επιβίωσης.

1.3.3 Η Συνάρτηση πυκνότητας πιθανότητας

Η συνάρτηση πυκνότητας πιθανότητας (σ.π.π., *probability density function*) είναι η άνευ όρων πιθανότητα το γεγονός να πραγματοποιηθεί σε χρόνο t.

Αν T η διάρκεια ζωής, συνεχής τυχαία μεταβλητή τότε έχουμε ότι έχει σ.π.π. $f(t)$ με $t \geq 0$, και η συνάρτηση κατανομής ορίζεται ως

$$F(t) = P[T \leq t] = \int_0^t f(x)dx. \tag{1.3.12}$$

1.3.4 Η Μέση Υπολειπόμενη Διάρκεια Ζωής

Η Μέση Υπολειπόμενη Διάρκεια Ζωής (*mean residual life*) η οποία συμβολίζεται με $mrl(t)$, είναι ο μέσος χρόνος επιβίωσης αυτών που έχουν ήδη επιβιώσει έως το χρόνο t. Δηλαδή η αναμενόμενη διάρκεια ζωής στην ηλικία t, ορίζεται ως:

$$mrl(t) = E[T - t|T \geq t] = \int_0^t \frac{(x-t)f(x)}{P(T \geq t)}dx \tag{1.3.13}$$

με

$$mrl\,(0) = \mu. \tag{1.3.14}$$

Παρατήρηση :

Γενικά αν ξέρουμε οποιαδήποτε από τις τέσσερις παραπάνω συναρτήσεις που παρουσιάστηκαν στις παραγράφους 1.3.1, 1.3.2, 1.3.3 και 1.3.4 μπορούμε εύκολα να υπολογίσουμε όλες τις υπόλοιπες.

1.4 Μέθοδοι εκτίμησης των Βασικών Συναρτήσεων

Υπάρχουν τρεις κατηγορίες στατιστικών μεθόδων για την εκτίμηση των βασικών συναρτήσεων που περιγράψαμε στην προηγούμενη παράγραφο, οι μη παραμετρικές, οι ημι-παραμετρικές και οι παραμετρικές μέθοδοι.

Στις μη παραμετρικές μεθόδους οι εμπειρικές συναρτήσεις που χρησιμοποιούνται ευρέως είναι η *Kaplan-Meier (KM)* , η *Nelson-Aalen (NA)* και η μέθοδος *Life-Table (LT)*.

Πιο συγκεκριμένα, ο *KM* εκτιμητής για την πιθανότητα επιβίωσης στον συγκεκριμένο χρόνο επιβίωσης είναι ένα γινόμενο της ίδιας εκτίμησης μέχρι τον προηγούμενο χρόνο και του παρατηρούμενου ποσοστού επιβίωσης για εκείνο τον δεδομένο χρόνο. Έτσι, η μέθοδος *KM* αναφέρεται επίσης ως μέθοδος ορίου προϊόντος. Η μέθοδος *NA* είναι ένας εκτιμητής που βασίζεται στις σύγχρονες τεχνικές διαδικασίες καταμέτρησης. Η *LT* είναι η εφαρμογή της μεθόδου *KM* στα δεδομένα επιβίωσης που έχουν ταξινομηθεί με διαστήματα.

Το πλεονεκτήματα των μεθόδων αυτών είναι ότι είναι πιο αποτελεσματικές όταν δεν υπάρχουν κατάλληλες γνωστές θεωρητικές κατανομές και το μειονέκτημα ότι είναι δύσκολο να ερμηνευτούν και ότι δίνουν ανακριβείς εκτιμήσεις.

Στις ημι-παραμετρικές μεθόδους το μοντέλο *Cox* είναι η συνηθέστερα χρησιμοποιούμενη προσέγγιση ανάλυσης παλινδρόμησης για δεδομένα επιβίωσης.

Πιο συγκεκριμένα, το μοντέλο του *Cox* βασίζεται στην υπόθεση της αναλογικής διακινδύνευσης και χρησιμοποιεί μερική συνάρτηση πιθανοφάνειας για την εκτίμηση των παραμέτρων. Η ανάλυση παλινδρόμησης *Cox* περιγράφεται ως ημιπαραμετρική μέθοδος, αφού η κατανομή του αποτελέσματος παραμένει άγνωστη ακόμα και αν βασίζεται σε ένα παραμετρικό μοντέλο παλινδρόμησης. Επιπρόσθετα, στη βιβλιογραφία προτείνονται επίσης πολλές χρήσιμες παραλλαγές του βασικού μοντέλου *Cox*, όπως τα υποτιθέμενα μοντέλα *Cox,* ο αλγόριθμος *CoxBoost* και το μοντέλο *Time-Dependent*.

Το πλεονέκτημα της μεθόδου αυτής είναι, ότι η γνώση της κατανομής των χρόνων επιβίωσης δεν απαιτείται. Το μειονέκτημα είναι ότι η κατανομή του αποτελέσματος δεν είναι γνωστή και άρα δεν είναι εύκολο να ερμηνευτεί.

Στις παραμετρικές μεθόδους χρησιμοποιούμε κυρίως τη μέθοδο της Γραμμικής Παλινδρόμησης και ο Επιταχυνόμενος Χρόνος Αποτυχίας (*Accelerated Failure Time, AFT*).

Το μοντέλο παλινδρόμησης *Tobit*, το μοντέλο *Buckley-James* και η *penalized* παλινδρόμηση είναι τα πιο συχνά χρησιμοποιούμενα γραμμικά μοντέλα για την ανάλυση επιβίωσης. Οι παραμετρικές μέθοδοι είναι πιο αποτελεσματικές και ακριβείς για τον υπολογισμό, όταν ο χρόνος για το ενδιαφέρον γεγονός ακολουθεί μια συγκεκριμένη κατανομή που καθορίζεται από ορισμένες παραμέτρους. Είναι σχετικά εύκολο να εκτιμηθούν οι χρόνοι με τα παραμετρικά μοντέλα, αλλά γίνεται δύσκολο ή και αδύνατο να γίνει με το μοντέλο *Cox*.

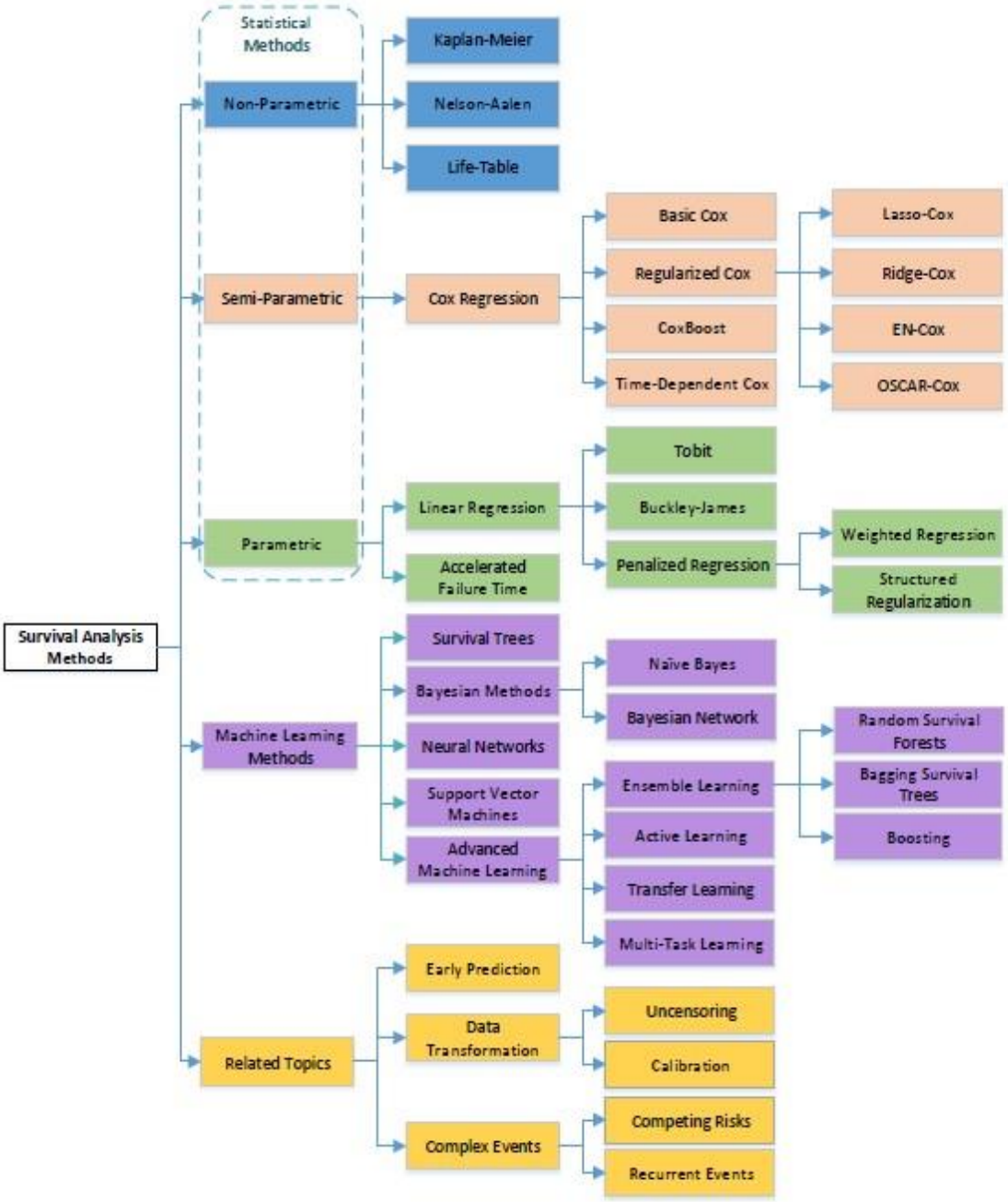
Το πλεονέκτημα της μεθόδου αυτής είναι η εύκολη ερμηνεία καθώς είναι πιο αποτελεσματική και ακριβής, όταν οι χρόνοι επιβίωσης ακολουθούν μια συγκεκριμένη κατανομή. Το μειονέκτημα είναι, ότι όταν η υπόθεση της κατανομής δεν είναι σωστή, μπορεί να είναι ασυνεπής και να μη δίνει τα βέλτιστα αποτελέσματα.

Πέρα όμως από τις στατιστικές μεθόδους που χρησιμοποιούμε για την εκτίμηση των Βασικών Συναρτήσεων της Ανάλυσης Επιβίωσης, τα τελευταία χρόνια έχει μπει δυναμικά και η μηχανική μάθηση (*machine learning*) στον τομέα αυτό, καθώς με τις μεθόδους της, μπορεί να διευκολύνει τους υπολογισμούς ειδικά στις περιπτώσεις που έχουμε αποκοπή των δεδομένων και θέλουμε και την εκτίμηση του χρόνου για το μοντέλο. Η μηχανική μάθηση μπορεί να μας βοηθήσει σε περιπτώσεις που έχουμε πολλά δεδομένα και ένα λογικό αριθμό από διαστάσεις, πράγμα βέβαια που δε συναντάμε πάντα στην Ανάλυση Επιβίωσης.

1.5 Ταξινόμηση Μεθόδων Ανάλυσης Επιβίωσης

Σε γενικές γραμμές, οι μέθοδοι ανάλυσης επιβίωσης μπορούν να ταξινομηθούν σε δύο κύριες κατηγορίες: τις στατιστικές μεθόδους και τις μεθόδους βασισμένες στην μηχανική μάθηση. Και οι δύο μέθοδοι έχουν τον κοινό στόχο να κάνουν προβλέψεις για τον χρόνο εκδήλωσης ενός γεγονότος. Ωστόσο, οι στατιστικές μέθοδοι εστιάζουν περισσότερο στις κατανομές των χρόνων των γεγονότων και τις στατιστικές ιδιότητες της εκτίμησης παραμέτρων. Οι μέθοδοι μηχανικής μάθησης συνήθως εφαρμόζονται σε προβλήματα μεγάλων διαστάσεων, ενώ οι στατιστικές μέθοδοι έχουν γενικά αναπτυχθεί για δεδομένα μικρών διαστάσεων. Επιπλέον, οι μέθοδοι μηχανικής μάθησης για την ανάλυση επιβίωσης προσφέρουν πιο αποτελεσματικούς αλγόριθμους καθώς ενσωματώνουν προβλήματα επιβίωσης και με στατιστικές μεθόδους και με μεθόδους μηχανικής μάθησης καθώς εκμεταλλεύονται τα πλεονεκτήματα των πρόσφατων εξελίξεων στη μηχανική μάθηση και τη βελτιστοποίηση για να μαθαίνουν τις εξαρτήσεις μεταξύ των συνδιακυμάνσεων και των χρόνων επιβίωσης με διαφορετικούς τρόπους.

Με βάση τις υποθέσεις και τη χρήση των παραμέτρων που χρησιμοποιούνται στο μοντέλο, οι παραδοσιακές στατιστικές μέθοδοι μπορούν να υποδιαιρεθούν σε τρεις κατηγορίες όπως αναφέρθηκε και στην προηγούμενη παράγραφο: (i) τα μη παραμετρικά μοντέλα, (ii) τα ημιπαραμετρικά μοντέλα και (iii) τα παραμετρικά μοντέλα. Οι αλγόριθμοι Μηχανικης Μάθησης, όπως τα Δέντρα Επιβίωσης, οι Μπευζιανοί μέθοδοι, τα Νευρωνικά Δίκτυα και οι Μηχανές Διανυσμάτων Υποστήριξης, τα οποία έχουν γίνει πιο δημοφιλή τα τελευταία χρόνια, περιλαμβάνονται σε ξεχωριστό κλάδο. Αρκετές προχωρημένες προσεγγίσεις μηχανικής μάθησης, συμπεριλαμβανομένων των Μεθόδων Μάθησης του Συνόλου, τη Διαδραστική μάθηση, τη Μεταφορά Μάθησης και τη Μάθηση Πολλαπλών Εργασιών. Η συνολική ταξινόμηση περιλαμβάνει επίσης μερικά θέματα της έρευνας που σχετίζονται με την ανάλυση επιβίωσης, όπως σύνθετα γεγονότα, μετασχηματισμός δεδομένων και πρόωρη πρόβλεψη. Μια πλήρης ταξινόμηση αυτών των μεθόδων ανάλυσης επιβίωσης είναι που φαίνεται στο παρακάτω σχήμα 1.3.



Σχήμα 1.3 Ταξινόμηση των μεθόδων που έχουν σχεδιαστεί για την Ανάλυση Επιβίωσης

1.6 Αξιολόγηση Απόδοσης με το δείκτη σύγκρισης

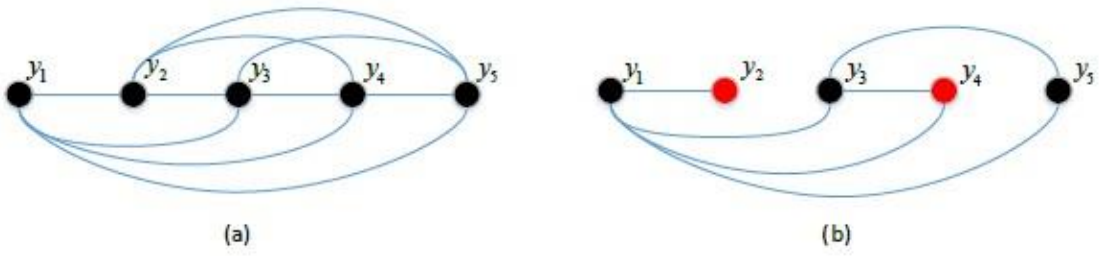
Στην Ανάλυση Επιβίωσης, τα κλασικά μέτρα αξιολόγησης της απόδοσης, όπως το R^2 και η τετραγωνική ρίζα του, δεν είναι κατάλληλα για να χρησιμοποιηθούν λόγω την παρουσίας αποκομμένων δεδομένων (*Heagerty and Zheng*, 2005). Έτσι, χρησιμοποιούμε πιο εξειδικευμένα μέτρα απόδοσης, ένα από τα οποία είναι και ο δείκτης σύγκρισης (*c-index, concordance index*).

Ένας κοινός τρόπος για την αξιολόγηση ενός μοντέλου, γενικά στην ανάλυση επιβίωσης, είναι να ληφθεί υπόψη ο σχετικός κίνδυνος ενός γεγονότος για διαφορετική περίπτωση, αντί για τους απόλυτους χρόνους επιβίωσης για κάθε περίπτωση. Αυτό μπορεί να γίνει με υπολογισμό της πιθανότητας σύγκρισης (*concordance probability*) ή του δείκτη σύγκρισης (*c-index*) (*Harrell et al.* 1984, *Harrell et al*. 1982, *Pencina and D'Agostino* 2004). Οι χρόνοι επιβίωσης δύο γεγονότων μπορούν να ταξινομηθούν για δύο περιπτώσεις:

(i) και οι δύο να είναι χωρίς αποκοπή,

(ii) ο παρατηρούμενος χρόνος ενός γεγονότος χωρίς αποκοπή είναι μικρότερος από τον χρόνο αποκοπής του γεγονότος με αποκοπή (*Steck et al.* 2008).

Αυτό μπορεί να απεικονιστεί με το διατεταγμένο γράφημα που δίνεται στο σχήμα 1.4. Το σχήμα 1.4 (a) και το σχήμα 1.4 (b) που απεικονίζουν τις πιθανές συγκρίσεις κατάταξης (που υποδηλώνονται από τις άκρες μεταξύ των γεγονότων) για τα δεδομένα επιβίωσης χωρίς και με αποκοπή, αντίστοιχα.



Σχήμα 1.4 Απεικόνιση των περιορισμών κατάταξης σε δεδομένα επιβίωσης για υπολογισμούς δείκτη σύγκρισης (*c-index)* $(y_1 < y_2 < y_3 < y_4)$. Εδώ, οι μαύρες κουκίδες δείχνουν τους παρατηρούμενους χρόνους γεγονότων και οι κόκκινες κουκίδες δείχνουν τις αποκομμένες παρατηρήσεις. Το σχήμα (α) μας δείχνει δεδομένα χωρίς αποκοπή και το (β) δεδομένα με αποκοπή.

Υπάρχουν $\binom{5}{2} = 10$ πιθανές συγκρίσεις ανά ζεύγη, για τις πέντε περιπτώσεις στα δεδομένα επιβίωσης χωρίς αποκοπή που φαίνονται στο σχήμα $4\,(a)$. Λόγω της παρουσίας αποκομμένων περιπτώσεων (παριστάνονται με κόκκινους κύκλους) στο Σχήμα $4\,(b)$, μόνο 6 από τις 10 συγκρίσεις είναι εφικτές. Θα πρέπει να επισημανθεί ότι ένα αποκομμένο γεγονός μπορεί να συγκριθεί μόνο με ένα προηγούμενο χωρίς αποκοπή (για παράδειγμα η y_2 με την y_1). Ωστόσο, οποιοδήποτε γεγονός με αποκοπή δεν μπορεί να συγκριθεί με περιπτώσεις με ή και χωρίς αποκοπή, εφόσον έχει υποστεί αποκοπή, δεδομένου ότι ο πραγματικός χρόνος του γεγονότος είναι άγνωστος (για παράδειγμα y_2 με y_3 και y_2 με y_4). Λαμβάνουμαι υπόψη τόσο τις παρατηρήσεις όσο και τις εκτιμήσεις δύο γεγονότων, $(y_1, \widehat{y_1})$ και $(y_2, \widehat{y_2})$, όπου το y_i αντιπροσωπεύει τον πραγματικό χρόνο παρατήρησης του γεγονότος και το $\widehat{y_i}$ την εκτιμώμενη τιμή, αντίστοιχα. Η πιθανότητα σύγκρισης μεταξύ τους μπορεί να υπολογιστεί ως:

$$c = P(\widehat{y_1} > \widehat{y_2} | y_1 \geq y_2). \tag{1.6.1}$$

Καθώς ο παραπάνω ορισμός δεν είναι απλός, στην πράξη, υπάρχουν πολλοί τρόποι υπολογισμού του δείκτη σύγκρισης:

1. Όταν το αποτέλεσμα του μοντέλου είναι ένα ποσοστό κινδύνου (όπως τα αποτελέσματα που προκύπτουν απ' τα μοντέλα του *Cox*), ο c- δείκτης μπορεί να υπολογιστεί χρησιμοποιώντας την παρακάτω σχέση:

   $$\hat{c} = \frac{1}{num} \sum_{i:\,\delta_i=1} \sum_{j:y_i<y_j} I[X_i\hat{\beta} > X_j\hat{\beta}], \tag{1.6.2}$$

   όπου $i, j \in \{1,2, \dots, N\}$, num είναι ο αριθμός όλων των συγκρίσιμων ζευγαριών, $I[\cdot]$ είναι η δείκτρια συνάρτηση και το $\hat{\beta}$ είναι παράμετρος εκτιμημένη απ' τα μοντέλα του *Cox.*
2. Στις μεθόδους επιβίωσης οι οποίες στοχεύουν στην άμεση εκμάθηση του χρόνου επιβίωσης, το $c-index$ πρέπει να υπολογίζεται ως εξής:

   $$\hat{c} = \frac{1}{num} \sum_{i:\,\delta_i=1} \sum_{j:y_i<y_j} I[S(\widehat{y_j}|X_j) > S(\widehat{y_i}|X_i)], \tag{1.6.3}$$

   όπου $S(\cdot)$ είναι η εκτιμώμενη πιθανότητα επιβίωσης.

Κεφάλαιο 2 – Εισαγωγή στις Μηχανές Διανυσμάτων Υποστήριξης

2.1 Γενικά περί Μηχανικής Μάθησης και Ανάλυσης Επιβίωσης

Η ακριβής πρόβλεψη του χρονικού διαστήματος που θα συμβεί ένα γεγονός είναι ένα πολύ σημαντικό πρόβλημα στην ανάλυση δεδομένων. Μία από τις πολλές προκλήσεις που πρέπει να αντιμετωπίσουμε είναι τα δεδομένα τα οποία υφίστανται κάποιου είδους αποκοπή. Η Ανάλυση Επιβίωσης έχει επεξεργαστεί πολλούς τρόπους με τους οποίους αντιμετωπίζουμε τέτοιου είδους δεδομένα, χρησιμοποιώντας τα παραδοσιακά εργαλεία της στατιστικής.

Επίσης πολλοί αλγόριθμοι μηχανικής μάθησης έχουν αναπτυχθεί ώστε να χειρίζονται αποτελεσματικά τα δεδομένα επιβίωσης και να αντιμετωπίζουν τα προβλήματα που προκύπτουν σε πραγματικά δεδομένα. Τα τελευταία χρόνια, χάρη στα πλεονεκτήματα των τεχνικών μηχανικής μάθησης, όπως η ικανότητά τους να μοντελοποιούν τις μη γραμμικές σχέσεις και η ποιότητα των συνολικών προβλέψεών τους, έχουν επιτύχει σημαντική επιτυχία σε διάφορους πρακτικούς τομείς. Στην ανάλυση επιβίωσης, η κύρια πρόκληση των μεθόδων μηχανικής μάθησης είναι η δυσκολία να αντιμετωπιστούν κατάλληλα οι αποκομμένες πληροφορίες και η χρονική εκτίμηση του μοντέλου.

2.2 Παρουσίαση βασικών Μεθόδων Μηχανικής Μάθησης

Οι βασικές Μέθοδοι Μηχανικής Μάθησης που χρησιμοποιούνται στην Ανάλυση επιβίωσης είναι τα Δέντρα Επιβίωσης (*Survival Trees*), οι Μπευζιανοί Μέθοδοι (*Bayesian Methods*), τα Τεχνητά Νευρωνικά Δίκτυα (*Artificial Neural Networks*), άλλες Προχωρημένες Προσεγγίσεις Μηχανικής Μάθησης (*Advanced Machine Learning Approaches*) και οι Μηχανές Διανυσμάτων Υποστήριξης (*Support Vector Machines*). Θα γίνει μια σύντομη περιγραφή της

λειτουργίας των τεσσάρων πρώτων μεθόδων και μια πιο αναλυτική περιγραφή του εργαλείου που μας ενδιαφέρει στην υπάρχουσα εργασία, των Μηχανών Διανυσμάτων Υποστήριξης (*Support Vector Machines*).

2.2.1 Δέντρα Επιβίωσης

Τα δέντρα επιβίωσης (*Survival Trees*) είναι μία μορφή ταξινόμησης και παλινδρόμησης που προσαρμόζονται για να χειρίζονται τα αποκομμένα δεδομένα. Η βασική διαίσθηση πίσω από τα μοντέλα των δένδρων είναι η κατανομή των δεδομένων με βάση ένα συγκεκριμένο κριτήριο διαίρεσης έτσι ώστε τα αντικείμενα που είναι παρόμοια μεταξύ τους με βάση το γεγονός που μας ενδιαφέρει, να τοποθετηθούν στον ίδιο κόμβο. Η αρχική απόπειρα χρήσης μιας δομής δέντρου για δεδομένα επιβίωσης έγινε από τους *Ciampi et αl*. (1981). Ωστόσο, στους *Gordon and Olshen* (1985) ανήκει η πρώτη αναφορά που συζητήθηκε η δημιουργία δένδρων επιβίωσης.

Η κύρια διαφορά μεταξύ ενός δένδρου επιβίωσης και του τυποποιημένου δέντρου αποφάσεων είναι η επιλογή του κριτηρίου διαίρεσης. Η μέθοδος του δέντρου αποφάσεων εκτελεί αναδρομικό διαχωρισμό στα δεδομένα, ορίζοντας ένα κατώφλι για κάθε χαρακτηριστικό, ωστόσο δεν μπορεί να εξετάσει, ούτε τις αλληλεπιδράσεις μεταξύ των χαρακτηριστικών, ούτε τις αποκομμένες πληροφορίες στο μοντέλο (*Safavian and Landgrebe*, 1991). Τα κριτήρια διαίρεσης που χρησιμοποιούνται για τα δέντρα επιβίωσης μπορούν να ομαδοποιηθούν σε δύο κατηγορίες:

(i) να μεγιστοποιηθεί η ετερογένεια μεταξύ κόμβων και
(ii) να ελαχιστοποιηθεί η ομοιογένεια εντός κόμβου.

Η πρώτη κατηγορία προσεγγίσεων ελαχιστοποιεί τη συνάρτηση απώλειας χρησιμοποιώντας το κριτήριο ομοιογένειας εντός κόμβου. Για την περίπτωση αυτή, κάποιοι ερευνητές μετρούσαν τις αποστάσεις ομοιογένειας και *Hellinger* (*Gordon and Olshen, 1985*) μεταξύ των εκτιμώμενων λειτουργιών διανομής χρησιμοποιώντας τη μετρική *Wasserstein*. Μια εκθετική συνάρτηση *loglikelihood* χρησιμοποιήθηκε από άλλους (*Davis and Anderson, 1989*), για αναδρομικό διαχωρισμό με βάση το άθροισμα υπολοίπων από το μοντέλο *Cox*.

Οι *Leblanc* και *Crowley* (*1992*) μέτρησαν την απόκλιση του κόμβου με βάση το πρώτο βήμα μιας διαδικασίας εκτίμησης πλήρους πιθανοφάνειας. Στη δεύτερη κατηγορία κριτηρίων διαίρεσης, χρησιμοποιήθηκαν στατιστικές δοκιμής *log-rank* για μέτρα ετερογένειας μεταξύ κόμβων (*Ciampi et al., 1986*). Αργότερα, οι *Ciampi et al.* (*1987*) πρότειναν μια στατιστική αναλογία πιθανοτήτων για να μετρηθεί η ανομοιότητα μεταξύ δύο κόμβων. Με βάση την τάξη στατιστικών δύο δειγμάτων *Tarone-Ware*, ο *Segal* (*Segal, 1988*) εισήγαγε μια διαδικασία μέτρησης της μεταξύ τους σχέσης. Η βελτιστοποίηση του επιπέδου της παράδοσης είναι η ικανότητά του να χειρίζεται τα αποκομμένα δεδομένα χρησιμοποιώντας τη δομή του δέντρου.

Μια άλλη πολύ σημαντική επιλογή κατασκευής του δέντρου επιβίωσης, είναι η επιλογή του τελικού δέντρου. Μπορούν να ακολουθηθούν διαδικασίες όπως η επιλογή προς τα πίσω ή η επιλογή προς τα εμπρός για την επιλογή του βέλτιστου δέντρου (*Bou-Hamad et al.*, 2011).

2.2.2 Μπευζιανοί Μέθοδοι

Το θεώρημα *Bayes* είναι μία από θεμελιώδεις αρχές της θεωρίας των πιθανοτήτων και των μαθηματικών στατιστικών, καθώς παρέχει μια σχέση μεταξύ της εκ των προτέρων και εκ των υστέρων πιθανότητας, με τέτοιο τρόπο έτσι ώστε να μπορούμε να δούμε τις διαφορές στις πιθανότητες, πριν και μετά την πραγματοποίησης του γεγονότος. Υπάρχουν δύο μοντέλα που χρησιμοποιούν το θεώρημα του *Bayes*, που ονομάζονται *Naïve Bayes* (NB) και Μπευζιανό Δίκτυο (*Bayesian Network (BN)*) (*Friedman et al.*, 1997). Και οι δύο αυτές προσεγγίσεις, μελετώνται συχνά στο πλαίσιο της κλινικής πρόβλεψης (*Kononenko* 1993, *Pepe* 2003, *Zupan et al*. 2000). Τα πειραματικά αποτελέσματα της χρήσης Μπευζιανών μεθόδων (*Bayesian Methods*) σε δεδομένα επιβίωσης δείχνουν ότι οι Μπευζιανές μέθοδοι έχουν καλές ιδιότητες τόσο ερμηνείας όσο και λογικής αβεβαιότητας (*Raftery et al.* 1995).

Το *Naíve Bayes* μοντέλο, μια γνωστή πιθανολογική μέθοδος στη μηχανική μάθηση, είναι ένας από τους πιο απλούς αλλά αποτελεσματικούς αλγόριθμους πρόβλεψης. Σε έρευνες (*Bellazzi and Zupan*, 2008),

χρησιμοποιήθηκε ένας *Naive Bayesian* ταξινομητής (*naive Bayesian classifier*) για να γίνουν προβλέψεις στην κλινική ιατρική, εκτιμώντας διάφορες πιθανότητες από τα δεδομένα. Άλλη μέθοδος, που χρησιμοποιήθηκε από τους (*Fard et al*., 2016) είναι η ενσωμάτωση Μπευζιανών μεθόδων με ένα μοντέλο AFT με την εκ των προτέρων πιθανότητα προκειμένου να προβλέψει τα δεδομένα επιβίωσης πρώιμου σταδίου για τη μελλοντική πιθανότητα επιβίωσης. Ένα μειονέκτημα της μεθόδου *Naïve Bayes* είναι ότι παίρνει την ανεξαρτησία των δεδομένων σαν δεδομένη, γεγονός το οποίο μπορεί να μην είναι αληθές για πολλά προβλήματα στην ανάλυση επιβίωσης.

Το Μπευζιανό Δίκτυο, στο οποίο τα χαρακτηριστικά μπορούν να σχετίζονται μεταξύ τους σε διάφορα επίπεδα, μπορεί να αντιπροσωπεύει γραφικά μια θεωρητική κατανομή σε ένα σύνολο μεταβλητών. Τα Μπευζιανά Δίκτυα μπορούν να αντιπροσωπεύουν οπτικά όλες τις σχέσεις μεταξύ των μεταβλητών, γεγονός που την καθιστά εύκολη στο να ερμηνευτούν τα τελικά αποτελέσματα από τον ερευνητή. Το BN μπορεί να αποκτήσει πληροφορίες γνώσης χρησιμοποιώντας διαδικασίες εκτίμησης των δομών δικτύου και των παραμέτρων από ένα δεδομένο σύνολο δεδομένων. Σε έρευνες (*Lisboa et al., 2003*), επίσης, έχει προταθεί ένα πλαίσιο Μπευζιανού νευρωνικού δικτύου για την επιλογή μοντέλου για «διαχρονικά (*longitudinal)* δεδομένα» χρησιμοποιώντας αυτόματο προσδιορισμό συνάφειας (*MacKay, 1995*). Έχει προταθεί (*Raftery, 1995*) ένα Μπευζιανό πρότυπο μέσου όρου για μοντέλα αναλογικού κινδύνου *Cox* και χρησιμοποιείται επίσης για την αξιολόγηση των παραγόντων *Bayes* στο πρόβλημα. Πιο πρόσφατα, σε έρευνες (*Fard et al., 2016*) προτάθηκε ένα νέο πλαίσιο το οποίο συνδυάζει τη δύναμη της αναπαράστασης του Μπευζιανού με το μοντέλο AFT, εξάγοντας τις προηγούμενες πιθανότητες σε μελλοντικά χρονικά σημεία.

2.2.3 Τα Τεχνητά Νευρωνικά Δίκτυα

Εμπνευσμένος από τα βιολογικά νευρικά συστήματα, το 1958 ο *Frank Rosenblatt* δημοσίευσε το πρώτο βιβλίο (*Rosenblatt*, 1958) σχετικά με το τεχνητό νευρωνικό δίκτυο (*Artificial Neural Network*, (ANN)). Σε αυτή την

προσέγγιση, οι απλοί τεχνητοί κόμβοι που δηλώνονται από τους "νευρώνες" συνδέονται με βάση μια ζυγισμένη σύνδεση για να σχηματίσουν ένα δίκτυο που προσομοιώνει ένα βιολογικό νευρωνικό δίκτυο. Ένας νευρώνας στο πλαίσιο αυτό είναι ένα στοιχείο υπολογιστικής που αποτελείται από σύνολα προσαρμοσμένων βαρών και δημιουργεί την έξοδο με βάση ένα συγκεκριμένο είδος λειτουργίας ενεργοποίησης. Τα τεχνητά νευρωνικά δίκτυα (ANN) έχουν χρησιμοποιηθεί ευρέως στην ανάλυση επιβίωσης. Στην βιβλιογραφία προτείνονται τρία είδη μεθόδων που χρησιμοποιούν τη μέθοδο του νευρικού δικτύου για την επίλυση των προβλημάτων ανάλυσης επιβίωσης:

i) Η ανάλυση επιβίωσης του νευρωνικού δικτύου έχει χρησιμοποιηθεί για την πρόβλεψη του χρόνου επιβίωσης ενός ατόμου απευθείας από τις δεδομένες εισροές.

ii) Μία επέκταση του μοντέλου *Cox* PH από τους *Faraggi* και *Simon* το 1995 στον μη γραμμικό ANN παράγοντα πρόβλεψης, όπου πρότειναν να μπει στο νευρωνικό δίκτυο το οποίο έχει ένα γραμμικό επίπεδο εξόδου και ένα λογιστικό κρυμμένο στοιχείο εξόδου. Η τεχνική αυτή έχει εφαρμοστεί από τους *Mariani et al*. (1997) για την εκτίμηση των προγνωστικών παραγόντων για την επανεμφάνιση του καρκίνου του μαστού. Αν και αυτές οι επεκτάσεις για το μοντέλο *Cox* επέτρεψαν τη διατήρηση των περισσότερων πλεονεκτημάτων ενός τυπικού μοντέλου PH, δεν ήταν ακόμα ο βέλτιστος τρόπος για να μοντελοποιηθεί η παραλλαγή της βασικής γραμμής (*Baesens et al.*, 2005).

iii) Πολλές προσεγγίσεις (*Liestbl et al.*, 1994- *Biganzoli et al*., 1998- *Brown et al.*, 1997- *Ravdin and Clark*, 1992- *Lisboa et al.*, 2003) παίρνουν την κατάσταση επιβίωσης ενός υποκειμένου, το οποίο μπορεί να εκπροσωπείται από την πιθανότητα επιβίωσης ή διακινδύνευσης, ως την έξοδο του νευρωνικού δικτύου. Σε έρευνες (*Biganzoli et al.*, 1998) έχει εφαρμοστεί η μέθοδος μερικής λογιστικής τεχνητής νευρωνικής δικτύωσης (*Partial Logistic Artificial Neural Network, PLANN*) για να αναλύσει τη σχέση μεταξύ των χαρακτηριστικών και των χρόνων επιβίωσης, ώστε να επιτευχθεί καλύτερη πρόβλεψη από το μοντέλο.

Τελευταία, τα νευρωνικά δίκτυα τροφοδοσίας προς τα εμπρός χρησιμοποιούνται για να αποκτήσουν ένα πιο ευέλικτο μη γραμμικό μοντέλο εξετάζοντας τις αποκομμένες πληροφορίες στα δεδομένα χρησιμοποιώντας μια γενίκευση τόσο των μοντέλων συνεχούς όσο και του διακριτού χρόνου (*Biganzoli et al.*, 1998). Σε έρευνα (*Lisboa et al.*, 2003) το PLANN επεκτάθηκε σε ένα Μπευζιανό νευρωνικό πλαίσιο με μεταβλητή κανονικοποίηση για να μεταφέρει την επιλογή του μοντέλου χρησιμοποιώντας αυτόματο προσδιορισμό της συνάφειας (*MacKay*, 1995).

2.2.4 Προχωρημένες Προσεγγίσεις Μηχανικής Μάθησης

Τα τελευταία χρόνια, όλο και πιο προχωρημένες προσεγγίσεις μηχανικής μάθησης (*Advanced Machine Learning Approaches*) έχουν δημιουργηθεί για να μπορέσουν να γίνουν σωστές προβλέψεις των αποκομμένων δεδομένων. Αυτές οι μέθοδοι έχουν πολλά πλεονεκτήματα στα δεδομένα διάρκειας επιβίωσης σε σύγκριση με τις μεθόδους που έχουν ήδη αναφερθεί. Κάποιες από αυτές είναι οι εξής:

1) **Μέθοδοι Μάθησης του Συνόλου** (*Ensemble learning methods*) (*Dietterich*, 2000) όπου δημιουργούν ταξινομητές που προβλέπουν της ετικέτες κλάσης για τα νέα δεδομένα, παίρνοντας σταθμισμένα αποτελέσματα πρόβλεψης από όλους τους ταξινομητές. Συχνά είναι δυνατή η κατασκευή καλών συνόλων και η επίτευξη καλύτερης προσέγγισης της άγνωστης λειτουργίας με τη μεταβολή των αρχικών σημείων, ειδικά όταν τα δεδομένα είναι ανεπαρκή. Για να ξεπεραστεί η αστάθεια μιας μεμονωμένης μεθόδου, προτάθηκαν από τον *Breiman* τα *Bagging* (*Breiman*, 1996) και τα *Random Forests* (*Breiman*, 2001), που, χρησιμοποιούνται συνήθως για να δημιουργήσουν το μοντέλο. Τέτοια μοντέλα συνόλου έχουν προσαρμοστεί με επιτυχία στην ανάλυση επιβίωσης. Πιο αναλυτικά:

Bagging Δέντρα Επιβίωσης:

Το *bagging* είναι ένα από τις παλαιότερες και πιο συνηθισμένες Μεθόδους Μάθησης του Συνόλου, η οποία συνήθως μειώνει τη διακύμανση των βασικών μοντέλων που χρησιμοποιούνται. Με τα *bagging* δέντρα επιβίωσης, η συνολική συνάρτηση επιβίωσης μπορεί να υπολογιστεί υπολογίζοντας κατά μέσο όρο τις προβλέψεις ενός μόνο δέντρου επιβίωσης, αντί να πάρει την πλειοψηφία (*Hothorn et al.*, 2004). Υπάρχουν κυρίως τρία βήματα σε αυτήν τη μέθοδο:

(i) Ο σχεδιασμός ενός *B* bootstrap δείγματος από τα δοσμένα δεδομένα.

(ii) Για κάθε δείγμα *bootstrap*, δημιουργείται ένα δέντρο επιβίωσης και εξασφαλίζεται ότι, για όλους τους τερματικούς κόμβους, ο αριθμός των συμβάντων είναι μεγαλύτερος ή ίσος με το κατώφλι d.

(iii) Με τον υπολογισμό του μέσου όρου των προβλέψεων των κόμβων φύλλων (*leaf nodes*), υπολογίζεται η συνολική συνάρτηση επιβίωσης *bootstrap*. Για κάθε κόμβο φύλλων η συνάρτηση επιβίωσης εκτιμάται χρησιμοποιώντας τον εκτιμητή *KM* και για όλες τις μεμονωμένες παρατηρήσεις στον ίδιο κόμβο γίνεται η υπόθεση ότι έχουν την ίδια συνάρτηση επιβίωσης.

Τυχαία Δάση Επιβίωσης:

Τυχαίο Δάσος Επιβίωσης (*Random Survival Forest* (*RSF*)) είναι μια μέθοδος μάθησης συνόλου που προτείνεται ειδικά για να κάνει προβλέψεις χρησιμοποιώντας τα δομημένα μοντέλα δέντρων (*Breiman*, 2001). Βασίζεται σε ένα πλαίσιο παρόμοιο με το *bagging*. Η κύρια διαφορά μεταξύ των τυχαίων δασών και του *bagging* είναι ότι σε έναν συγκεκριμένο κόμβο, τα τυχαία δέντρα αντί να χρησιμοποιούν όλα τα χαρακτηριστικά, χρησιμοποιούν μόνο ένα τυχαίο υποσύνολο των υπολειπόμενων χαρακτηριστικών, για να επιλέξουν τα χαρακτηριστικά που βασίζονται στο κριτήριο διαίρεσης.

Φαίνεται ότι η τυχαιοποίηση μπορεί να μειώσει τη συσχέτιση μεταξύ των δένδρων και έτσι να βελτιωθεί η απόδοση της πρόβλεψης. Τα τυχαία δάση επιβίωσης (*Ishwaran et al*., 2008) επέκτεινε ο *Breiman* με τη μέθοδο του για τυχαία δάση, χρησιμοποιώντας δάσος από δέντρα επιβίωσης για πρόβλεψη. Υπάρχουν κυρίως τέσσερα βήματα στο RSF:

(i) Ο σχεδιασμός *B bootstrap* δειγμάτων τυχαία από το δοσμένο σύνολο δεδομένων. Αυτό επίσης ονομάζεται out-of-bag (OOB) δεδομένα, επειδή περίπου το 37% των δεδομένων αποκλείεται σε κάθε δείγμα.

(ii) Για κάθε δείγμα, γίνεται η κατασκευή ενός δέντρου επιβίωσης, επιλέγοντας τυχαία τα χαρακτηριστικά και διαιρώντας τον κόμβο χρησιμοποιώντας το υποψήφιο χαρακτηριστικό, το οποίο μπορεί να μεγιστοποιήσει τη διαφορά επιβίωσης μεταξύ των «παιδικών» (*child*) κόμβων.

(iii) Γίνεται η δημιουργία του δέντρου σε πλήρες μέγεθος, με τον περιορισμό, ότι ο τερματικός κόμβος έχει μεγαλύτερο ή ίσο με ένα συγκεκριμένο μοναδικό αριθμό θανάτων.

(iv) Με χρήση του μη παραμετρικού εκτιμητή *Nelson-Aalen*, υπολογίζεται η σωρευτική συνάρτηση κινδύνου (CHF) του συνόλου των δεδομένων OOB, λαμβάνοντας το μέσο όρο της CHF κάθε δέντρου. Επιπλέον, οι συγγραφείς στο (*Ishwaran et al*., 2011) παρέχουν έναν αποτελεσματικό τρόπο για την εφαρμογή των RSF για προβλήματα ανάλυσης επιβίωσης μεγάλων διαστάσεων μέσω κανονικοποίησης των δασών.

Boosting:

Ο αλγόριθμος *boosting* είναι μια από τις ευρέως χρησιμοποιούμενες μεθόδους συνόλου που έχουν σχεδιαστεί να συνδυάσουν τις βασικές μεταβλητές εκμάθησης σε ένα

σταθμισμένο άθροισμα που αντιπροσωπεύει το τελικό αποτέλεσμα της ισχυρής μεταβλητής εκμάθησης. Επαναλαμβάνει τα καταλλήλως καθορισμένα υπολείμματα βάσει του αλγόριθμου καθοδικής κλίσης (*gradient descent algorithm*) (*Hothorn et al.,* 2006 και *Bühlmann and Hothorn,* 2007). Οι συγγραφείς στο (*Hothorn et al.*, 2006) επεκτείνουν τον αλγόριθμο ενίσχυσης κλίσης για να ελαχιστοποιήσουν τη σταθμισμένη συνάρτηση κινδύνου

$$\hat{\beta}_{\breve{U},X} = \arg min_{\beta} \sum_{i=1}^{N} w_i(\breve{U}_i - h(X_i|\,\beta)), \tag{2.2.1}$$

όπου $\breve{U}$ είναι μια μεταβλητή ψευδοπρόβλεψης με

$$\boldsymbol{\breve{U}_i} = -\frac{\partial L(y_i, \varphi)}{\partial \varphi}\Big|_{\varphi = \widehat{f_m}(X_i)}, \beta \tag{2.2.2}$$

είναι ένα διάνυσμα παραμέτρων,

$h(\cdot\,|\beta_{U,X})$ είναι η πρόβλεψη που έγινε από την παλινδρόμηση του U χρησιμοποιώντας μία βασική μεταβλητή εκμάθησης. Στη συνέχεια, τα βήματα για τη βελτιστοποίηση του προβλήματος είναι τα ακόλουθα:

(i) Αρχικοποιούνται τα εξής $\boldsymbol{\breve{U}_i} = y_i (i = 1, \dots, N)$, $m = 0$ και $\widehat{f_0}(\cdot\,|\hat{\beta}_{\breve{U},X})$ και γίνεται καθορισμός του αριθμού των επαναλήψεων σε M $(M > 1)$.

(ii) Γίνεται προσαρμογή $h(\cdot\,|\hat{\beta}_{U,X})$ μετά την ενημέρωση των υπολειμμάτων $\boldsymbol{\breve{U}_i}$ $(i = 1, \dots, N)$.

(iii) Επαναληπτικά γίνεται ενημέρωση της

$$\widehat{f_{m+1}}(\cdot) = \widehat{f_m}(\cdot) + uh(\cdot\,|\tilde{\beta}_{U,X}) \tag{2.2.3}$$

όπου το $0 < u \leq 1$ αντιπροσωπεύει το μέγεθος του βήματος.

(iv) Επαναλαμβάνονται τα βήματα (ii) και (iii) έως ότου $m = M$.

2) **Διαδραστική μάθηση** (*Active Learning*). Η μέθοδος αυτή που βασίζεται στα δεδομένα με τις αποκομμένες παρατηρήσεις μπορεί να

είναι πολύ χρήσιμη για την ανάλυση επιβίωσης, καθώς οι απόψεις ενός εμπειρογνώμονα στον τομέα μπορούν να ενσωματωθούν στα μοντέλα. Ο μηχανισμός διαδραστικής μάθησης επιτρέπει στο μοντέλο επιβίωσης να επιλέξει ένα υποσύνολο υποκειμένων, μαθαίνοντας από ένα περιορισμένο σύνολο επισημασμένων υποκειμένων πρώτα και στη συνέχεια να ζητήσει από τον ειδικό να βρει την κατάσταση επιβίωσης πριν εξεταστεί σαν όλο. Η ανατροφοδότηση από τον ειδικό είναι ιδιαίτερα χρήσιμη για τη βελτίωση του μοντέλου σε πολλούς τομείς εφαρμογής πραγματικού κόσμου (*Vinzamuri et al*., 2014). Ο στόχος της ενεργητικής μάθησης για προβλήματα ανάλυσης επιβίωσης είναι να οικοδομηθεί ένα μοντέλο παλινδρόμησης επιβίωσης με την χρησιμοποίηση των αποκομμένων δεδομένων, χωρίς τη διαγραφή ή την τροποποίησή τους.

3) **Μεταφορά Μάθησης** (*Transfer Learning*). Η συλλογή επισημασμένων πληροφοριών σε προβλήματα επιβίωσης είναι πολύ χρονοβόρα, δηλαδή, πρέπει να περιμένουμε για την εμφάνιση του συμβάντος από έναν επαρκή αριθμό εκπαιδευτικών περιπτώσεων για την κατασκευή αξιόπιστων μοντέλων. Μια λύση γι' αυτό το ανεπαρκές πρόβλημα δεδομένων είναι η απλή ενσωμάτωση των δεδομένων από συναφή προβλήματα σε μια ενοποιημένη μορφή και η δημιουργία προτύπων πρόβλεψης σε τέτοια ολοκληρωμένα δεδομένα. Ωστόσο, οι προσεγγίσεις αυτές συχνά δεν αποδίδουν καλά επειδή το κύριο πρόβλημα (για το οποίο πρόκειται να γίνουν οι προβλέψεις) θα κατακλυστεί από βοηθητικά δεδομένα με διαφορετικές κατανομές. Σε τέτοια σενάρια, η μεταφορά γνώσης μεταξύ σχετικών προβλημάτων θα παράγει συνήθως καλύτερα αποτελέσματα σε σύγκριση με μια προσέγγιση ενσωμάτωσης δεδομένων. Η μέθοδος εκμάθησης μεταφοράς έχει μελετηθεί εκτενώς για την επίλυση τυπικών προβλημάτων ταξινόμησης και παλινδρόμησης (*Pan and Yang* 2010). Ένα κανονικοποιημένο μοντέλο *Cox PH* (*Li et al*., 2016a), που ονομάζεται *Transfer-Cox*,

προτείνεται να βελτιώσει την απόδοση πρόβλεψης του μοντέλου *Cox* στον πρόβλημα μέσω της μεταφοράς γνώσης από την πηγή στα μοντέλα επιβίωσης που βασίζονται σε πολλαπλά σύνολα δεδομένων μεγάλης διαστάσεως. Το μοντέλο *Transfer-Cox* χρησιμοποιεί τη νόρμα $l_{2,1}$ για να τιμωρήσει το άθροισμα των λειτουργιών απώλειας (μερική αρνητική *log-likelihood*) τόσο για τις πηγές όσο και για τους τομείς στόχων. Έτσι, όχι μόνο το μοντέλο θα επιλέξει σημαντικά χαρακτηριστικά, αλλά θα μάθει επίσης μια κοινή αντιπροσώπευση σε όλους τους τομείς προέλευσης και στόχου του προβλήματος, για να βελτιώσει την απόδοση του μοντέλου.

4) **Μάθηση Πολλαπλών Εργασιών** (*Multi-task Learning*). Στο (*Li et al*. 2016b), το πρόβλημα πρόβλεψης του χρόνου επιβίωσης αναδιατυπώθηκε ως πρόβλημα εκμάθησης πολλών καθηκόντων. Στα δεδομένα επιβίωσης, ο πίνακας με τα επισημασμένα αποτελέσματα είναι ελλιπής δεδομένου ότι η ετικέτα κάθε αποκομμένης παρατήρησης δεν είναι διαθέσιμη μετά τον αντίστοιχο χρόνο αποκοπής της. Επομένως, οι τυπικές μέθοδοι μάθησης πολλαπλών εργασιών δεν είναι κατάλληλες να χειριστούν τα αποκομμένα δεδομένα. Για να αντιμετωπιστεί αυτό το πρόβλημα, το Μοντέλο Μάθησης Πολλαπλών Εργασιών για την Ανάλυση Επιβίωσης (*Multi-task Learning Model For Survival Analysis,* MTLSA) μεταφράζει τις πρωτότυπες ετικέτες γεγονότων σε έναν NxK δείκτη πίνακα I, όπου το $K = max\ (y_i), \forall\ i = 1, \dots, N$ είναι ο μέγιστος χρόνος παρακολούθησης όλων των περιπτώσεων στο σύνολο δεδομένων. Το στοιχείο I_{ij} $(i = 1, \dots, N, j = 1, \dots, K)$, του δείκτη πίνακα θα είναι 1 εάν συνέβη το γεγονός πριν από το χρόνο y_j για την περίπτωση i, διαφορετικά θα είναι 0. Ένα από τα κύρια πλεονεκτήματα της προσέγγισης MTLSA είναι ότι μπορεί να εντοπίσει την εξάρτηση μεταξύ των αποτελεσμάτων σε διάφορα χρονικά σημεία, χρησιμοποιώντας μια κοινή αντιπροσώπευση σε όλες τις σχετικές εργασίες της μετατροπής, η οποία θα μειώσει το

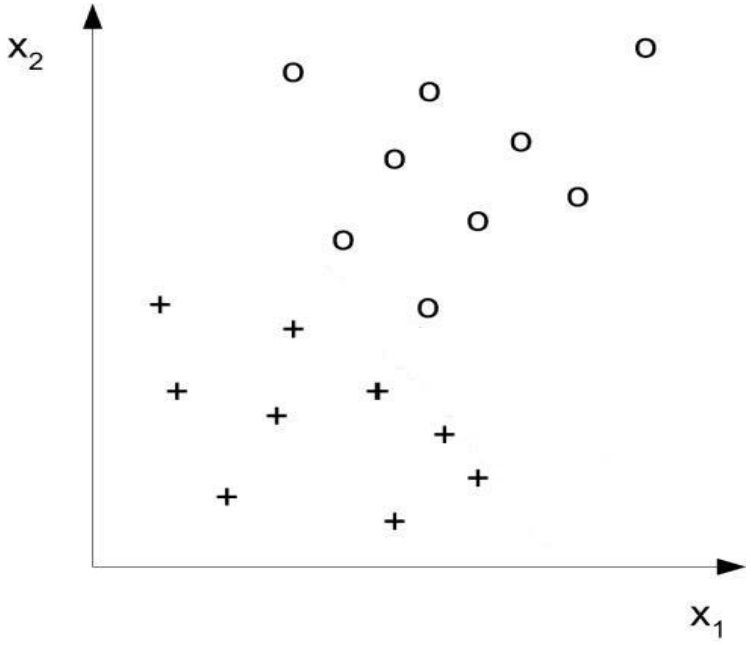
σφάλμα πρόβλεψης σε κάθε εργασία. Επιπλέον, το μοντέλο μπορεί να μάθει ταυτόχρονα από τις αποκομμένες και μη παρατηρήσεις βασιζόμενο στο δείκτη πίνακα. Ένα σημαντικό χαρακτηριστικό των μη επαναλαμβανόμενων γεγονότων, δηλαδή των γεγονότων τα οποία αν συμβούν μία φορά δεν ξανασυμβαίνουν, κωδικοποιείται μέσω της μη αρνητικής μη-αυξανόμενης λίστας περιορισμών δομής. Στον αλγόριθμο MTLSA χρησιμοποιείται η ποινή της νόρμας $l_{2,1}$ για να μάθουν μια κοινή αντιπροσώπευση σε συναφείς εργασίες και ως εκ τούτου να υπολογίσουν τη συγγένεια μεταξύ των επιμέρους μοντέλων που έχουν κατασκευαστεί για τους διάφορους μεμονωμένους χρόνους γεγονότων.

2.3 Περιγραφή των Μηχανών Διανυσμάτων Υποστήριξης

2.3.1 Γενική Ιδέα των Μηχανών Διανυσμάτων Υποστήριξης

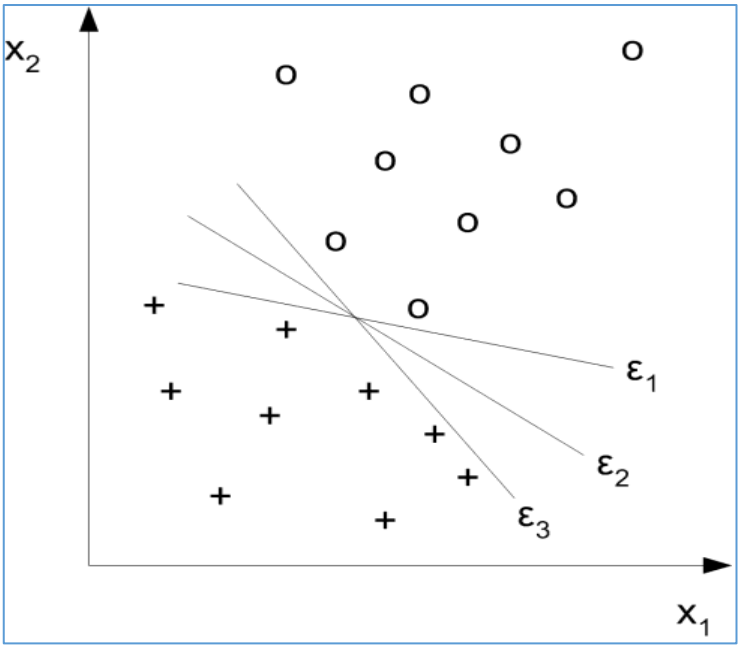
Οι Μηχανές Διανυσμάτων Υποστήριξης (*Support Vector Machines, SVM*) είναι μοντέλα τα οποία λύνουν προβλήματα ταξινόμησης προτύπων. Σκοπός είναι να δημιουργηθεί ένα υπερεπίπεδο το οποίο θα μπορεί να ταξινομεί τις υπάρχουσες παρατηρήσεις σε δύο κατηγορίες έτσι ώστε το περιθώριο διαχωρισμού των δύο επιφανειών να μεγιστοποιείται. Η λειτουργία των *SVMs* στηρίζεται στη δημιουργία των διανυσμάτων στήριξης (*Support Vectors*), τα οποία είναι ένα μικρό υποσύνολο των δοσμένων δεδομένων του προβλήματος και αποτελούν τα δεδομένα εκπαίδευσης του μοντέλου (*training dataset*). Τα μοντέλα αυτά έχουν μικρούς χρόνους εκπαίδευσης.

Τα *SVMs* ουσιαστικά αποτελούν ταξινομητές που διαχωρίζουν τα δεδομένα μας σε δύο διακριτές κατηγορίες. Για παράδειγμα, για να καταλάβουμε τη γενική ιδέα, ας υποθέσουμε ότι θέλουμε να κατηγοριοποιήσουμε σε δύο κατηγορίες, τα «+» και «ο» του Σχήματος 2.1.



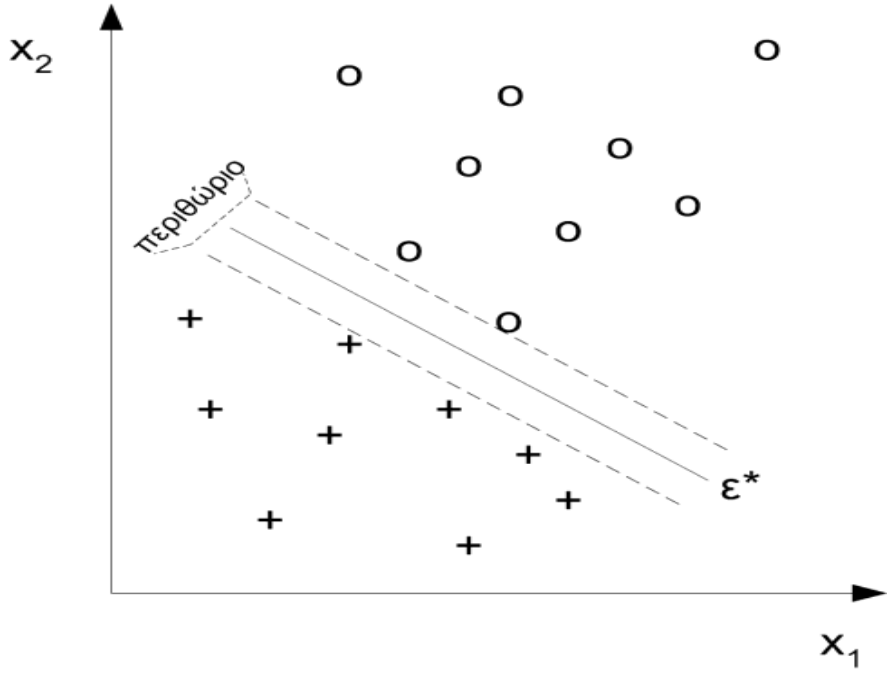
Σχήμα 2.1 Δεδομένα Κατηγοριοποίησης με ετικέτες «+» και «ο».

Συνεπώς ψάχνουμε να βρούμε μια ευθεία γραμμή η οποία να είναι ικανή να διαχωρίσει τη μία κατηγορία από την άλλη. Φαίνεται όμως στο Σχήμα 2.2 ότι υπάρχει παραπάνω από μία ευθεία οι οποίες μπορούν να διαχωρίσουν τα δεδομένα μου.



Σχήμα 2.2 Κατηγοριοποίηση των «+» και «ο» μέσω των ευθειών $\varepsilon_1, \varepsilon_2, \varepsilon_3$.

Άρα τα *SVMs* σκοπεύουν στην αναζήτηση της ευθείας αυτής η οποία θα διαχωρίζει τα δεδομένα μου με τέτοιο τρόπο έτσι ώστε το περιθώριο της ευθείας μεταξύ των κατηγοριών να μεγιστοποιείται, όπως φαίνεται στο Σχήμα 2.3.



Σχήμα 2.3 Η βέλτιστη ευθεία απόφασης ε^*

Σαφώς και δεν είναι όλες οι περιπτώσεις τόσο απλές καθώς μπορεί τα δεδομένα μας μπορεί να μην είναι γραμμικά διαχωρίσιμα. Παρακάτω παρουσιάζεται η διαδικασία δημιουργίας των SVM σε γραμμικά και μη, διαχωρίσιμα προβλήματα.

2.3.2 Γραμμικά Διαχωρίσιμα Προβλήματα

Ας υποθέσουμε ότι αντιμετωπίζουμε το πρόβλημα ταξινόμησης δύο κλάσεων C_0 και C_1 που είναι γραμμικά διαχωρίσιμες. Έτσι υπάρχουν ένα διάνυσμα w και ένα κατώφλι w_0 τέτοια ώστε:

$$\begin{cases} \boldsymbol{w^T x} + w_0 > 0\,, & \alpha\nu\ \boldsymbol{x} \in C_0 \\ \boldsymbol{w^T x} + w_0 < 0\,, & \alpha\nu\ \boldsymbol{x} \in C_1 \end{cases} \qquad (2.3.1)$$

Επειδή υπάρχουν πολλές λύσεις, δηλαδή άπειρα ζεύγη $(\boldsymbol{w}, w_0)$ για το διαχωρισμό των κλάσεων, θεωρούμε κάποιο κριτήριο αξιολόγησης των λύσεων.

Το κριτήριο αυτό είναι το περιθώριο ταξινόμησης (*margin*) γ μεταξύ των δύο κλάσεων, το οποίο ορίζεται ως το άθροισμα $\gamma = \gamma_0 + \gamma_1$ μεταξύ των δύο παρακάτω περιθωρίων (το γ_0 για την κλάση C_0 και το γ_1 για την κλάση C_1), όπου

$$\gamma_0 = \min_{x \in C_0} \frac{-(\boldsymbol{w^T x} + w_0)}{||w||} \tag{2.3.2}$$

$$\gamma_1 = \min_{x \in C_1} \frac{\boldsymbol{w^T x} + w_0}{||\boldsymbol{w}||}. \tag{2.3.3}$$

Τα πρότυπα x της κλάσης C_0 καθώς και τα πρότυπα x' της κλάσης C_1 , για τα οποία επιτυγχάνεται η ελάχιστη απόσταση (η ελάχιστη απόλυτη τιμή των γ_0 και γ_1 στους παραπάνω τύπους), λέγονται διανύσματα υποστήριξης (*support vectors*).

Στη συνέχεια θεωρούμε το «κανονικό διαχωριστικό υπερεπίπεδο» για το οποίο

α) το κατώφλι w_0 τοποθετείται ακριβώς στη μέση ανάμεσα στις δύο κλάσεις ($\gamma_0 = \gamma_1$) β) η κλιμάκωση των w και w_0 είναι τέτοια, ώστε:

$$\begin{cases} \boldsymbol{w^T x} + w_0 \leq -1, & \alpha\nu\ \boldsymbol{x} \in C_0 \\ \boldsymbol{w^T x} + w_0 > 1, & \alpha\nu\ \boldsymbol{x} \in C_1 \end{cases} \tag{2.3.4}$$

και για το οποίο προκύπτει $\gamma_0 = \gamma_1 = \frac{1}{||\boldsymbol{w}||}$ και άρα $\gamma = \frac{2}{||w||}$.

Το πρόβλημα συνεπώς είναι ο προσδιορισμός του ελαχίστου της συνάρτησης

$$J(\boldsymbol{w}, w_0) = \frac{1}{2}\left|\left|\boldsymbol{w}\right|\right|^2 \tag{2.3.5}$$

υπό τους περιορισμούς των P ανισοτήτων που δίνονται παρακάτω

$$d_i(w^T x_i + w_0) \geq 1, \text{ με } i = 1, \dots, P \tag{2.3.6}$$

Για την ελαχιστοποίηση της παραπάνω συνάρτησης χρησιμοποιούνται πολλαπλασιαστές Lagrange δημιουργώντας τη συνάρτηση κόστους

$L^d(\lambda_1,\ldots,\lambda_p) = -L(\lambda_1,\ldots,\lambda_p)$, η οποία πρέπει να ελαχιστοποιηθεί ως προς τα λ_i, όπου:

$$L(\lambda_1,\ldots,\lambda_p) = \sum_{i-1}^{P}\lambda_i - \frac{1}{2}\sum_{i-1}^{P}\sum_{j-1}^{P}\lambda_i\lambda_j d_i d_j \boldsymbol{x}_i^T \boldsymbol{x}_j. \tag{2.3.7}$$

Έτσι έχουμε να λύσουμε το δυϊκό του αρχικού προβλήματος, που είναι ο υπολογισμός του ελαχίστου της συνάρτησης L^d ως προς τα $\lambda_1,\ldots,\lambda_p$ υπό τους περιορισμούς:

$$\textstyle\sum_{i-1}^{P}\lambda_i d_i = 0,\ \lambda_i \geq 0,\ i = 1,\ldots,P. \tag{2.3.8}$$

Αποδεικνύεται ότι το διάνυσμα λύσης $\boldsymbol{w}$ είναι ένας θετικός γραμμικός συνδυασμός των διανυσμάτων υποστήριξης

$$\boldsymbol{w} = \sum\nolimits_{i\in I_{sv}} \lambda_{0,i} d_i \boldsymbol{x_i}, \tag{2.3.9}$$

όπου I_{SV} είναι το σύνολο που αποτελείται από τα διανύσματα υποστήριξης, και $\lambda_{0,i}$ είναι οι βέλτιστοι πολλαπλασιαστές *Lagrange* που βρίσκονται από τη λύση της (2.3.8).

Η βέλτιστη διαχωριστική επιφάνεια είναι η ακόλουθη:

$$g(\boldsymbol{x})^* = \sum\nolimits_{i\in I_{sv}} \lambda_{0,i} d_i\, \boldsymbol{x}_i^T \boldsymbol{x} + w_0, \tag{2.3.10}$$

και το κατώφλι υπολογίζεται ως εξής:

$$w_0 = \frac{1}{|I_{SV}|}\sum\nolimits_{i\in I_{sv}} \frac{1}{d_i} - \boldsymbol{w}^T \boldsymbol{x}_i\ . \tag{2.3.11}$$

2.3.3 Μη Γραμμικά Διαχωρίσιμα Προβλήματα

Όταν έχουμε μη γραμμικά προβλήματα, δηλαδή προβλήματα στα οποία οι κλάσεις δεν είναι γραμμικά διαχωρίσιμες, το πρόβλημα, είναι ο προσδιορισμός του ελαχίστου της συνάρτησης

$$J_{NS}(\boldsymbol{w}, w_0) = \frac{1}{2}||\boldsymbol{w}||^2 + c\sum_{i-1}^{P}\xi_i, \tag{2.3.12}$$

υπό τους περιορισμούς των παρακάτω P ανισοτήτων

$$d_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) \geq 1 - \xi_i, \mu\varepsilon\ \xi_i \geq 0 \tag{2.3.13}$$

όπου $i = 1, \dots, P$, η μεταβλητή ξ_i λέγεται μεταβλητή χαλαρότητας και η παράμετρος C επιλέγεται από το χρήστη και είναι το βάρος του κόστους των λάθος ταξινομήσεων. Το παραπάνω πρόβλημα μετασχηματίζεται στο ακόλουθο δυϊκό πρόβλημα υπολογισμού του ελαχίστου της συνάρτησης L^d ως προς τα $\lambda_1, \dots, \lambda_p$, όπως στην

$$L(\lambda_1, \dots, \lambda_p) = \sum_{i-1}^{P}\lambda_i - \frac{1}{2}\sum_{i-1}^{P}\sum_{j-1}^{P}\lambda_i\lambda_j d_i d_j \boldsymbol{x}_i^T\,\boldsymbol{x_j} \tag{2.3.14}$$

υπό τους περιορισμούς

$$\textstyle\sum_{i-1}^{P}\lambda_i d_i = 0,\ 0 \leq \lambda_i \leq c,\ i = 1, \dots, P. \tag{2.3.15}$$

Το διάνυσμα λύσης βρίσκεται ότι είναι το ίδιο όπως πριν :

$$\boldsymbol{w} = \sum_{i \in I_{sv}} \lambda_{0,i} d_i \boldsymbol{x}_i. \tag{2.3.16}$$

Σε αυτή την περίπτωση, επειδή χρησιμοποιούνται γραμμικές συναρτήσεις για διαχωρισμό μη γραμμικών κλάσεων, είναι πιθανόν πολλά πρότυπα να ταξινομηθούν λανθασμένα.

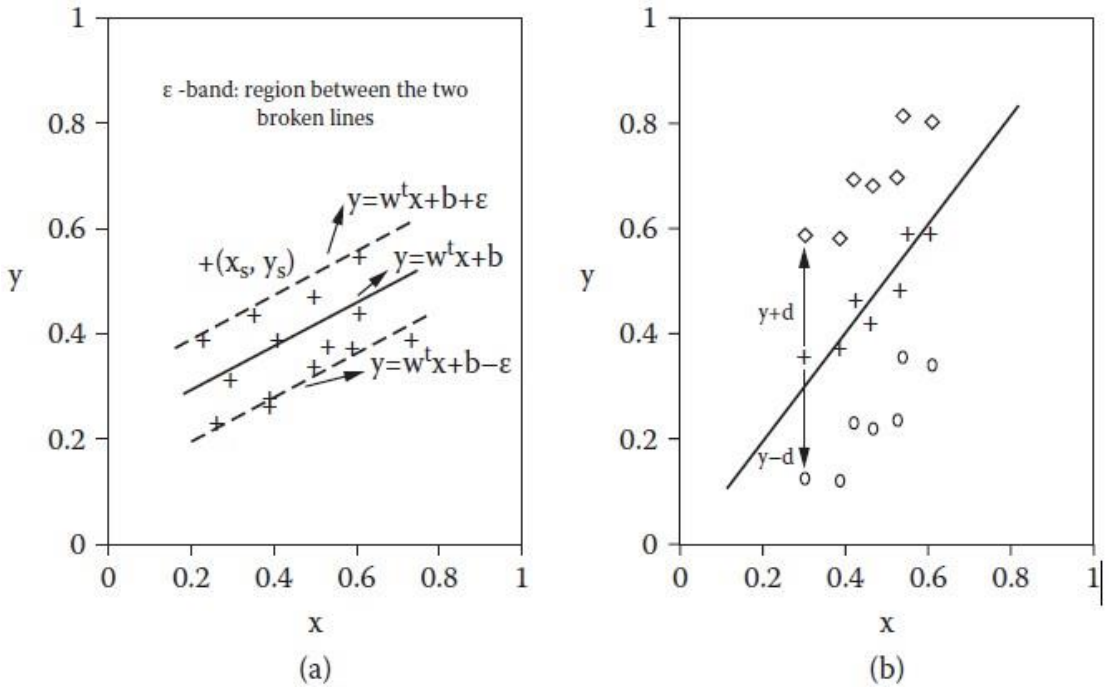
Κεφάλαιο 3– Ανάλυση Παλινδρόμησης με Μηχανές Διανυσμάτων Υποστήριξης στην Ανάλυση Επιβίωσης

3.1 ε- Ανάλυση Παλινδρόμησης με Μηχανές Διανυσμάτων Υποστήριξης (ε-SVR)

Οι μηχανές διανυσμάτων υποστήριξης έχουν επεκταθεί για να λύσουν το πρόβλημα της παλινδρόμησης με το δοσμένο σύνολο δεδομένων $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ που λαμβάνεται από μία λανθάνουσα συνάρτηση όπου $\boldsymbol{x_i}$ δηλώνει το διάνυσμα του δείγματος, το y_i η αντίστοιχη απόκριση και N είναι ο συνολικός αριθμός των δειγμάτων. Αρχικά θα πρέπει να δοθεί προσοχή σε δύο βασικές έννοιες στην SVR. Το ένα είναι η ε-ζώνη (***ε-band***) και η άλλη είναι η ε-μη ευαισθητοποιημένη συνάρτηση απώλειας (***ε-insensitive loss function***). Όπως και με την Ταξινόμηση με Μηχανές Διανυσμάτων Υποστήριξης (*Support Vector Classification*, SVC), τα αληθινά δεδομένα αρχικά απεικονίζονται μη γραμμικά σε έναν χώρο χαρακτηριστικών μεγάλων διαστάσεων και στη συνέχεια μια γραμμική συνάρτηση είναι προσαρμοσμένη για την προσέγγιση της λανθάνουσας συνάρτησης μεταξύ $\boldsymbol{X}$ και $\boldsymbol{y}$. Η κύρια ιδέα είναι να βρεθεί μια συνάρτηση $f(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} + b$ που προσεγγίζει τα $y_1, y_2, \ldots y_N$, η οποία να έχει τα περισσότερα ε παράγωγα από τις πραγματικές τιμές y_i και να είναι όσο το δυνατόν πιο «επίπεδη» (για να αποφευχθεί η υπερφόρτωση).

3.1.1 Η ε-ζώνη και η ε-μη ευαισθητοποιημένη συνάρτηση κόστους

Ας πάρουμε μια συνάρτηση μιας μεταβλητής, για παράδειγμα, όπως φαίνεται στο Σχήμα 3.1. Η ε-ζώνη αναφέρεται στην περιοχή μεταξύ των δύο διακεκομμένων γραμμών. Αυτή η περιοχή μπορεί να βρεθεί με τη μετακίνησης της συμπαγούς γραμμής μεταξύ της μετατόπισης ε πάνω και κάτω. Εδώ, ε είναι ένας προκαθορισμένος θετικός αριθμός.



Σχήμα 3.1: Σχέδιο (α): η γραφική παράσταση της ε-ζώνης της μονοδιάστατης συνάρτησης με προκαθορισμένο ε. Σχέδιο (b): Μετατροπή του προβλήματος παλινδρόμησης σε πρόβλημα κατάταξης. Για κάθε δείγμα x_i στα δεδομένα εκπαίδευσης (τα συν της γραφικής παράστασης), το αντίστοιχο y_i προστίθεται με ένα θετικό αριθμό d για να δημιουργήσει ένα νέο δείγμα (x_i, y_i^1) που ανήκει στην Κλάση 1. Ομοίως, το y_i μπορεί επίσης να αφαιρεθεί από το ίδιο d για να παραχθεί ένα άλλο νέο δείγμα (x_i, y_i^{-1}) που ανήκει στην κλάση -1. Επαναλαμβάνοντας αυτή τη διαδικασία, τα N δείγματα για παλινδρόμηση διπλασιάζονται και ταξινομούνται σε δύο κατηγορίες. Μετά το πρόβλημα παλινδρόμησης μετατρέπεται στο δυαδικό πρόβλημα ταξινόμησης. Έτσι ο αλγόριθμος για το SVC μπορεί να εφαρμοστεί εδώ για να λυθεί το πρόβλημα παλινδρόμησης.

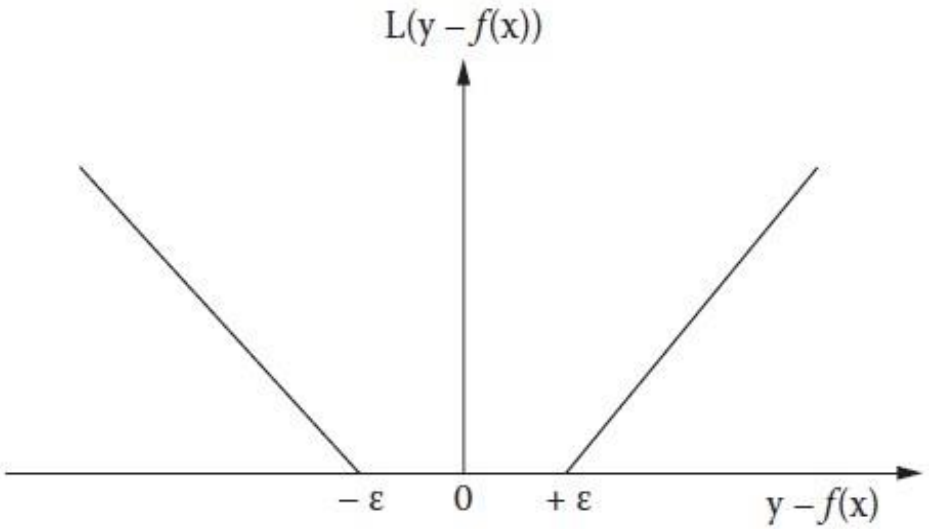
Η ε-μη ευαισθητοποιημένη συνάρτηση κόστους έχει την τμηματική μορφή:

$$L(y - f(\boldsymbol{x}), \varepsilon) = \begin{cases} |y - f(\boldsymbol{x})| - \varepsilon, & |y - f(\boldsymbol{x})| \geq \varepsilon \\ 0, & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases} \qquad (3.1.1)$$

Δηλαδή, μόνο τα σημεία δεδομένων έξω από την ε-ζώνη (π.χ., $(\boldsymbol{x_s}, y_s)$ στο Σχήμα 3.1α) προκαλούν απώλεια. Αυτού του είδους η συνάρτηση απώλειας απεικονίζεται στο Σχήμα 3.2.

3.1.2 Γραμμική ε-SVR

Προκειμένου να επιλυθεί το πρόβλημα παλινδρόμησης, οι *Cortes and Vapnik (1995)* μετέτρεψαν έξυπνα το πρόβλημα παλινδρόμησης σε πρόβλημα ταξινόμησης. Έπειτα η συνάρτηση παλινδρόμησης μπορεί να υπολογιστεί χρησιμοποιώντας τον ίδιο αλγόριθμο όπως περιγράφεται στο SVC. Το πρώτο βήμα που ασχολείται το πρόβλημα παλινδρόμησης είναι η μετατροπή του σε πρόβλημα ταξινόμησης, όπως φαίνεται στο Σχήμα 3.1 (b).



Σχήμα 3.2 Μια απεικόνιση της συνάρτησης ε-μη ευαισθητοποιημένης συνάρτησης.

Έχοντας το σύνολο δεδομένων εκπαίδευσης $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, ο γραμμικός ε-SVR αλγόριθμος στοχεύει θεωρητικά, στην επίλυση του προβλήματος βελτιστοποίησης, το οποίο μπορεί να γραφτεί μέσα στην ακόλουθη φόρμα με μια ε-μη ευαισθητοποιημένη συνάρτηση κόστους:

$$min\frac{1}{2}||w||^2 + \frac{C}{N}\sum_{i=1}^{N} L(y_i - f(x_i), \varepsilon), \tag{3.1.2}$$

όπου C είναι μια προκαθορισμένη παράμετρος κανονικοποίησης. Το παραπάνω πρόβλημα ελαχιστοποίησης μπορεί να εκφραστεί περαιτέρω στην ακόλουθη μορφή με τη χαλαρή μεταβλητή ξ_i^* ως εξής:

$$min\, L(\boldsymbol{w}, b, \xi^{(*)}) = \frac{1}{2}||w||^2 + \frac{C}{N}\sum_{i=1}^{N}(\xi_i + \xi_i^*), \tag{3.1.3}$$

το οποίο υπακούει σε:

$$(\boldsymbol{w}^T\boldsymbol{x}_i + b) - y_i \leq \varepsilon + \xi_i, \quad i = 1,2,..,N$$
$$y_i - (\boldsymbol{w}^T\boldsymbol{x}_i + b) \leq \varepsilon + \xi_i^*, i = 1,2,..,N \tag{3.1.4}$$
$$\xi_i^{(*)} \geq 0, \qquad i = 1,2,..,N$$

Με τη βοήθεια της μεθόδου των πολλαπλασιαστών *Lagrange* και του αλγόριθμου *QP*, η συνάρτηση παλινδρόμησης μπορεί να οριστεί ως:

$$f(\boldsymbol{x}) = \sum_{i=1}^{N}(\boldsymbol{\alpha}_{i,f}^* - \boldsymbol{a}_{i,f})(\boldsymbol{x}_i{}^T\boldsymbol{x}) + b_f \tag{3.1.5}$$

$$b_f = y_j - \sum_{i=1}^{N}(\boldsymbol{\alpha}_{i,f}^* - \boldsymbol{a}_{i,f})(\boldsymbol{x}_i{}^T\boldsymbol{x_j}) + \varepsilon \tag{3.1.6}$$

όπου $\boldsymbol{\alpha}_{i,f}^*$ και $\boldsymbol{\alpha}_{i,f}$, είναι οι βελτιστοποιημένοι πολλαπλασιαστές *Lagrange*, αντίστοιχα.

3.1.3 ε-SVR βασισμένη σε πυρήνα

Είναι γνωστό ότι τα μη γραμμικά σύνολα δεδομένων, είναι η συνηθέστερη περίπτωση στα πραγματικά δεδομένα. Επομένως, είναι απαραίτητο να επεκταθεί η γραμμική ε-SVR σε μη γραμμική παλινδρόμηση. Εισάγοντας τη συνάρτηση του πυρήνα, η αρχική είσοδος ήταν πρώτα απεικονισμένη μη γραμμικά στο χώρο των χαρακτηριστικών και το προκύπτον ε-SVR γίνεται τόσο ευέλικτο ώστε να μπορεί να χρησιμοποιηθεί για να αντιμετωπίσει το περίπλοκο μη γραμμικό πρόβλημα παλινδρόμησης στη χημεία.

Καθώς η διαδικασία δημιουργίας της τελικής συνάρτησης απόφασης είναι αρκετά παρόμοια με την γραμμική περίπτωση, δίνουμε εδώ μόνο την τελική μαθηματική μορφή:

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} (\boldsymbol{\alpha}_{i,f}^{*} - \boldsymbol{a}_{i,f}) K(\boldsymbol{x_i x}) + b_f \tag{3.1.7}$$

$$b_f = y_j - \sum_{i=1}^{N} (\boldsymbol{\alpha}_{i,f}^{*} - \boldsymbol{a}_{i,f})(\boldsymbol{x}_i{}^T \boldsymbol{x_j}) + \varepsilon \tag{3.1.8}$$

όπου $\boldsymbol{\alpha}_{i,f}^{*}$ και $\boldsymbol{\alpha}_{i,f}$, είναι οι βελτιστοποιημένοι πολλαπλασιαστές *Lagrange*, αντίστοιχα.

Επιπλέον, θα πρέπει να επισημάνουμε ότι το ν-*SVR* ($\nu \in [0,1]$) θα μπορούσε να αντιμετωπίζεται ως μια τροποποιημένη εκδοχή του αρχικού ε-*SVR* με βάση την συνάρτηση πυρήνα. Στο ν-SVR, το ν είναι μια παράμετρος σπανιότητας. Μετά την επιλογή του ν, η τιμή ε ρυθμίζεται αυτόματα από τον αλγόριθμο. Εδώ ν είναι το κάτω όριο του ποσοστού των διανυσμάτων υποστήριξης στα συνολικά δείγματα και το ανώτερο όριο του κλάσματος των σφαλμάτων στον ίδιο χρόνο.

3.2 Εισαγωγή στην Παλινδρόμηση Διανυσμάτων Υποστήριξης

Σε αυτό το κεφάλαιο ξεκινάμε μια σύντομη παρουσίαση για τα υπάρχοντα μοντέλα επιβίωσης βασισμένα σε *SVMs*. Εφόσον το αποτέλεσμα αυτού του τύπου μοντέλων επιβίωσης δεν μπορεί γενικά να ερμηνευτεί ως χρόνος αποτυχίας, θα υποδείξουμε την έκβαση του μοντέλου ως δείκτη πρόγνωσης $u(x)$ αντί της πρόβλεψης του μοντέλου. Για το μοντέλο *Cox* αυτό αντιστοιχεί στο $u\,(x) = w^T\, x$.

Η εξαιρετική απόδοση των Μηχανών Διανυσμάτων Υποστήριξης (*SVMs*) για ταξινόμηση και παλινδρόμηση οδήγησε στο ερώτημα κατά πόσο αυτό το είδος μοντέλων μπορεί να επεκταθεί προς άλλα στατιστικά προβλήματα. Κατά την ανάλυση των δεδομένων επιβίωσης, ενδιαφερόμαστε για το χρονικό διάστημα ανάμεσα σε ένα συγκεκριμένο σημείο εκκίνησης και την εμφάνιση ενός προκαθορισμένου γεγονότος. Μια πρώτη προσέγγιση θα μπορούσε κανείς να σκεφτεί είναι να χρησιμοποιήσει μοντέλα παλινδρόμησης για να μοντελοποιήσει το χρόνο της υποτροπής. Ωστόσο, τα δεδομένα

επιβίωσης περιέχουν συνήθως σημεία δεδομένων με ελλιπείς πληροφορίες, αποκαλούμενα αποκομμένα δεδομένα. Οι δεξιά αποκομμένες παρατηρήσεις είναι παρατηρήσεις για τις οποίες ο ακριβής χρόνος του συμβάντος δεν είναι γνωστός, αλλά είναι γνωστό ένα κατώτερο χρονικό όριο. Μόνο η ενσωμάτωση παρατηρήσεων για τις οποίες είναι γνωστός ο ακριβής χρόνος αποτυχίας θα οδηγούσε σε υποτιμημένους χρόνους επιβίωσης. Οι διάφορες προτάσεις για τον τρόπο υπολογισμού του προβλήματος της επιβίωσης με τη βοήθεια μηχανών διανυσμάτων υποστήριξης αναπτύσσονται στο κεφάλαιο αυτό.

Τα *SVMs* αρχικά είχαν προταθεί για να λύσουν προβλήματα ταξινόμησης (*classification problems*). Στη συνέχεια, αυτά τα μοντέλα επεκτάθηκαν για να είναι εφαρμόσιμα σε προβλήματα ανάλυσης παλινδρόμησης. Η Παλινδρόμηση Διανυσμάτων Υποστήριξης (*Support Vector Regression*, SVR) έχει εφαρμοστεί εκτενώς στη βιβλιογραφία για την πρόβλεψη απόκρισης (*Kazem et al. 2013*, *Fonseca et al 2012*, *Wang et al. 2012, Yijun et al. 2006*). Ωστόσο, υπάρχουν μελέτες που χρησιμοποιούν την *SVR* για ανάλυση επιβίωσης. Αυτό οφείλεται εν μέρει στη μεταβλητή απόκρισης (χρόνος επιβίωσης) στην ανάλυση επιβίωσης, συμπεριλαμβανομένων των αποκομμένων παρατηρήσεων που δεν μπορεί, όμως, να μοντελοποιήσει την παραδοσιακή *SVR*. Ωστόσο, τα επιθυμητά χαρακτηριστικά της *SVR* οδήγησαν τους ερευνητές να την επεκτείνουν ώστε να μπορούν να εφαρμοστούν σε προβλήματα επιβίωσης.

Οι *Yijun et al*. (2006) θεωρούσαν τον χρόνο επιβίωσης ως κατηγορική μεταβλητή και χρησιμοποίησαν το μοντέλο ταξινόμησης διανυσμάτων υποστήριξης για ανάλυση επιβίωσης. Το μοντέλο *SVR* για ανάλυση επιβίωσης προτείνεται από τους *Shivaswamy et al.* (2007). Οι συγγραφείς διερεύνησαν την απόδοση ορισμένων ανταγωνιστικών μοντέλων *SVR* για πραγματικά σύνολα δεδομένων με διαφορετικά ποσοστά αποκοπής. Ο *Ding* (2011) έχει επίσης συζητήσει την πιθανή εφαρμογή του *SVM* στην ανάλυση επιβίωσης. Εφάρμοσε το μοντέλο SVR με διαφορετικούς πυρήνες σε ορισμένα πραγματικά σύνολα δεδομένων. Οι *Khan* και *Zubek* (2008) σύγκριναν το μοντέλο *SVR* με το *Cox* για πέντε πραγματικά σύνολα δεδομένων. Οι *Van Belle et al. (2011c)* και οι *Van Belle et al.* (2010) πρότειναν μια προσέγγιση της *SVR* που χρησιμοποιεί περιορισμούς ταξινόμησης και παλινδρόμησης για δεξιά

αποκομμένα δεδομένα. Οι συγγραφείς σύγκριναν τις επιδόσεις των μοντέλων *SVR* και *Cox* τόσο για κλινικά όσο και για *micro-array* δεδομένα. Επίσης, συζήτησαν ένα τροποποιημένο μοντέλο *SVR* για άλλους τύπους αποκοπής (*Van Belle et al.*, 2011a). Οι *Du και Dua* (2011) εφάρμοσαν *SVR* με δύο μεθόδους επιλογής χαρακτηριστικών, συγκεκριμένα την επιλογή μεμονωμένων χαρακτηριστικών και την επιλογή υποσυνόλου χαρακτηριστικών και συζήτησαν την επίδρασή τους στην απόδοση των *Cox* και *SVR* σε σύνολα δεδομένων για τον καρκίνο του μαστού.

3.3 *SVR* μοντέλα για την Ανάλυση επιβίωσης

3.3.1 Το τυπικό μοντέλο *SVR* για αποκομμένα δεδομένα

3.3.1.1 Περιγραφή των δεδομένων

Στη μάθηση με επίβλεψη δίνεται ένα σύνολο επισημασμένων παρατηρήσεων ως δεδομένα, όπου ένα παράδειγμα αποτελείται από ένα πίνακα δεδομένων (επεξηγηματικές μεταβλητές, χαρακτηριστικά) συν ένα στόχο (μεταβλητή απόκρισης). Ανάλογα με τον τύπο του στόχου λαμβάνουμε διαφορετικά προβλήματα. Έχουμε τις εξής κατηγορίες στόχων:

(i) **Στόχοι Σημείων**: Αυτή είναι η περίπτωση της τυπικής παλινδρόμησης όπου κάθε διάνυσμα $\boldsymbol{x_i} \in \mathbb{R}^m$ έχει ένα στόχο σημείου $y_i \epsilon \mathbb{R}$.

(ii) **Στόχοι Δυαδικής Κλάσης**: Οι Στόχοι Δυαδικής Κλάσης (*Binary Class Labels*) συνήθως σημειώνονται με $y_i \in \{\pm 1\}$ ενώ τα χαρακτηριστικά εξακολουθούν να είναι όπως στην παλινδρόμηση, δηλαδή, $\boldsymbol{x_i} \epsilon \mathbb{R}^m$.

(iii) **Διαστήματα Στόχοι**: Στις περιπτώσεις αυτές έχουμε και το άνω και το κάτω φράγμα στο στόχο. Η τριάδα (x_i, l_i, u_i) με $x_i \epsilon \mathbb{R}^m$, $l_i \epsilon \mathbb{R}$, $u_i \epsilon \mathbb{R}$, όπου τα $l_i < u_i$, υποδηλώνουν ένα στόχο διαστήματος.

(iv) **Χρόνοι Επιβίωσης**: Αυτές είναι οι περιπτώσεις όπου μια μη αποκομμένη παρατήρηση στην ανάλυση επιβίωσης είναι ίδια με

τον στόχο που ορίζεται παραπάνω, ενώ μια σωστά αποκομμένη παρατήρηση γράφεται ως $(x_i, l_i, +\infty)$ του οποίου ο χρόνος επιβίωσης είναι μεγαλύτερος από $l_i \epsilon \mathbb{R}$. Τέλος, για λόγους πληρότητας, οι αριστερά αποκομμένες παρατηρήσεις είναι γραμμένες ως $(x_i, -\infty, u_i)$ των οποίων ο στόχος είναι το πολύ $u_i \epsilon \mathbb{R}$. Οι στόχοι διαστήματος είναι μια γενική περιγραφή των παραπάνω παρατηρήσεων. Ας υποθέσουμε ότι υπάρχει ένα σύνολο δεδομένων $(x_i, l_i, u_i)\ _{i=1}^{n}$ από n παρατηρήσεις με στόχους διαστήματος όπου $l_i < u_i$. Ο στόχος είναι να μάθουμε μια συνάρτηση $f: \mathbb{R}^m \to \mathbb{R}$ έτσι ώστε οι τιμές της συνάρτησης να προσεγγίζουν τις τιμές στόχους. Αυτό που περιγράφουμε παρακάτω είναι τα μέτρα απόδοσης για μάθηση από ένα τέτοιο σύνολο δεδομένων.

3.3.1.2 Μέσο Απόλυτο Σφάλμα

Στην ιδανική περίπτωση, η συνάρτηση παλινδρόμησης $f: \mathbb{R}^m \to \mathbb{R}$ θα πρέπει να δώσει την καλύτερη υπόθεση σχετικά με την τιμή στόχου μιας τιμής x της $f(x)$ αφού μάθει από τα δεδομένα εκπαίδευσης. Για την αξιολόγηση της απόδοσης της τιμής στα διαστήματα, μπορεί να χρησιμοποιηθεί ο ορισμός του Μέσου Απόλυτου Σφάλματος (*Average Absolute Error,* AEE) το οποίο μετρά το απόλυτο σφάλμα εκτός του διαστήματος-στόχου και δίνεται από τον εξής τύπο:

$$AAE = \frac{1}{n}\sum_{i=1}^{n} \max\big(0, l_i - f(x_i)\big) + max(0, f(x_i) - u_i) \qquad (3.3.1)$$

3.3.1.3 Παρουσίαση του τυπικού μοντέλο SVR για αποκομμένα δεδομένα

Ας υποθέσουμε αρχικά ότι έχουμε ένα σύνολο δεδομένων (x_i, l_i, u_i) όπως αυτά περιεγράφηκαν παραπάνω. Σε αυτή την υπόθεση, χρειαζόμαστε το x_i να είναι εντός του διαστήματος (l_i, u_i). Όσο το $f(x_i)$ είναι μεταξύ l_i και u_i δεν υπάρχει κάποια «ποινή» για τα δεδομένα μας. Επιβάλλουμε ποινή εάν το $f(x_i)$

είναι μεγαλύτερο από u_i ή εάν είναι μικρότερο από l_i. Έτσι, η συνάρτηση κόστους για αυτή την περίπτωση γίνεται:

$$c[f(x_i), l_i, u_i] = \max[l_i - f(x_i), f(x_i) - u_i]. \tag{3.3.2}$$

Η απώλεια είναι ακριβώς το απόλυτο σφάλμα που ορίζεται στην (3.2.1). Να σημειωθεί ότι όταν $l_i = -\infty$ ή $u_i = +\infty$, η συνάρτηση απώλειας γίνεται μονόπλευρη. Ας χωρίσουμε τους δείκτες των στόχων μας $\{1,2,\dots n\}$, σε τρία διαφορετικά σύνολα ως εξής:

$$l_u \stackrel{\text{def}}{=} \{i \mid l_i > -\infty, \quad u_i < +\infty\}, \tag{3.3.3}$$

$$l_r \stackrel{\text{def}}{=} \{i \mid u_i = +\infty\}, \tag{3.3.4}$$

$$l_l \stackrel{\text{def}}{=} \{i \mid l_i = -\infty\}. \tag{3.3.5}$$

Να σημειωθεί ότι δεν υπάρχει επικάλυψη μεταξύ του l_r και του l_l, δεδομένου ότι κανένας στόχος δεν γίνεται άπειρος και στις δύο πλευρές. Το σύνολο l_u περιέχει τους δείκτες εκείνων των στόχων, που έχουν τόσο ένα κατώτατο φράγμα όσο και ένα ανώτατο φράγμα, ενώ τα l_r και l_l περιέχουν τους δείκτες των περιπτώσεων που είναι δεξιά και αριστερά αποκομμένες αντίστοιχα. Ορίζουμε επίσης τα σύνολα:

$$L \stackrel{\text{def}}{=} l_u \cup l_r, \tag{3.3.6}$$

$$\text{και } U \stackrel{\text{def}}{=} l_u \cup l_l, \tag{3.3.7}$$

όπου το L περιέχει τους δείκτες εκείνων των περιπτώσεων των οποίων οι στόχοι έχουν ένα κατώτατο φράγμα, ενώ το U περιέχει τους δείκτες εκείνων που έχουν ένα ανώτατο φράγμα. Προκύπτουν, λοιπόν, οι παρακάτω σχέσεις για την παλινδρόμηση με αποκομμένες παρατηρήσεις:

$$\min_{w,b,\xi,\xi*} \frac{1}{2}||w||^2 + c\left(\sum_{i\epsilon U} \xi_i + \sum_{i\epsilon L} \xi_i^*\right), \tag{3.3.8}$$

το οποίο υπακούει σε: $w^T x_i + b - u_i \le \xi_i^*, \forall i \in U,$ (3.3.9)

$$l_i - w^T x_i - b \le \xi_i, \forall i \in L, \tag{3.3.10}$$

όπου $\xi_i \geq 0\ \forall i\ \in U$ και $\xi_i^* \geq 0\ \forall i\ \in L$.

Από τις παραπάνω σχέσεις, μπορούμε να συμπεράνουμε ότι χρησιμοποιούμε όλες τις διαθέσιμες πληροφορίες του συνόλου των δεδομένων. Εισάγοντας έτσι τους πολλαπλασιαστές *Lagrange* $a_i^* \geq 0$ για τις ανισότητες στο $(3.3.9)$ και $a_i \geq 0$ για τις ανισότητες στο $(3.3.10)$, το διπλό της παραπάνω σύνθεσης μπορεί να αποδειχθεί ότι είναι:

$$\min_{\alpha,\alpha^*} \frac{1}{2} \sum_{i,j=1}^{n} (a_i - a_i^*)(a_j - a_j^*)\, k\,(x_i, x_j) - \sum_{i\epsilon L} l_i a_i + \sum_{i\epsilon U} u_i a_i^* \qquad (3.3.11)$$

το οποίο υπακούει σε: $$\sum_{i\epsilon L} a_i - \sum_{i\epsilon U} a_i^* = 0, \qquad (3.3.12)$$

με $a_i\ \geq 0\,, a_i^* \leq c,\ \forall 1 \leq i \leq n$ και $\xi_i\ \geq 0\ \forall\ i\ \in U$, όπου $a_i = 0\ \forall i\ \in L$, $a_i^* = 0\ \forall\ i\ \in U$ είναι ψευδομεταβλητές και $k(x_i, x_j)$= $x_i^T x_j$. Το εσωτερικό γινόμενο $x_i^T x_j$ μπορεί να αντικατασταθεί από μια συνάρτηση πυρήνα (*kernel function*) για να αποκτήσει μια μη γραμμική συνάρτηση απεικόνισης. Στη βέλτιστη λύση, με τις ψευδομεταβλητές, η τιμή συνάρτησης του x παριστάνεται με $f(x)\ = \sum_{i=1}^{n}(a_i - a_i^*)\, k\,(x_i, x) + b$. Να σημειωθεί ότι συνήθως ένα μικρό κλάσμα $\{a_i - a_i^*\}$ είναι μη μηδενικό.

Γενικά, η κλασική περίπτωση εφαρμογής του *SVR* χωρίς αποκομμένα δεδομένα, αποτελούν μόνο ειδικές περιπτώσεις των εξισώσεων $(3.3.8), (3.3.9), (3.3.10)$. Είναι εύκολο να δούμε ότι, δεδομένου ενός συνόλου δεδομένων παλινδρόμησης $(x_i, y_i,)\,_{i=1}^{n}$, μετατρέποντας κάθε δείγμα σε $(x_i, y_i - \varepsilon, y_i + \varepsilon)$ όπου $\varepsilon >\ 0$ προέρχεται από την ε-συνάρτηση κόστους χωρίς ευαισθησία η οποία μειώνεται σε *SVR.*

3.3.2 Το μοντέλο SVCR για την Ανάλυση επιβίωσης

Στο τυπικό *SVR* η πρόβλεψη ορίζεται ως ένας γραμμικός συνδυασμός μετασχηματισμού των μεταβλητών $\varphi\,(x)$, με $\varphi\,(\cdot)$ τη συνάρτηση με τα χαρακτηριστικά, συν μια σταθερά b:

$$y = w^T \varphi(x) + b. \tag{3.3.13}$$

Για να ληφθεί αυτή η εκτίμηση, οι συντελεστές w και η σταθερά b πρέπει να υπολογιστούν. Για να αποκτήσουν καλές ιδιότητες για να γίνει γενίκευση, οι συντελεστές παραμένουν μικροί. Τα *SVMs* διατυπώνονται ως προβλήματα βελτιστοποίησης, όπου μια λειτουργία κόστους θα πρέπει να ελαχιστοποιείται κάτω από ορισμένους περιορισμούς. Οι περιορισμοί περιλαμβάνουν ότι η εκτίμηση για το $\hat{y}$ θα πρέπει να είναι μεγαλύτερη από $y - \varepsilon$, με $\varepsilon > 0$. Επίσης, το $-\hat{y}$ θα πρέπει να είναι μεγαλύτερο από το $-y - \varepsilon^*$, με $\varepsilon^* > 0$. Η συνάρτηση κόστους «τιμωρεί» τις μεγάλες τιμές ε και ε^* έτσι ώστε οι εκτιμήσεις που προκύπτουν για το $\hat{y}$ να είναι κοντά στις παρατηρούμενες τιμές y. Τυπικά, το πρόβλημα διατυπώνεται ως εξής:

$$\min_{w,b,\varepsilon,\varepsilon^*} \frac{1}{2} w^T w \; + \gamma \sum_{i=1}^{n} (\varepsilon_i + \varepsilon_i^*), \tag{3.3.14}$$

το οποίο υπόκειται σε:

$$\begin{cases} w^T \varphi(x_i) \; + \; b \; \geq y_i - \varepsilon_i, & \forall \; i = 1, 2, \dots, n \\ -w^T \varphi(x_i) - \; b \; \geq \; -y_i - \varepsilon_i^* \,, & \forall \; i \; = 1, 2, \dots, n \\ \varepsilon_i \; \geq 0, & \forall \; i = 1, 2, \dots, n \\ \varepsilon_i^* \geq 0, & \forall \; i = 1, 2, \dots, n \end{cases} \tag{3.3.15}$$

Η παράμετρος γ είναι μια αυστηρή θετική σταθερά κανονικοποίησης και ε και ε^* είναι χαλαρές (*slack*) μεταβλητές που επιτρέπουν σφάλματα στις εκτιμήσεις των δεδομένων «εκπαίδευσης». Το εκτιμώμενο αποτέλεσμα για ένα νέο σημείο x^* βρίσκεται στη συνέχεια ως εξής:

$$y(x^*) = \textstyle\sum_{i=1}^{n} (a_i - \alpha_i^*) \varphi(x_i)^T \varphi(x^*) + b, \tag{3.3.16}$$

Όπου a_i και α_i^* είναι οι πολλαπλασιαστές *Lagrange*. Ένα πλεονέκτημα των μοντέλων που βασίζονται σε *SVM*, είναι ότι η συνάρτηση χαρακτηριστικών $\varphi(\cdot)$ δεν χρειάζεται να οριστεί. Σύμφωνα με το θεώρημα *Mercer* (1909), το $\varphi(x_i)^T \varphi(x_j)$ μπορεί να γραφτεί ως

$$k(x_i, x_j) = \; \varphi(x_i)^T \varphi(x_j) \tag{3.3.17}$$

με την προϋπόθεση ότι ο $k\,(\cdot,\cdot)$ είναι ένας θετικός καθορισμένος πυρήνας. Συχνά χρησιμοποιούμενοι πυρήνες είναι οι εξής:

- Ο γραμμικός πυρήνας (*linear kernel*):

$$k(x, z) = \; x^T z \,. \tag{3.3.18}$$

- Ο πολυωνυμικός πυρήνας (*polynomial kernel*) βαθμού a:

$$k(x,z) = (\tau + x^T z)^\alpha \text{ , με } \tau \geq 0. \quad (3.3.19)$$

- Ο ακτινικός πυρήνας (*RBF kernel*):

$$k(x,z) = \exp\left(-\frac{\left||x-z|\right|_2^2}{\sigma^2}\right). \quad (3.3.20)$$

Ένας πυρήνας για κλινικά δεδομένα (*Daemen and De Moor*,2009) επίσης έχει προταθεί ως πρόσθετος πυρήνας $k(x,z) = \sum_{p=1}^{d} k_p(x,z)$, όπου ο $k_p(\cdot,\cdot)$ εξαρτάται από τον τύπο των μεταβλητών p. Για συνεχείς και κανονικές μεταβλητές, ο $k_p(\cdot,\cdot)$ ορίζεται ως:

$$k_p(x_p, z_p) = \frac{c - |x_p - z_p|}{c}, \quad (3.3.21)$$

με x_p την p-οστή συμμεταβλητή της παρατήρησης x, $c = max_p - min_p$, με min_p και max_p η ελάχιστη και μέγιστη τιμή της p-οστής συμμεταβλητής στο σύνολο δεδομένων «εκπαίδευσης». Για τα κατηγορικά και δυαδικά δεδομένα, ο $k_p(\cdot,\cdot)$ ορίζεται ως

$$k_p(x_p, z_p) = \begin{cases} 1, \alpha\nu\ x_p = z_p \\ 0, \alpha\nu\ x_p \neq z_p \end{cases}. \quad (3.3.22)$$

Το πρόβλημα της χρησιμοποίησης του *SVR* για αποκομμένα δεδομένα έγκειται στην αβεβαιότητα σχετικά με τα αποτελέσματα του y. Οι πρώτες προσεγγίσεις σε αποκομμένα δεδομένα χρησιμοποιώντας στρατηγικές παλινδρόμησης είτε παραλείπουν τις αποκομμένες παρατηρήσεις, με αποτέλεσμα υποτιμημένους χρόνους αποτυχίας, είτε αντιμετωπίζουν τις αποκομμένες παρατηρήσεις ως μη γνωστές (*nonevents*), οδηγώντας σε μοντέλα με μεροληψία. Για να χρησιμοποιήσουμε, λοιπόν, όλες τις διαθέσιμες πληροφορίες (οι αποκομμένες παρατηρήσεις δίνουν πληροφορίες για το διάστημα στο οποίο συμβαίνει το συμβάν), ο *Shivaswamy et al.* (2007) πρότεινε μια προσέγγιση *SVR* για αποκομμένα δεδομένα (*SVCR*). Για τις μη αποκομμένες παρατηρήσεις, οι ίδιοι περιορισμοί λαμβάνονται υπόψη όπως στο κλασικό μοντέλο *SVR*. Για τις δεξιά αποκομμένες παρατηρήσεις, είναι γνωστό ότι η αποτυχία δεν συνέβη πριν από τον χρόνο αποκοπής τους. Συνεπώς, ο πρώτος περιορισμός της σχέσης (3.3.15) εξακολουθεί να ισχύει.

Ωστόσο, ο δεύτερος περιορισμός της (3.3.15) είναι υπερβολικά περιοριστικός για τις δεξιά αποκομμένες παρατηρήσεις. Το μοντέλο που προτείνεται από τον *Shivaswamy* είναι το εξής:

SVCR:

$$\min_{w,b,\varepsilon,\varepsilon^*} \frac{1}{2} w^T w + \gamma \sum_{i=1}^{n} (\varepsilon_i + \varepsilon_i^*), \tag{3.3.23}$$

το οποίο υπόκεινται σε:

$$\begin{cases} w^T \varphi(x_i) + b \geq y_i - \varepsilon_i, & \forall\, i = 1, 2, \dots, n \\ -\delta_i (w^T \varphi(x_i) + b) \geq -\delta_i y_i - \varepsilon_i^*, & \forall\, i = 1, 2, \dots, n \\ \varepsilon_i \geq 0, & \forall\, i = 1, 2, \dots, n \\ \varepsilon_i^* \geq 0, & \forall\, i = 1, 2, \dots, n \end{cases} \tag{3.3.24}$$

Στην σχέση (3.3.23), n είναι το μέγεθος του δείγματος και η παράμετρος γ είναι μια θετική σταθερά κανονικοποίησης. Τα ε_i και ε_i^* είναι χαλαρές μεταβλητές και επιτρέπουν τα λάθη στην πρόβλεψη. Οι μεγάλες τιμές των χαλαρών μεταβλητών «τιμωρούνται» από τη συνάρτηση κόστους. Ο δείκτης πρόγνωσης, η πρόβλεψη του μοντέλου, για ένα νέο σημείο x^* υπολογίζεται ως εξής:

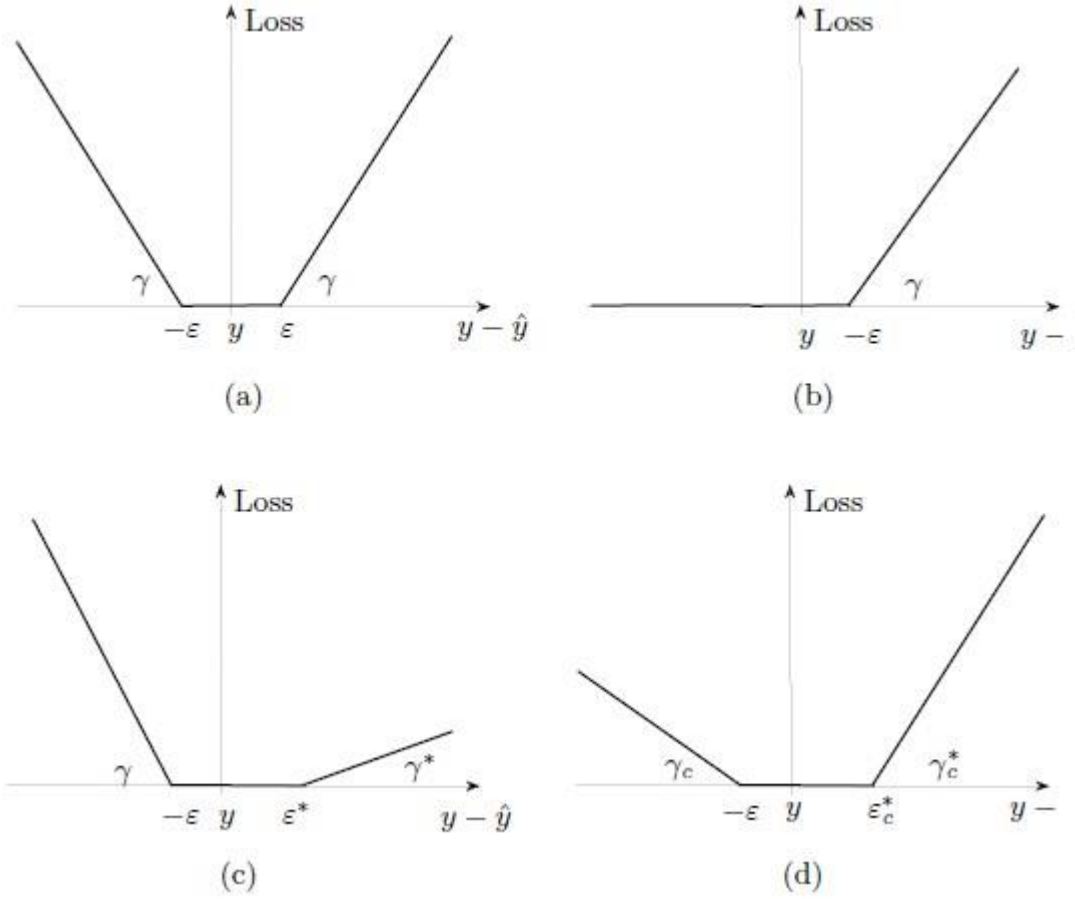
$$u(x^*) = \sum_i (a_i - \delta_i \alpha_i^*) \varphi(x_i)^T \varphi(x^*) + b, \tag{3.3.25}$$

όπου a_i και α_i^* είναι οι πολλαπλασιαστές *Lagrange*.

3.3.3 Το μοντέλο SVRc για την Ανάλυση επιβίωσης

Μια δεύτερη πρόταση για την παλινδρόμηση για τα κολοβά δεδομένα έγινε από τους τον *Khan* και *Zubek (*2008), (*Support Vector Regression for Censored Data*, *SVRc*). Η διαφορά μεταξύ των δύο μοντέλων έγκειται στην εφαρμοζόμενη «ποινή» ή το κόστος εσφαλμένων προγνωστικών δεικτών

(βλέπε σχήμα 3.3), οι οποίες μπορούν να ερμηνευτούν ως προβλεπόμενοι χρόνοι αποτυχίας για προσεγγίσεις παλινδρόμησης.



Σχήμα 3.3 Οι συναρτήσεις κόστους όπως ορίζονται από το *SVCR (Shivaswamy et al., 2007)* στα (a) και (b) και στο *SVRc* (*Khan and Zubek*, 2008) στα (c)και (d).Οι συναρτήσεις κόστους για τα γεγονότα σημειώνονται στα στοιχεία (a) και (c). Η συνάρτηση κόστους για τα δεξιά σημεία αποκοπής δεδομένων απεικονίζονται στα στοιχεία (b) και (d). Και οι δύο μέθοδοι διαφέρουν με τον τρόπο που υπολογίζεται η συνάρτηση κόστους. Ένα σημαντικό μειονέκτημα της μεθόδου *SVRC* είναι η ανάγκη να οριστούν 4 παράμετροι.

Η μέθοδος *SVCR* «τιμωρεί» τις λανθασμένες εκτιμήσεις το ίδιο εάν η εκτίμηση ήταν υψηλότερη ή χαμηλότερη από τον παρατηρούμενο χρόνο αποτυχίας και τιμωρεί τις λανθασμένες εκτιμήσεις για σωστά αποκομμένα δεδομένα μόνο εάν η εκτίμηση είναι χαμηλότερη από τον παρατηρούμενο χρόνο αποκοπής. Ο συλλογισμός είναι ότι είναι γνωστό μόνο ότι ο χρόνος αποτυχίας είναι υψηλότερος από το παρατηρούμενο αποτέλεσμα. Επιπλέον, η «ποινή» για λανθασμένες προβλέψεις είναι η ίδια, ανεξάρτητα από το αν συνίσταται σε μια εκτίμηση για αποκομμένους ή μη, χρόνους αποτυχίας. Αντίθετα, η *SVRc* εφαρμόζει διαφορετικές «ποινές» για τέσσερις πιθανές περιπτώσεις:

(i) ποινή γ για συμβάντα με εκτιμώμενη επιβίωση μικρότερη από την παρατηρούμενη επιβίωση.

(ii) ποινή γ^* για συμβάντα με εκτιμώμενη επιβίωση μεγαλύτερη από την παρατηρούμενη επιβίωση, με $\gamma^* > \gamma$

(iii) ποινή γ_c για σωστά αποκομμένα δεδομένα με εκτιμώμενη επιβίωση μικρότερη από τον παρατηρούμενο χρόνο αποκοπής και

(iv) ποινή γ_c^* για σωστά αποκομμένα δεδομένα με εκτιμώμενη επιβίωση μεγαλύτερη από τον παρατηρούμενο χρόνο αποκοπής, με $\gamma_c^* < \gamma_c$. Επιπλέον, αυτό το μοντέλο παρέχει διαφορετικές ε-«ποινές» για γεγονότα και σωστά αποκομμένα δεδομένα και για εκτιμήσεις υψηλότερες και χαμηλότερες από τον παρατηρούμενο χρόνο. Συνεπώς, το κύριο μειονέκτημα της τελευταίας μεθόδου είναι ο μεγάλος αριθμός υπερπαραμέτρων.

Το βασικό ζήτημα για την εφαρμογή του SVR στην Ανάλυση Επιβίωσης είναι η αδυναμία διαχείρισης των διαφορών μεταξύ των αποκομμένων περιπτώσεων και των συμβάντων. Ο στόχος των τιμών της παλινδρόμησης για τα γεγονότα είναι αρκετά σίγουρος. Ο πραγματικός χρόνος μπορεί να έχει συμβεί λίγο πριν από την καταγεγραμμένη παρατήρηση. Οι αποκομμένες τιμές του στόχου είναι εξαιρετικά αβέβαιες. Η βασική ιδέα του SVRc είναι να ληφθούν υπόψη αυτές οι διαφορές, τροποποιώντας ασύμμετρα την ε-μη ευαισθητοποιημένη συνάρτηση κόστους. Τέσσερις νέες τιμές των δύο παραμέτρων C και ε εισάγονται για αποκομμένες και μη παρατηρήσεις.

Η μέθοδος *SVRc* εισάγει τέσσερεις νέες παραμέτρους γ_n^* και γ_n που αντικαθιστούν το γ, και ε_n^* και ε_n που αντικαθιστούν ε. Το *SVRc* υπολογίζει και την φυσική αριστερή αποκοπή των περιστατικών, δηλαδή των ασθενών που πάσχουν από μια ασθένεια προτού αυτή διαγνωσθεί κατά τη διάρκεια επίσκεψης στο γιατρό.

Η παράμετρος ε_n^* ορίζει το αποδεκτό περιθώριο του σφάλματος εάν η προβλεπόμενη τιμή του μοντέλου είναι μεγαλύτερη από την τιμή του στόχου $(f(x) > y)$. Εάν ισχύει αυτό, η συνάρτηση της ποινής ελέγχεται από το γ_n^*. Η παράμετρος ε_n ορίζει το αποδεκτό περιθώριο σφάλματος η προβλεπόμενη τιμή

του μοντέλου είναι μικρότερη από την τιμή του στόχου $(f(x) < y)$. Εάν ισχύει αυτό, η συνάρτηση της ποινής ελέγχεται από το γ_n. Οι προτεινόμενες σχέσεις μεταξύ αυτών των παραμέτρων είναι $\varepsilon_n > \varepsilon_n^*$ και $\gamma_n < \gamma_n^*$ για να ληφθεί υπόψη και η φυσική αριστερή αποκοπή των γεγονότων. Συνεπώς, εάν το μοντέλο προβλέπει ότι ένα συμβάν εμφανίζεται πριν από την πραγματική τιμή-στόχο, υπάρχει ένα σχετικά μεγαλύτερο περιθώριο σφάλματος και μια μικρότερη ποινή. Ωστόσο, αν δεν είναι επιθυμητό να ληφθεί υπόψη η φυσική αριστερή αποκοπή των γεγονότων, οι παράμετροι μπορούν απλά να επιλεχθούν ακολουθώντας τις σχέσεις $\varepsilon_n = \varepsilon_n^*$ και $\gamma_n^* = \gamma_n$.

Οι αποκομμένες περιπτώσεις αντιμετωπίζονται παρόμοια με τα γεγονότα. Ο αλγόριθμος *SVRc* εισάγει τέσσερεις νέες παραμέτρους γ_c^* και γ_c που αντικαθιστούν το γ, και ε_c^* και ε_c που αντικαθιστούν ε. Ο αλγόριθμος υπολογίζει τη δεξιά αποκομμένη φύση των δειγμάτων, δηλαδή των ασθενών που αντιμετωπίζουν το γεγονός που μας ενδιαφέρει, μετά τον τελευταίο καταγεγραμμένο χρόνο χωρίς ασθένεια. Η παράμετρος ε_c^* ορίζει το αποδεκτό περιθώριο σφάλματος εάν η προβλεπόμενη τιμή του μοντέλου είναι μεγαλύτερη από την τιμή του στόχου $(f(x) > y)$. Εάν ισχύει αυτό, η συνάρτηση της ποινής ελέγχεται από το γ_c^*. Η παράμετρος ε_c ορίζει το αποδεκτό περιθώριο σφάλματος εάν η προβλεπόμενη τιμή του μοντέλου είναι μικρότερη από την τιμή του στόχου $(f(x) < y)$. Εάν ισχύει αυτό, η συνάρτηση της ποινής ελέγχεται από το γ_c. Η προτεινόμενη σχέση μεταξύ αυτών των παραμέτρων είναι $\varepsilon_c < \varepsilon_c^*$ και $\gamma_c > \gamma_c^*$ για να υπολογίσουμε τις δεξιά αποκομμένες περιπτώσεις. Αν η πρόβλεψη ότι το γεγονός συμβαίνει μετά την τιμή στόχου, υπάρχει ένα σχετικά μεγαλύτερο περιθώριο σφάλματος και μικρότερη ποινή.
Ο γενικός αλγόριθμος, ο οποίος καλύπτει τις περιπτώσεις αποκομμένων και μη παρατηρήσεων είναι ο παρακάτω:

$$\min_{w,b,} \frac{1}{2}||w||^2 + \sum_{i=1}^{n} (\gamma_i \xi_i + \gamma_i^* \xi_i^*), \qquad (3.3.26)$$

δοθέντος των περιορισμών:

$$y_i - (w^T \varphi(x_i) + b) \leq \varepsilon_i + \xi_i \,, \forall\, i = 1, 2, \dots, n \tag{3.3.27}$$

$$(w^T \varphi(x_i) + b) - y_i \leq \varepsilon_i^* + \xi_i^* \,, \forall\, i = 1, 2, \dots, n \tag{3.3.28}$$

$$\xi_i^{(*)} \geq 0, \forall\, i = 1, 2, \dots, n \tag{3.3.29}$$

όπου:

Εφόσον υπάρχει αποκοπή έχουμε ότι: $s_i = 1$, (3.3.30)

Και όταν δεν υπάρχει αποκοπή έχουμε ότι: $s_i = 0$, (3.3.31)

Ισχύουν επίσης:

$$\gamma_i^{(*)} = s_i \gamma_c^{(*)} + (1 - s_i)\gamma_n^{(*)} \tag{3.3.32}$$

$$\varepsilon_i^{(*)} = \varepsilon_i \gamma_c^{(*)} + (1 - s_i)\varepsilon_n^{(*)}. \tag{3.3.33}$$

Οι προτεινόμενες σχέσεις μεταξύ των παραμέτρων είναι $\gamma_c^* < \gamma_n < \gamma_n^* = \gamma_c$ και $\varepsilon_c^* > \varepsilon_n > \varepsilon_n^* = \varepsilon_c$. Οι ποινές για τις αποκομμένες προβλέψεις μικρότερες από τον στόχο ή το γεγονός είναι ισοδύναμες και η μεγαλύτερες, καθώς είναι σαφώς εσφαλμένες. Υπάρχει μια μικρή προσαρμογή για μη-αποκομμένες προβλέψεις πριν από το στόχο λόγω της αριστερά αποκομμένης φύσης των γεγονότων και η μέγιστη ανοχή γίνεται για αποκομμένες προβλέψεις μετά το στόχο επειδή μπορεί να είναι σωστές.

Αυτή η ενημέρωση της συνάρτησης κόστους *SVR* για διαφορετικούς εσφαλμένους και δομικούς παραμέτρους κινδύνου και της ασύμμετρης σχέσης μεταξύ τους είναι η βασική πρωτοβουλία του *SVRc.* Ο αλγόριθμος διατηρεί όλα τα πλεονεκτήματα του συμβατικού *SVR*, όπως την απεικόνιση μέσω πυρήνων (όλοι οι έγκυροι πυρήνες SVM παραμένουν χρησιμοποιήσιμοι), τη συγκέντρωση σε σημαντικές περιπτώσεις, τη μειωμένη ευαισθησία σε υπερφόρτωση, την ικανότητα να δουλεύει με περισσότερα χαρακτηριστικά κ.λπ., και μπορεί τώρα να εφαρμοστεί στην ανάλυση επιβίωσης

3.3.4 Το μοντέλο SVR που χρησιμοποιεί τη συνάρτηση MRL

Το μοντέλο *SVR* για την Ανάλυση Επιβίωσης που αναλύσαμε στην ενότητα (3.3.4) χρησιμοποιεί μια απλή συνάρτηση απώλειας για σφάλματα που προκύπτουν από την πρόβλεψη αποκομμένων παρατηρήσεων. Αυτή η συνάρτηση απώλειας «τιμωρεί» το μοντέλο μόνο όταν οι αποκομμένες παρατηρήσεις προβλέπονται μικρότερες από τον χρόνο αποκοπής τους. Προτείνεται λοιπόν, ένα άλλο μοντέλο *SVR* το οποίο χρησιμοποιεί συνάρτηση απώλειας δύο όψεων. Αυτό το μοντέλο υποθέτει ότι ο χρόνος εκδήλωσης για μια αποκομμένη παρατήρηση είναι ίσος με το άθροισμα του χρόνου αποκοπής του και του μέσου υπολειπόμενου χρόνου ζωής (*Mean Residual Lifetime*, MRL). Για τα άτομα ηλικίας x, το *MRL* υπολογίζει την αναμενόμενη διάρκεια ζωής τους (*Klein and Moeschberger*, 2003), χρησιμοποιώντας τον ακόλουθο τύπο:

$$MRL(x) = \frac{\int_x^\infty S(t)dt}{S(x)} \tag{3.3.34}$$

όπου, το $S(x)$ είναι η συνάρτηση επιβίωσης. Μία κλασσική εκτιμήτρια της συνάρτησης επιβίωσης είναι η *Kaplan–Meier*. Ως εκ τούτου, το μοντέλο «τιμωρείται» επίσης όταν προβλέπει ότι οι αποκομμένες παρατηρήσεις είναι μεγαλύτερες από το άθροισμα του χρόνου αποκοπής και του *MRL*. Αυτό το μοντέλο ονομάζεται *SSVR-MRL* και είναι το εξής:

SSVR-MRL:

$$\min_{w,b,\varepsilon,\varepsilon^*,\xi} \frac{1}{2}w^T w \ + \gamma \sum_{i=1}^{n}(\varepsilon_i + \varepsilon_i^*) + \mu \sum_{i=1}^{n} \xi_i, \tag{3.3.35}$$

το οποίο, $\forall\ i = 1, 2, \dots, n$, υπόκεινται σε:

$$\begin{cases} w^T\varphi(x_i) \ + \ b \ \geq y_i - \varepsilon_i, \\ -\delta_i(w^T\varphi(x_i) \ + \ b) \ \geq \ -\delta_i y_i - \varepsilon_i^*, \\ (\delta_i - 1)(w^T\varphi(x_i) \ + \ b) \ \geq \ (\delta_i - 1)(y_i + MRL_i), \\ \varepsilon_i \ \geq 0, \\ \varepsilon_i^* \geq 0, \\ \xi_i \geq 0 \end{cases} \tag{3.3.36}$$

Η πρόβλεψη του μοντέλου, για ένα νέο σημείο x^* υπολογίζεται ως εξής:

$$u(x^*) = \sum_i [a_i - \delta_i \alpha_i^* + (\delta_i - 1)\beta_i]\varphi(x_i)^T \varphi(x^*) + b \quad (3.3.37)$$

Όπου a_i, α_i^* και β_i είναι οι πολλαπλασιαστές *Lagrange*.

3.3.5 Το γραμμικό μοντέλο επιβίωσης-SVR με περιορισμούς θετικότητας

Αυτό το μοντέλο χρησιμοποιείται για την επιλογή χαρακτηριστικών. Σε αυτό το μοντέλο, προστίθεται ένας περιορισμός στο μοντέλο *SVCR* για να διασφαλιστεί η θετικότητα από τα βάρη. Η επιλογή χαρακτηριστικών γίνεται σε αυτό το μοντέλο περιορίζοντας τα βάρη w για να δεχτούν θετικές τιμές. Σε αυτή τη μέθοδο, απαιτείται ένα βήμα επεξεργασίας στο σύνολο δεδομένων πριν από την «εκπαίδευση» του μοντέλου. Ας υποθέσουμε ότι το x^p συμβολίζει το p-οστό χαρακτηριστικό των δεδομένων εισόδου. Η αντιστοιχία μεταξύ κάθε x^p και του χρόνου που θα γίνει το γεγονός υπολογίζεται και κάθε x^p με συμφωνία μικρότερη από 0.5 αλλάζει σε $-xp$. Αυτό το μοντέλο ονομάζεται *SSVRP*. Μετά την τροποποίηση του συνόλου δεδομένων, το μοντέλο εκπαιδεύεται σε σύνολο δεδομένων. Η εκτίμηση προκύπτει από την επίλυση του ακόλουθου προβλήματος βελτιστοποίησης: *SSVRP*:

SSVRP:

$$\min_{w,b,\varepsilon,\varepsilon^*} \frac{1}{2} w^T w + \gamma \sum_{i=1}^{n} (\varepsilon_i + \varepsilon_i^*), \quad (3.3.38)$$

το οποίο υπόκεινται σε:

$$\begin{cases} w^T x_i + b \geq y_i - \varepsilon_i, & \forall\, i = 1, 2, \dots, n \\ -\delta_i (w^T x_i + b) \geq -\delta_i y_i - \varepsilon_i^*, & \forall\, i = 1, 2, \dots, n \\ w_p \geq 0, & \forall\, p = 1, 2, \dots, n \\ \varepsilon_i \geq 0, & \forall\, i = 1, 2, \dots, n \\ \varepsilon_i^* \geq 0, & \forall\, i = 1, 2, \dots, n \end{cases} \quad (3.3.39)$$

Αυτός ο περιορισμός κάνει τα εκτιμώμενα βάρη να είναι κοντά στο μηδέν για τις μη σχετικές μεταβλητές και να είναι υψηλότερες για τις σχετικές μεταβλητές.

3.3.6 Η μέθοδος L1-SVR

Η μέθοδος *L1-SVR* είναι ένα άλλο μοντέλο *SVR* που χρησιμοποιείται για την επιλογή των χαρακτηριστικών. Σε αυτό το μοντέλο, η «ποινή» L1, $\sum_{i=1}^{n} |w_i|$ χρησιμοποιείται αντί του όρου $w^T w$. Αυτό το μοντέλο οδηγεί σε αραιές λύσεις. Επομένως, επιλέγει λιγότερα χαρακτηριστικά από το κλασικό *SVR* όπως αυτό παρουσιάστηκε στην ενότητα (3.3.1.3). Στην παρούσα εργασία, το *L1-SVR* για ανάλυση επιβίωσης ονομάζεται *L1-SSVR*.

3.3.7 Το μοντέλο επιβίωσης-SVR χρησιμοποιώντας περιορισμούς ταξινόμησης (*ranking*) και παλινδρόμησης

Οι *Van Belle et al.* (2011) πρότειναν μια προσέγγιση του *SVR* που χρησιμοποιεί περιορισμούς ταξινόμησης (*ranking*) και παλινδρόμησης. Αυτό το μοντέλο ονομάζεται *SSVR2*. Η τυπική μέθοδος *SVR (SVCR)* περιλαμβάνει μόνο τους περιορισμούς παλινδρόμησης, αλλά το *SSVR2* περιλαμβάνει και τους περιορισμούς παλινδρόμησης και ταξινόμησης. Σε αυτή τη μέθοδο, οι παρατηρήσεις είναι διατεταγμένες ανάλογα με το γεγονός ή τους χρόνους αποκοπής τους.

Στη συνέχεια ορίζονται τα συγκρίσιμα ζεύγη παρατηρήσεων. Ένα ζεύγος δεδομένων ορίζεται ότι είναι συγκρίσιμο όταν είναι γνωστή η σειρά των χρόνων των γεγονότων. Για παράδειγμα, εάν ο χρόνος του ασθενή A είναι αποκομμένος στο χρόνο a και ο χρόνος του ασθενή B δεν έχει αποκοπεί και το γεγονός έχει συμβεί σε χρόνο b με $a < b$, οι δύο αυτοί χρόνοι δεν μπορούν να συγκριθούν καθώς δεν είναι γνωστό ποιο γεγονός συνέβη νωρίτερα. Δεδομένου ότι ο χρόνος συμβάντος για μια αποκομμένη παρατήρηση δεν είναι γνωστός, ένα ζεύγος δεδομένων είναι συγκρίσιμο εάν και οι δύο παρατηρήσεις είναι χωρίς αποκοπή ή εάν μόνο ένας από τους χρόνους να έχει αποκοπεί, με τον χρόνο αποκοπής να είναι μεταγενέστερος από τον χρόνο συμβάντος της μη αποκομμένης παρατήρησης. Η μέθοδος *SVR* με περιορισμούς ταξινόμησης (*ranking*) περιλαμβάνει μια «ποινή» για κάθε συγκρίσιμο ζεύγος παρατηρήσεων για τις οποίες η σειρά στην πρόβλεψη του μοντέλου (δείκτης

πρόβλεψης) διαφέρει από την παρατηρούμενη σειρά. Ο αριθμός των συγκρίσεων μειώνεται συγκρίνοντας κάθε παρατήρηση i με την συγκρίσιμη γειτονική παρατήρηση με τον μεγαλύτερο χρόνο επιβίωσης μικρότερου από y_i, το οποίο θα συμβολίζεται με $y_i(j)$. Αυτό το μοντέλο για αποκομμένα δεδομένα διατυπώνεται ως εξής:

$$\min_{w,b,\varepsilon,\varepsilon^*,\xi} \frac{1}{2} w^T w \ + \gamma \sum_{i=1}^{n} (\varepsilon_i + \varepsilon_i^*) + \mu \sum_{i=1}^{n} \xi_i, \tag{3.3.40}$$

το οποίο,

$\forall\, i = 1, 2, \dots, n,$

υπόκεινται σε:

$$\begin{cases} w^T \varphi(x_i) \ + \ b \ \geq y_i - \varepsilon_i, \\ -\delta_i (w^T \varphi(x_i) \ + \ b) \ \geq \ -\delta_i y_i - \varepsilon_i^*, \\ (\varphi(x_i) - \varphi(x_{j(i)}) \ \geq \ y_i - y_{j(i)} - \xi_i \\ \varepsilon_i \ \geq 0, \\ \varepsilon_i^* \geq 0, \\ \xi_i \geq 0 \end{cases}, \tag{3.3.41}$$

Οι παράμετροι γ και μ σε μοντέλα *SVR* συντονίστηκαν χρησιμοποιώντας το τριπλό κριτήριο διασταυρούμενης επικύρωσης. Διαφορετικά μοντέλα *SVR* χρησιμοποιήθηκαν για ανάλυση επιβίωσης χρησιμοποιώντας τεχνητά και πραγματικά σύνολα δεδομένων.

Η πρόβλεψη του μοντέλου, για ένα νέο σημείο x^* υπολογίζεται ως εξής:

$$u(x^*) = \sum_i (a_i \, [\varphi(x_i) - \varphi(x_{j(i)})]^T + (\beta_i - \delta_i \beta_i^*) \varphi(x_i)^T] \varphi(x^*) + b, \tag{3.3.42}$$

Όπου a_i, α_i^* και β_i είναι οι πολλαπλασιαστές *Lagrange*.

Κεφάλαιο 4– Ταξινόμηση (*ranking*) με Μηχανές Διανυσμάτων Υποστήριξης στην Ανάλυση Επιβίωσης

4.1 Εισαγωγή στα *Ranking SVM*

Το έργο της εκτίμησης ενός μοντέλου μετασχηματισμού βασισμένου σε ένα σύνολο δεδομένων που αντικατοπτρίζει την επιβίωση των ασθενών ενσωματώθηκε σε ένα πλαίσιο μηχανικής μάθησης και *SVMs*. Η βασική ιδέα είναι η αναδιατύπωση του έργου ως προβλήματος κατάταξης μέσω του δείκτη σύγκρισης, ένα πρόβλημα το οποίο μπορεί να επιλυθεί αποτελεσματικά σε ένα πλαίσιο που θα χρησιμοποιεί κατασκευαστικές μεθόδους ελαχιστοποίησης του κινδύνου και κυρτές τεχνικές βελτιστοποίησης. Για να καθοριστεί η επιθυμητή κατάταξη, κάθε παρατήρηση συγκρίνεται με την πλησιέστερη γειτονική. Το κύριο πλεονέκτημα αυτής της προσέγγισης είναι η χρήση μη γραμμικών πυρήνων υλοποιώντας αυτόματα μη παραμετρικά αποτελέσματα των συμμεταβλητών. Ένα μειονέκτημα των μοντέλων κατάταξης είναι ότι η κατάταξη καθορίζεται σε μια αυθαίρετη επιλογή ζευγών παρατηρήσεων. Επομένως, διερευνάται κατά πόσον η απόδοση αυξάνεται όταν διαχωρίζουμε τις παρατηρήσεις με βάση τα χρονικά διαστήματα των χρόνων εκδήλωσης των γεγονότων, μια τεχνική που μπορεί να εφαρμοστεί με τη χρήση εικονικών παρατηρήσεων.

4.2 Μοντέλα *Ranking SVR* (RSVR) για την Ανάλυση Επιβίωσης

4.2.1 Τυπικό μοντέλο *SVM* με βάση περιορισμούς *ranking*

Όταν δεν χρησιμοποιούνται τα μοντέλα παλινδρόμησης, τα προβλήματα στην Ανάλυση Επιβίωσης συχνά μοντελοποιούνται σε προβλήματα ταξινόμησης που απαντούν στο ερώτημα εάν ο ασθενής επιβιώνει σε ένα προκαθορισμένο από πριν διάστημα (π.χ. επιβιώνει 5 έτη μετά τη χειρουργική

επέμβαση). Ωστόσο, για να μπορέσουν να συμπεριληφθούν όσο το δυνατόν περισσότερα γεγονότα, αυτός ο προκαθορισμένος χρόνος πρέπει να ληφθεί πολύ αργά. Αντίθετα, θα θέλαμε μια χρονική στιγμή νωρίς, προκειμένου να διατηρήσουμε όσο περισσότερους ασθενείς, δεδομένου ότι όλοι οι ασθενείς που έχουν αποκοπεί πριν από αυτό το διάστημα θα χαθούν για την ανάλυση.

Ένα δεύτερο πρόβλημα αυτής της προσέγγισης είναι ότι η εγκυρότητα της μεθόδου μειώνεται όταν όλο και περισσότεροι ασθενείς αποκόπτονται νωρίτερα. Για να ξεπεραστούν τα προβλήματα που περιεγράφηκαν παραπάνω, προτάθηκε να διατυπωθεί το πρόβλημα της επιβίωσης ως ένα πρόβλημα ταξινόμησης (*ranking problem*) από *Van Belle et al.* (2007) και τους *Evers* και *Messow* (2008) και προτάθηκε υπολογιστική απλοποίηση από τους *Van Belle et al.* (2008). Η ιδέα πίσω από τη διατύπωση του προβλήματος επιβίωσης ως πρόβλημα ταξινόμησης είναι ότι σε κλινικές εφαρμογές πολλές φορές αυτό που ενδιαφέρει είναι η δημιουργία ομάδων κινδύνου (*risk groups*). Σε αυτή την περίπτωση, δηλαδή, δεν ενδιαφέρει πρωτίστως η εκτίμηση του χρόνου επιβίωσης, αλλά το εάν ο ασθενής έχει υψηλό ή χαμηλό κίνδυνο εμφάνισης του γεγονότος, έτσι ώστε να δοθεί η κατάλληλη θεραπεία. Για να επιτευχθεί αυτός ο στόχος, χρησιμοποιείται μια μέθοδος *Ranking SVM*, παρόμοια στο μοντέλο *ranksvm* για μάθηση κατάταξης ή προτίμησης *Herbrich et al.* (2000). Η μέθοδος που τίθενται από *Van Belle et al.* (2007) και τους *Evers* και *Messow* (2008) περιλαμβάνει κανονικοποίηση ως συνήθως και επιβολή «ποινής» σε κάθε συγκρίσιμο ζεύγος στο οποίο τα σημεία δεδομένων για τα οποία η σειρά στο δείκτη πρόγνωσης διαφέρει από την παρατηρούμενη σειρά. Ένα ζεύγος δεδομένων $\{(x_i, y_i, \delta_i), (x_j, y_j, \delta_j)\}$ είναι συγκρίσιμο όταν είναι γνωστή η χρονική σειρά που συνέβησαν τα γεγονότα (π.χ. δύο γεγονότα, ή ένα γεγονός και ένας δεξιά αποκομμένος χρόνος, όπου ο αποκομμένος αυτός χρόνος είναι αργότερα από τον χρόνο που συνέβη το άλλο γεγονός). Για την ακρίβεια, η συγκρισιμότητα ενός ζεύγους δεδομένων $\{(x_i, y_i, \delta_i), (x_j, y_j, \delta_j)\}$ ορίζεται ως:

$$comp(i,j) = \begin{cases} 1, & \alpha\nu \begin{cases} \delta_i = 1 \; \kappa\alpha\iota \; \delta_j = 1 \\ \delta_i = 1 \; \kappa\alpha\iota \; \delta_j = 0 \; \kappa\alpha\iota \; y_i \leq y_j \end{cases} \\ 0, & \sigma\varepsilon \; \kappa\acute{\alpha}\theta\varepsilon \; \acute{\alpha}\lambda\lambda\eta \; \pi\varepsilon\rho\acute{\iota}\pi\tau\omega\sigma\eta \end{cases} \qquad (4.2.1)$$

Το μοντέλο για αποκομμένα δεδομένα διατυπώνεται ως εξής:

RANKSVMC:

$$\min_{w,\varepsilon} \frac{1}{2} w^T w \; + \gamma \sum_{i=1}^{n} \sum_{\substack{j:\ y_i > y_j, \\ comp(i,j)=1}} \varepsilon_{ij}, \tag{4.2.2}$$

το οποίο υπόκεινται σε:

$$\begin{cases} w^T[\varphi(x_i) - \varphi(x_j)] \geq 1 - \varepsilon_{ij}, \\ \forall\, i = 1, 2, \dots, n, \qquad \forall\, j:\ y_i > y_j \ \kappa\alpha\iota\ comp(i,j) = 1 \\ \varepsilon_{ij} \geq 0, \forall\, i = 1, 2, \dots, n, \forall\, j:\ y_i > y_j \ \kappa\alpha\iota\ comp(i,j) = 1 \end{cases} \tag{4.2.3}$$

Παρατήρηση: Η προσθήκη ενός σταθερού όρου b εδώ δεν είναι χρήσιμη εδώ, αφού μόνο οι διαφορές στον δείκτη πρόγνωσης λαμβάνονται υπόψη. Η πρόβλεψη του μοντέλου για ένα σημείο x^* υπολογίζεται ως εξής:

$$u(x^*) = \sum_{i=1}^{n} \sum_{\substack{j:\ y_i > y_j, \\ comp(i,j)=1}} \alpha_{ij} \left[\varphi(x_i) - \varphi(x_{j(i)})\right]^T \varphi(x^*), \tag{4.2.4}$$

Όπου a_{ij} είναι οι πολλαπλασιαστές *Lagrange.*

Ένα μειονέκτημα της μεθόδου αυτής είναι ότι είναι ένα πρόβλημα τετραγωνικού προγραμματισμού (*Quadratical Programming, QP*) με $O(n^2)$ περιορισμούς το οποίο πρέπει να λυθεί. Για να ξεπεραστεί αυτό το πρόβλημα,

η απλοποιημένη προσέγγιση που επιλύει ένα πρόβλημα QP με $O(n)$ περιορισμούς προτάθηκε στο *Van Belle et al.* (2008). Η μείωση βρέθηκε συγκρίνοντας κάθε σημείο δεδομένων με το πλησιέστερο συγκρίσιμο ζεύγος αντί να συγκρίνεται με όλα τα συγκρίσιμα σημεία δεδομένων. Το μοντέλο αυτό θα αναφέρεται ως *RANKSVMC*.

4.2.2 Πρώτο Μοντέλο SVM με βάση περιορισμούς κατάταξης

Στην παρούσα ενότητα θα περιγραφεί το πρώτο Μοντέλο SVM των *Van Belle et al. (2011c)* με βάση περιορισμούς κατάταξης. Ένα σημαντικό αποτέλεσμα για τον τρόπο που σχετίζουμε τα τυπικά στατιστικά μοντέλα επιβίωσης όπως το μοντέλο της αναλογικής διακινδύνευσης του *Cox* (1972) και το μοντέλο της επιταχυνόμενης διακοπής (*accelerated failure time model)* των

Kalbfleisch και *Prentice* (2002) με μοντέλα επιβίωσης βασισμένα σε *SVM*, βρίσκεται σε μοντέλα μετασχηματισμού των *Van Belle et al.* (2009). Ένα μοντέλο επιβίωσης παρόμοιο με το μοντέλο *RAKNSVMC* με διαφορά που βρίσκεται στη δεξιά πλευρά του πρώτου συνόλου περιορισμών των σχέσεων της (4.2.3), όπου το 1 αντικαθίσταται από το $y_{(i)} - y_{j(i)}$, όπου το $j(i)$ ορίζεται ως:

$$j(i) = arg\max_{j} j \tag{4.2.5}$$

Το οποίο υπόκεινται σε:

$$\begin{cases} comp(i,j) = 1, & (4.2.6) \\ y_j < y_i & (4.2.7) \end{cases}$$

για δεδομένη τιμή i. Στην περίπτωση χωρίς αποκοπή για ένα συγκεκριμένο i έχουμε ότι οι τιμές γίνονται $comp(i,j) = 1,\ \forall\, i = 1, 2, \dots, n$ με $j \neq i$, έτσι ώστε $j(i) = i - 1$. Αυτό το μοντέλο ορίζεται ως εξής:

MODEL 1:

$$\min_{w,\varepsilon} \frac{1}{2} w^T w + \gamma \sum_{i=1}^{n} \varepsilon_i, \tag{4.2.8}$$

Το οποίο υπόκεινται σε:

$$\begin{cases} w^T[\varphi(x_i) + \varphi(x_{j(i)})] \geq y_i - y_{j(i)} - \varepsilon_i, \forall\, i = 1, 2, \dots, n, \\ \varepsilon_i \geq 0, \forall\, i = 1, 2, \dots, n \end{cases} \tag{4.2.9}$$

Διατυπώνοντας τη σχέση *Lagrangian* για τη σχέση (4.2.8) έχουμε τα εξής:

$$\begin{aligned} L(w, \varepsilon, \alpha, \beta) = \frac{1}{2} w^T w + \gamma \sum_{i=1}^{n} \varepsilon_i - \sum_{i=1}^{n} \alpha_i\, w^T(\varphi(x_i) - \varphi(x_{j(i)})) \\ - \sum_{i=1}^{n} a_i(-y_i + y_{j(i)} + \varepsilon_i) - \sum_{i=1}^{n} \beta_i \varepsilon_i, \end{aligned} \tag{4.2.10}$$

και λύνοντας τις εξισώσεις *Karush- Kuhn- Tucker* (*KKT*) (*Boyd και Vandenberghe, 2004*) έχουμε:

$$\begin{cases} \dfrac{dL}{dw} = 0 \Rightarrow w = \sum_{i=1}^{n} \alpha_i \left(\varphi(x_i) - \varphi(x_{j(i)})\right), \\ \dfrac{dL}{d\varepsilon_i} = 0 \Rightarrow \gamma = \alpha_i + \beta_i, \forall\, i = 1, 2, \dots, n \\ \alpha_i \geq 0, \quad \forall\, i = 1, 2, \dots, n \\ \beta_i \geq 0, \quad \forall\, i = 1, 2, \dots, n \\ \sum_{i=1}^{n} \alpha_i \left(w^T(\varphi(x_i) - \varphi(x_{j(i)})) - y_i + y_{j(i)} + \varepsilon_i\right) = 0, \\ \sum_{i=1}^{n} \beta_i \varepsilon_i = 0, \end{cases} \tag{4.2.11}$$

Μετά την εξάλειψη των πρωταρχικών μεταβλητών και του β_i έχουμε:

$$\min_{\alpha_i} \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{n} \alpha_i \alpha_k \left(\varphi(x_i) - \varphi(x_{j(i)})\right)^T \left(\varphi(x_k) - \varphi(x_{l(k)})\right) - \sum_{i=1}^{n} \alpha_i (y_i - y_{j(i)}), \tag{4.2.12}$$

το οποίο υπόκεινται σε $\gamma \leq \alpha_i \leq 0, \forall\, i = 1, 2, \dots, n$

Η πρόβλεψη του μοντέλου, για ένα νέο σημείο x^* υπολογίζεται ως εξής:

$$u(x^*) = \sum_{i=1}^{n} \alpha_i \left[\varphi(x_i) - \varphi(x_{j(i)})\right]^T \varphi(x^*) \tag{4.2.13}$$

Όπου a_i, α_i^* και β_i είναι οι πολλαπλασιαστές *Lagrange*.

4.2.3 Δεύτερο Μοντέλο SVM με βάση περιορισμούς κατάταξης.

Στην παρούσα ενότητα θα περιγραφεί το δεύτερο Μοντέλο SVM των *Van Belle et al. (2011c)* με βάση περιορισμούς κατάταξης. Συγκεκριμένα, παρουσιάζονται εδώ δύο διαφορετικά μοντέλα μετασχηματισμού που χρησιμοποιούν μοντέλα επιβίωσης *SV*. Αρχικά, θα γίνει μείωση της πολυπλοκότητας της κατάταξης με βάση πρόταση των *Van Belle et al.* (2008). Στη συνέχεια, θα διερευνηθεί κατά πόσον η απόδοση του μοντέλου μπορεί να

βελτιωθεί αντικαθιστώντας αυτή την αυθαίρετη επιλογή συγκρίσεων με άγνωστες εικονικές παρατηρήσεις. Θα υποθέσουμε ότι:

$$u = w^T \varphi(x) + b, \tag{4.2.14}$$

εκτός αν ορίζεται διαφορετικά. Στην ανάλυση επιβίωσης το πρωταρχικό ερώτημα που πρέπει να απαντηθεί δεν είναι η εκτίμηση του χρόνου αποτυχίας όσο το δυνατόν ακριβέστερα, αλλά να οριστούν οι κίνδυνοι των παρατηρήσεων. Αυτό ακριβώς συμβαίνει στο μοντέλο του *Cox*: ένας προγνωστικός δείκτης $u(x)$ εκτιμάται και η συνάρτηση που συνδέει τον δείκτη πρόγνωσης και τον πραγματικό χρόνο επιβίωσης παραμένει απροσδιόριστη. Τα μοντέλα που βασίζονται στο πυρήνα για την ανάλυση επιβίωσης επίσης αγνοούν το $h(\cdot)$, αλλά επιπλέον όταν η εκτίμηση του δείκτη πρόγνωσης είναι «ελεύθερη», δημιουργούνται πιο ευέλικτα μοντέλα. Η βασική ιδέα του *SVM* επιβίωσης των *Van Belle et al.* (*2009*), είναι ότι ένας δείκτης πρόγνωσης μπορεί να βρεθεί με την ελαχιστοποίηση του εμπειρικού κινδύνου της λάθος κατάταξης (*misranking*) δύο παρατηρήσεων. Αντί να γίνει προσπάθεια να εκτιμηθεί ο χρόνος επιβίωσης, αναζητείται ο δείκτης πρόγνωσης που εκτιμά τη σειρά με την οποία τα αντικείμενα / άτομα υποτροπιάζουν. Το μοντέλο, λοιπόν, είναι το εξής:

$$\min_{w,\varepsilon} \sum_{i=1}^{n} \varepsilon_i, \tag{4.2.15}$$

Το οποίο υπόκεινται σε:

$$\begin{cases} u(x_i) - u(x_{j(i)}) + \varepsilon_i \geq \rho_i, & \forall\, i = 1, 2, \dots, n, \quad (4.2.16) \\ \varepsilon_i \geq 0, & \forall\, i = 1, 2, \dots, n \quad (4.2.17) \end{cases}$$

όπου ρ_i θετικές σταθερές και $u = w^T \varphi(x)$.

Πρέπει να λυθούν δύο ακόμα προβλήματα:

(i) πώς να ορίσουμε το ρ_i και

(ii) πώς να συμπεριληφθεί η κανονικοποίηση.

Το μοντέλο *RANKSVMC* είναι μια ειδική περίπτωση της παραπάνω διατύπωσης παίρνοντας $\rho_i = 1$ και κάνοντας κανονικοποίηση χρησιμοποιώντας το αξίωμα του μέγιστου περιθωρίου (*Vapnik,* 1998). Οι μέθοδοι που αναφέρονται παρακάτω, αναλύονται περαιτέρω στο έργο του (*Van Belle et al.,* 2009), χρησιμοποιώντας τη σταθερά *Lipschitz* αντί για το μέγιστο περιθώριο για σκοπούς κανονικοποίησης. Με τον τρόπο αυτό λύνεται η

ερώτηση ως προς τον τρόπο καθορισμού της ρ, δεδομένου ότι αυτή η μεθοδολογία παίρνει

$$\rho_i = y_i - y_{j(i)} \tag{4.2.18}$$

Δεδομένου ότι το *MODEL 1* είναι κατασκευασμένο ελαχιστοποιώντας τον εμπειρικό κίνδυνο για λανθασμένες κατατάξεις, η αξία του μοντέλου πρόγνωσης έχει σχετική σημασία. Το μοντέλο πρόγνωσης για το οποίο ο δείκτης πρόγνωσης μπορεί να ερμηνευτεί ως επί, μπορεί να συμπεριλάβει και περιορισμούς παλινδρόμησης όπως και στις περιπτώσεις (*Shivaswamy et al.,* 2007, *Kahn and Zubek*, 2008). Το *MODEL 1* τροποποιείται στο παρακάτω *MODEL 2* ως εξής:

MODEL 2:

$$\min_{w,b,\varepsilon,\xi^*,\xi} \frac{1}{2} w^T w \; + \gamma \sum_{i=1}^{n} \varepsilon_i + \mu \sum_{i=1}^{n} (\xi_i + \xi_i^*), \tag{4.2.19}$$

Το οποίο υπόκεινται σε:

$$\begin{cases} w^T[\varphi(x_i) \; + \varphi(x_{j(i)})] \; \geq y_i - y_{j(i)} - \varepsilon_i, \forall \; i = 1, 2, \dots, n \, , \\ \quad w^T \varphi(x_i) \; + \; b \; \geq y_i - \xi_i \, , \qquad \forall \; i = 1, 2, \dots, n \\ -\delta_i (w^T \varphi(x_i) \; + \; b) \; \geq \; -\delta_i y_i - \xi_i^* \, , \qquad \forall \; i \; = 1, 2, \dots, n \\ \qquad \varepsilon_i \; \geq 0, \qquad \forall \; i = 1, 2, \dots, n \\ \qquad \xi_i^* \geq 0, \qquad \forall \; i = 1, 2, \dots, n \\ \qquad \xi_i \; \geq 0, \qquad \forall \; i = 1, 2, \dots, n \end{cases} \tag{4.2.20}$$

Ο πρώτος περιορισμός είναι ο ίδιος όπως στο *MODEL 1* και είναι ο περιορισμός κατάταξης που βελτιστοποιεί το δείκτη σύγκρισης. Ο δεύτερος περιορισμός απαιτεί ότι ο δείκτης πρόγνωσης θα πρέπει να είναι μεγαλύτερος από τον παρατηρούμενο χρόνο αποτυχίας. Ο τρίτος περιορισμός απαιτεί ο δείκτης πρόγνωσης u να είναι μικρότερος από το y. Ο πολλαπλασιασμός με τη δείκτρια αποκοπής δ_i εξασφαλίζει ότι αυτός ο περιορισμός δεν θα λαμβάνονται υπόψη για τα δεξιά αποκομμένα δεδομένα. Λαμβάνοντας μαζί τον δεύτερο και τον τρίτο περιορισμό, η τιμή του δείκτη πρόγνωσης u στοχεύει στον χρόνο αποτυχίας. Ωστόσο, για τα δεξιά αποκομμένα δεδομένα, ο τρίτος περιορισμός δεν χρειάζεται να ληφθεί υπόψη.

Διατυπώνοντας τη σχέση *Lagrangian* για τη σχέση (4.2.19) έχουμε τα εξής:

$$L(w,\varepsilon,b,\alpha,\beta) = \frac{1}{2}w^T w + \gamma\sum_{i=1}^{n}\varepsilon_i + \mu\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$
$$-\sum_{i=1}^{n}\alpha_i\,(w^T(\varphi(x_i) - \varphi(x_{j(i)})) - y_i + y_{j(i)} + \varepsilon_i)$$
$$-\sum_{i=1}^{n}\beta_i(w^T\varphi(x_i) + b - y_i + \xi_i) \qquad (4.2.21)$$
$$-\sum_{i=1}^{n}\beta_i^*(-\delta_i(w^T\varphi(x_i) + b - y_i) + \xi_i^*)$$
$$-\sum_{i=1}^{n}\eta_i\varepsilon_i - \sum_{i=1}^{n}\nu_i\xi_i - \sum_{i=1}^{n}\nu_i^*\xi_i^*,$$

και λύνοντας τις εξισώσεις *Karush- Kuhn- Tucker* (*KKT*) έχουμε:

$$
\begin{cases}
\dfrac{dL}{dw} = 0 \Rightarrow w = \displaystyle\sum_{i=1}^{n} \alpha_i \left(\varphi(x_i) - \varphi(x_{j(i)})\right) + \sum_{i=1}^{n} (\beta_i - \delta_i \beta_i^*)\varphi(x_i)\,, \\
\dfrac{dL}{db} = 0 \Rightarrow \displaystyle\sum_{i=1}^{n} (-\beta_i + \delta_i \beta_i^*) = 0, \\
\dfrac{dL}{d\varepsilon_i} = 0 \Rightarrow \gamma = \alpha_i + \eta_i, \forall\, i = 1, 2, \dots, n \\
\dfrac{dL}{d\xi_i} = 0 \Rightarrow \mu = \beta_i + \nu_i, \forall\, i = 1, 2, \dots, n \\
\dfrac{dL}{d\xi_i^*} = 0 \Rightarrow \mu = \beta_i^* + \nu_i^*, \forall\, i = 1, 2, \dots, n \\
\alpha_i, \beta_i, \beta_i^*, \eta_i, \xi_i, \xi_i^* \geq 0, \qquad \forall\, i = 1, 2, \dots, n \\
\displaystyle\sum_{i=1}^{n} \alpha_i \left(w^T(\varphi(x_i) - \varphi(x_{j(i)})) - y_i + y_{j(i)} + \varepsilon_i\right) = 0, \\
\displaystyle\sum_{i=1}^{n} \beta_i \varepsilon_i = 0, \\
\displaystyle\sum_{i=1}^{n} \beta_i (w^T \varphi(x_i) + b - y_i + \xi_i) = 0, \\
\displaystyle\sum_{i=1}^{n} \beta_i^* (-\delta_i (w^T \varphi(x_i) + b - y_i) + \xi_i^*) = 0, \\
\displaystyle\sum_{i=1}^{n} \eta_i \varepsilon_i = 0, \\
\displaystyle\sum_{i=1}^{n} \nu_i \xi_i = 0, \\
\displaystyle\sum_{i=1}^{n} \nu_i^* \xi_i^* = 0
\end{cases}
\tag{4.2.22}
$$

Μετά την εξάλειψη όλων των άγνωστων εκτός των α_i, β_i και β_i^* η λύση βρίσκεται:

$$\min_{\alpha_i,\beta,\beta^*} \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{n}\alpha_i\alpha_k\,(\varphi(x_i)-\varphi(x_{j(i)}))^T(\varphi(x_k)-\varphi(x_{l(k)}))$$

$$+\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{n}(\beta_i\beta_k+\delta_i\delta_k\,\beta_i^*\beta_k^*)\varphi(x_i)^T\varphi(x_k)$$

$$-\sum_{i=1}^{n}\alpha_i(y_i-y_{i-1}) \qquad (4.2.23)$$

$$+\sum_{i=1}^{n}\sum_{k=1}^{n}(\beta_i-\delta_k\,\beta_k^*)a_k\varphi(x_i)^T(\varphi(x_k)-\varphi(x_{l(k)}))$$

$$+\sum_{i=1}^{n}\sum_{k=1}^{n}\beta_i\delta_i\,\beta_i^*\varphi(x_i)^T\varphi(x_k)-\sum_{i=1}^{n}\beta_i y_i+\sum_{i=1}^{n}\delta_i\beta_i^* y_i$$

το οποίο υπόκεινται σε:

$$\begin{cases}\gamma \le \alpha_i \le 0, & \forall\, i=1,2,\dots,n\\ \mu \le \beta_i \le 0, & \forall\, i=1,2,\dots,n\\ \mu \le \beta_i^* \le 0, & \forall\, i=1,2,\dots,n\end{cases} \qquad (4.2.24)$$

Η πρόβλεψη του μοντέλου, για ένα νέο σημείο x^* υπολογίζεται ως εξής:

$$u(x^*)=\sum_{i=1}^{n}(\alpha_i\,[\varphi(x_i)-\varphi(x_{j(i)})]^T+(\beta_i-\delta_i\beta_i^*)\,\varphi(x_i)^T)\varphi(x^*)+b \qquad (4.2.25)$$

Όπου a_i, β_i και β_i^* είναι οι πολλαπλασιαστές *Lagrange*.

4.2.4 Τρίτο Μοντέλο SVM με βάση περιορισμούς κατάταξης.

Στην παρούσα ενότητα θα περιγραφεί το τρίτο Μοντέλο *SVM* των *Van Belle et al. (2011c)* με βάση περιορισμούς κατάταξης. Συγκεκριμένα, τόσο το *MODEL 1* όσο και το *MODEL 2* περιορίζουν τον αριθμό των συγκρίσεων ή τους περιορισμούς κατάταξης σε $O(n)$ συγκρίνοντας κάθε παρατήρηση με το πλησιέστερο (σε παρατηρούμενο χρόνο) συγκρίσιμο σημείο. Δεδομένου ότι η επιλογή αυτή είναι μάλλον αυθαίρετη, τίθεται το ερώτημα εάν οι επιδόσεις του μοντέλου μπορούν να βελτιωθούν με τη δημιουργία k εικονικών παρατηρήσεων με άγνωστη χρησιμότητα v_l (επίσης ονομαζόμενες τιμές κατωφλίου) και γνωστό χρόνο αποτυχίας z_l, με $l = 1,\dots,k$. Οι τιμές z κατανέμονται ομοιόμορφα στο παρατηρούμενο χρονικό διάστημα. Η ελάχιστη τιμή του z ισούται με το μηδέν, καθώς αυτό είναι το χαμηλότερο δυνατό αποτέλεσμα. Η μέγιστη τιμή του

z είναι μεγαλύτερη από το μεγαλύτερο παρατηρούμενο χρόνο. Όλοι οι χρόνοι επιβίωσης και αποκοπής βρίσκονται, λοιπόν, εντός του διαστήματος (z_1, z_k). Η συμπερίληψη των κατωφλιών είναι ένας κομψός τρόπος για να γίνει μια μη τυχαία επιλογή συγκρίσεων. Ας υποθέσουμε ότι τα δεδομένα μας δεν είναι αποκομμένα. Στο *MODEL 2* θα γίνει σύγκριση μεταξύ κάθε i παρατήρησης και της παρατήρησης με τον μεγαλύτερο χρόνο αποτυχίας για την οποία ο χρόνος αποτυχίας είναι μικρότερος από το y_i. Στο *MODEL 3*, αυτές οι συγκρίσεις γίνονται κάνοντας λιγότερο αυθαίρετες, συγκρίνοντας κάθε παρατήρηση με μια εικονική παρατήρηση.

Η τροποποίηση του *MODEL 2* με τη συμπερίληψη αυτών των κατωφλιών οδηγεί στο εξής μοντέλο:

MODEL 3:

$$\min_{w,b,\varepsilon,\varepsilon^*,\xi,\xi^*} \frac{1}{2} w^T w \ + \gamma \sum_{i=1}^{n} (\varepsilon_i + \varepsilon_i^*) + \mu \sum_{i=1}^{n} (\xi_i + \xi_i^*), \qquad (4.2.26)$$

Το οποίο υπόκεινται σε:

$$\begin{cases} w^T \varphi(x_i) + b - l_i^T v \geq y_i - l_i^T z - \varepsilon_i, & \forall\, i = 1, 2, \dots, n \\ -\delta_i [w^T \varphi(x_i) \ + \ b - r_i^T (v - z) - y_i] \geq -\varepsilon_i^*, \forall\, i \ = 1, 2, \dots, n \\ w^T \varphi(x_i) \ + \ b \ \geq y_i - \xi_i \,, \forall\, i = 1, 2, \dots, n \\ -\delta_i (w^T \varphi(x_i) \ + \ b) \ \geq \ -\delta_i y_i - \xi_i^* \,, \forall\, i \ = 1, 2, \dots, n \\ \varepsilon_i \ \geq 0, \qquad \forall\, i = 1, 2, \dots, n \\ \varepsilon_i^* \geq 0, \qquad \forall\, i = 1, 2, \dots, n \\ \xi_i \ \geq 0, \qquad \forall\, i = 1, 2, \dots, n \\ \xi_i^* \geq 0, \qquad \forall\, i \ = 1, 2, \dots, n \\ v_l - v_{l-1} \geq 0, \ \ \forall\, l \ = 1, 2, \dots, k \end{cases} \qquad (4.2.27)$$

Στις παραπάνω σχέσεις τα l_i και r_i παριστάνουν διανύσματα με μήκος k τα οποία περιέχουν όλες τις μηδενικές τιμές, εκτός από την τιμή $\bar{l} = arg \max_l l$, με την προϋπόθεση ότι $z_l < y_i$ για τα οποία το $l_i(\bar{l}) = 1$ και $\bar{l} = arg \min_l l$ με την προϋπόθεση ότι $z_l > y_i$ για τα οποία $r_i(\bar{l}) = 1$. Ο πρώτος και ο δεύτερος περιορισμός από τη σχέση $(4.2.27)$ είναι παρόμοιοι με τους αντίστοιχους περιορισμούς της σχέσης $(4.2.20)$, αντικαθιστώντας τις αυθαίρετες συγκρίσεις με την πλησιέστερη γειτονική παρατήρηση συγκρίνοντας με τα πλησιέστερα κατώφλια. Ο τελευταίος περιορισμός της σχέσης $(4.2.27)$ εξασφαλίζει ότι τα κατώφλια αυξάνονται.

Διατυπώνοντας τη σχέση *Lagrangian* για τη σχέση (4.2.26) έχουμε τα εξής:

$$
\begin{aligned}
&L(w,\varepsilon,\varepsilon^*,\xi,\xi^*,v,b;\alpha,\alpha^*,\beta,\beta^*,\nu,\nu^*,\eta,\eta^*,\theta) = \\
&\quad \frac{1}{2}w^T w + \gamma\sum_{i=1}^{n}(\varepsilon_i+\varepsilon_i^*) + \mu\sum_{i=1}^{n}(\xi_i+\xi_i^*) \\
&\quad -\sum_{i=1}^{n}\alpha_i\,(w^T\varphi(x_i)+b-l_i^T(v-z)-y_i+\varepsilon_i) \\
&\quad -\sum_{i=1}^{n}\alpha_i^*\,(-\delta_i[w^T\varphi(x_i)\;+\;b-r_i^T(v-z)-y_i)+\varepsilon_i^* \\
&\quad -\sum_{i=1}^{n}\beta_i(w^T\varphi(x_i)\;+\;b-y_i+\xi_i) \\
&\quad -\sum_{i=1}^{n}\beta_i^*(-\delta_i(w^T\varphi(x_i)\;+\;b-y_i)+\xi_i^*) \\
&-\sum_{i=1}^{n}\theta_l(v_l-v_{l-1}) - \sum_{i=1}^{n}\eta_i\varepsilon_i - \sum_{i=1}^{n}\eta_i^*\varepsilon_i^* - \sum_{i=1}^{n}\nu_i\xi_i - \sum_{i=1}^{n}\nu_i^*\xi_i^*.
\end{aligned}
\tag{4.2.28}
$$

και λύνοντας τις εξισώσεις *Karush- Kuhn- Tucker* (*KKT*) έχουμε:

$$
\begin{cases}
\dfrac{dL}{dw} = 0 \Rightarrow w = \sum_{i=1}^{n} [\alpha_i + \beta_i - \beta_i(\alpha_i^* + \beta_i^*)]\varphi(x_i)\,, \\
\dfrac{dL}{db} = 0 \Rightarrow \sum_{i=1}^{n} [\alpha_i + \beta_i - \beta_i(\alpha_i^* + \beta_i^*)] = 0, \\
\dfrac{dL}{d\varepsilon_i} = 0 \Rightarrow \gamma = \alpha_i + \eta_i, \forall\, i = 1, 2, \dots, n \\
\dfrac{dL}{d\varepsilon_i^*} = 0 \Rightarrow \gamma = \alpha_i^* + \eta_i^*, \forall\, i = 1, 2, \dots, n \\
\dfrac{dL}{d\xi_i} = 0 \Rightarrow \mu = \beta_i + \nu_i, \forall\, i = 1, 2, \dots, n \\
\dfrac{dL}{d\xi_i^*} = 0 \Rightarrow \mu = \beta_i^* + \nu_i^*, \forall\, i = 1, 2, \dots, n \\
\dfrac{dL}{dv_l} = 0 \Rightarrow \sum_{i=1}^{n} (\alpha_i l_{i,l} - \delta_i \alpha_i^* r_{i,l} - \theta_l + \theta_{l-1}) = 0, \forall\, l = 1, 2, \dots, k \\
\alpha_i, \alpha_i^*, \beta_i, \beta_i^*, \eta_i, \eta_i^*\, \xi_i, \xi_i^* \geq 0, \qquad \forall\, i = 1, 2, \dots, n \\
\theta_l \geq 0, \forall\, l = 1, 2, \dots, k \\
\sum_{i=1}^{n} \alpha_i\, (w^T \varphi(x_i) + b - l_i^T (v - z) - y_i + \varepsilon_i) = 0, \\
\sum_{i=1}^{n} \alpha_i^*\, (-\delta_i [w^T \varphi(x_i)\ +\ b - r_i^T (v - z) - y_i) + \varepsilon_i^* = 0 \\
\sum_{i=1}^{n} \beta_i (w^T \varphi(x_i)\ +\ b - y_i + \xi_i)\ = 0, \\
\sum_{i=1}^{n} \beta_i^* (-\delta_i w^T \varphi(x_i)\ +\ b - y_i + \xi_i^*) = 0, \\
\sum_{i=1}^{n} \eta_i \varepsilon_i = 0, \\
\sum_{i=1}^{n} \eta_i^* \varepsilon_i^* = 0\,, \\
\sum_{i=1}^{n} \nu_i \xi_i = 0, \\
\sum_{i=1}^{n} \nu_i^* \xi_i^* = 0\,, \\
\sum_{i=1}^{n} \theta_l (v_l - v_{l-1}) = 0,
\end{cases}
\tag{4.2.29}
$$

με $l_{i,l}$ και $r_{i,l}$ το l-οστό στοιχείο των διανυσμάτων l_i και r_i αντιστοίχως. Μετά εξάλειψη ορισμένων μεταβλητών, η λύση βρίσκεται ως εξής:

$$\min_{a,a^*,\beta,\beta^*,\theta} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i\alpha_j + \beta_i\beta_j)\varphi(x_i)^T\varphi(x_j)$$
$$+\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i\delta_j(a_i^*a_j^* + \beta_i^*\beta_j^*)\varphi(x_i)^T\varphi(x_j)$$
$$+\sum_{i=1}^{n}\sum_{j=1}^{n}(a_i\beta_j - \delta_j\left((a_i+\beta_i)(\alpha_j^*+\beta_j^*)\right))\varphi(x_i)^T\varphi(x_j) \qquad (4.2.30)$$
$$-\sum_{i=1}^{n}\sum_{k=1}^{n}\delta_i\,\alpha_i^*\,\beta_j^*\varphi(x_i)^T\varphi(x_j) - \sum_{i=1}^{n}(\alpha_i - \delta_i\alpha_i^*)y_i$$
$$+\sum_{i=1}^{n}(l_i - r_i)^T z\; - \sum_{i=1}^{n}\beta_i y_i + \sum_{i=1}^{n}\delta_i\beta_i^* y_i$$

Το οποίο υπόκεινται σε:

$$\begin{cases} \gamma \le \alpha_i \le 0, & \forall\, i = 1,2,\dots,n, \\ \gamma \le \alpha_i^* \le 0, & \forall\, i = 1,2,\dots,n, \\ \mu \le \beta_i \le 0, & \forall\, i = 1,2,\dots,n\,, \\ \mu \le \beta_i^* \le 0, & \forall\, i = 1,2,\dots,n \\ \sum_{i=1}^{n}(\alpha_i l_{i,l} - \delta_i\alpha_i^* r_{i,l}) - \theta_l + \theta_{l-1}\,, \forall\, l = 1,2,\dots,k \end{cases} \qquad (4.2.31)$$

Η πρόβλεψη του μοντέλου, για ένα νέο σημείο x^* υπολογίζεται ως εξής:

$$u(x^*) = \sum_{i=1}^{n}(a_i + \beta_i - \delta_i(\alpha_i^* + \beta_i^*))\varphi(x_i)^T\varphi(x^*) + b \qquad (4.2.32)$$

Όπου a_i, β_i,α_i^* και β_i^* είναι οι πολλαπλασιαστές *Lagrange*.

Κεφάλαιο 5- Παραδείγματα και Εφαρμογές των Μηχανών Διανυσμάτων Υποστήριξης στην Ανάλυση Επιβίωσης

5.1 Παρουσίαση του κώδικα της R σε πραγματικά δεδομένα

5.1.1 Περιγραφή των δεδομένων

Στο παρόν κεφάλαιο θα χρησιμοποιήσουμε τα δεδομένα από το *Veteran's Administration Study (Kalbfleisch and Prentice, 2002)* για την εφαρμογή μοντέλων που παρουσιάστηκαν στα παραπάνω κεφάλαια. Τα δεδομένα αφορούν μια έρευνα σε άνδρες με προχωρημένο, μη εγχειρίσιμο καρκίνο του πνεύμονα. Σε αυτή την κλινική δοκιμή τα δεδομένα ήταν τυχαιοποιημένα είτε σε κανονική (*standard*) είτε σε δοκιμαστική (*test*) χημειοθεραπεία. Το κύριο σημείο για τη λήξη της χημειοθεραπείας ήταν ο χρόνος θανάτου. Μόνο 9 από τους 137 χρόνους επιβίωσης είναι αποκομμένοι. Όπως συμβαίνει συνήθως σε τέτοιου είδους μελέτες, υπάρχει μεγάλη ανομοιογένεια μεταξύ των ασθενών, σε θέματα που αφορούν για παράδειγμα την έκταση και την παθολογία των ασθενειών, την προηγούμενη θεραπεία που είχε εφαρμοστεί για τη νόσο, τα δημογραφικά χαρακτηριστικά και την αρχική κατάσταση υγείας. Τα δεδομένα που χρησιμοποιήθηκαν, περιέχουν μεταβλητές οι οποίες περιλαμβάνουν πληροφορίες οι οποίες μετράνε μερικές πτυχές αυτής της ανομοιογένειας των χαρακτηριστικών των ασθενών:

1. *Karno*: Αρχικά ένα μέτρο τυχαιοποίησης, είναι μία μεταβλητή η οποία αναφέρεται στην κατάσταση απόδοσης του ασθενούς, είναι η κλίμακα Καρνόφσκι (*Karnofsky rating*), όπου τιμές 10-30 παίρνουν οι ασθενείς οι οποίοι ήταν συνεχώς νοσηλευόμενοι, 40-60 αυτοί που ήταν με μερικό περιορισμό και 70-90 αυτοί οι οποίοι ήταν ικανοί να προσέχουν οι ίδιοι τον εαυτό τους.
2. *Diagtime*: Μία άλλη μεταβλητή αναφέρεται στον χρόνο (σε μήνες) από τη διάγνωση μέχρι την τυχαιοποίηση.
3. *Age*: Η ηλικία (σε χρόνια).
4. *Prior*: Η ύπαρξη προηγούμενης θεραπείας (0 = όχι, 10 = ναι).

5. *Celltype*: Ο ιστολογικός τύπος του όγκου: πλακώδης (*squamous*), μικρά κύτταρα (*small cell*), αδενοκύτταρα (*adeno*), μεγάλα κύτταρα (*large cell*).
6. *Status*: Θεραπεία (0 = τυπική, 1 = δοκιμαστική).

5.1.2 Εφαρμογή των Μοντέλων της *R*

Στο παρόν κεφάλαιο θα γίνει μια αναλυτική περιγραφή του κώδικα της R στην εφαρμογή των μοντέλων του πακέτου *survivalsvm* της *R* και συγκεκριμένα του μοντέλου *Regression* που αντιστοιχεί στο *SVCR* (3.3.2) της παρούσας εργασίας, του *Van Belle1* που αντιστοιχεί στο *RANKSVMC* (4.2.1), του *VanBelle2* που αντιστοιχεί στο *MODEL1* (4.2.2), και του *Hybrid* το οποίο αντιστοιχεί στο *MODEL2* (4.2.3) που παρουσιάστηκαν στα παραπάνω κεφάλαια, με βάση και τον κώδικα που παρουσιάζεται από τους *Fouodo et al*., (2018).

Η συνάρτηση *survivalsvm* προσαρμόζει ένα νέο μοντέλο επιβίωσης και οι τάξεις κινδύνου προβλέπονται με τη χρήση της γενικής λειτουργίας της *R*. Το κοινό και στα τέσσερα μοντέλα που θα χρησιμοποιηθούν είναι ότι επιλύεται ένα τετραγωνικό πρόβλημα βελτιστοποίησης όταν πηγαίνουμε στον δισδιάστατο χώρο. Στη συνάρτηση *survivalsvm*, επιλύουμε αυτό το πρόβλημα βελτιστοποίησης χρησιμοποιώντας δύο τετραγωνικούς προγραμματιστές: το *ipop* από το πακέτο *kernlab* (*Karatzoglou et al*., 2004) και το *pracma* από το πακέτο *pracma* (*Borchers*, 2016). Η συνάρτηση *pracma* περικλείει το *quadprog* από το πακέτο *quadprog* (*Berwin A*., 2013), το οποίο εφαρμόζεται στη *Fortran* για την επίλυση των τετραγωνικών προβλημάτων όπως περιγράφεται από τους *Goldfarb* και *Idnani* (1982). Επομένως, ο πίνακας πυρήνα θεωρείται θετικά ημι-ορισμένος. Εάν δεν συμβαίνει αυτό, η συνάρτηση *nearPD* από το πακέτο *Matrix* (*Bates and Maechler*, 2016) χρησιμοποιείται για να προσαρμόσει τον πίνακα πυρήνα στον πλησιέστερο θετικά ημι-ορισμένο πίνακα. Σε αντίθεση με το *quadprog*, το *ipop* είναι γραμμένο σε καθαρή *R*. Επομένως, ο χρόνος εκτέλεσης του *ipop* αναμένεται να είναι μεγαλύτερος από αυτόν του *quadprog* για την επίλυση του ιδίου προβλήματος βελτιστοποίησης. Ωστόσο, ένα πλεονέκτημα του *ipop* είναι ότι ο πίνακας πυρήνα δεν χρειάζεται να τροποποιηθεί κατά την

επίλυση του προβλήματος βελτιστοποίησης. Καθώς είναι στην ευχέρεια του χρήστη της εντολής *survivalsvm* να επιλεχθεί ο τρόπος επίλυσης του τετραγωνικού προβλήματος, στην παρούσα εργασία θα χρησιμοποιηθεί η εντολή *quadprog*.

Τα μοντέλα *Vanbelle1* και *Vanbelle2* και το *Hybrid* υπολογίζουν τις διαφορές μεταξύ συγκρίσιμων σημείων δεδομένων. Έχουμε διαθέσιμες τρεις επιλογές για τον προσδιορισμό των συγκρίσιμων ζευγών μέσω του *survivalsvm*: το *makediff1* το οποίο δεν υποθέτει ότι το πρώτο σημείο δεδομένων είναι μη αποκομμένο, το *makediff2* το οποίο υπολογίζει τις διαφορές μόνο σε μη αποκομμένα δεδομένα και το *makediff3* το οποίο χρησιμοποιεί τον ορισμό που περιεγράφηκε στα κεφάλαια που δόθηκαν τα μοντέλα και είναι αυτό που θα χρησιμοποιήσουμε.

Το πακέτο R *survivalsvm* επιτρέπει επίσης στον χρήστη να επιλέξει έναν από τους τέσσερις πυρήνες: το γραμμικό, τον πρόσθετο (*Daemen και De Moor*, 2009),την ακτινική βάση και τον πολυωνυμικό, με την ονομασία *lin_kernel*, *add_kernel*, *rbf_kernel* και *poly_kernel*, αντίστοιχα. Οι πυρήνες μπαίνουν στο μοντέλο μέσω της εντολής *kernel* του *survivalsvm*. Στην εργασία αυτή, θα γίνει εφαρμογή των τεσσάρων μοντέλων που αναφέρθηκαν αρχικά με τη χρήση των πυρήνων *lin_kernel*, *add_kernel* και *rbf_kernel* για κάθε μοντέλο αντίστοιχα.

Το μοντέλο της προσέγγισης *Regression* της *R*, βασίζεται στο *SV Regression* (*Vapnik*, 1998) και στοχεύει στην εύρεση μιας συνάρτησης που εκτιμάει τους χρόνους επιβίωσης ως συνεχείς τιμές απόκρισης y_i με τη χρήση των συμμεταβλητών x_i. Μια πολύ απλή παραλλαγή αυτής της προσέγγισης είναι να αγνοήσουμε όλες τις αποκομμένες παρατηρήσεις και να λύσουμε το πρόβλημα ως ένα απλό *SVR*. Σαφώς όμως, μια τέτοια διατύπωση συνεπάγεται απώλεια πληροφοριών, δεδομένου ότι δεν λαμβάνει υπόψη την ιδιαιτερότητα των δεδομένων επιβίωσης. Οι *Shivaswamy et al.* (2007) όπως παρουσιάσαμε αναλυτικά και στο τρίτο κεφάλαιο, βελτίωσαν το μοντέλο *SVR* συμπεριλαμβάνοντας την αποκοπή των δεδομένων. Για τις αποκομμένες παρατηρήσεις, ο χρόνος έως το συμβάν μετά την αποκοπή είναι άγνωστος και επομένως προβλέψεις με μεγαλύτερο από τον χρόνο αποκοπής δεν πρέπει να τιμωρούνται. Ωστόσο, οι προβλέψεις επιβίωσης οι οποίες είναι χαμηλότερες

από ότι ο χρόνος αποκοπής τιμωρούνται ως συνήθως. Στα μη αποκομμένα δεδομένα, οι ακριβείς χρόνοι επιβίωσης είναι γνωστοί και όπως στο πρότυπο SVR, σε όλες τις προβλέψεις επιβίωσης που είναι μικρότερες ή μεγαλύτερες από την παρατηρούμενη επιβίωση επιβάλλονται ποινές. Συνεπώς το πρώτο μοντέλο που εφαρμόζουμε είναι το *SVCR* έχοντας τον παρακάτω αλγόριθμο:

Η συνάρτηση *surv* από το πακέτο *survival* χρησιμεύει για να κατασκευάσουμε ουσιαστικά το αντικείμενο survival.

```
library(survivalsvm)
> library(survival)
> data(veteran, package = "survival")
```

Στη συνέχεια χωρίζουμε τα δεδομένα μας σε ένα εκπαιδευόμενο (*training*) και ένα δοκιμαστικό (*test*) σύνολο δεδομένων, με τις εξής εντολές:

```
set.seed(123)
> n <- nrow(veteran)
> train.index <- sample(1:n, 0.7 * n, replace = FALSE)
> test.index <- setdiff(1:n, train.index)
```

Έπειτα καλούμε το μοντέλο παλινδρόμησης επιβίωσης που επιθυμούμε, το οποίο για εμάς περιλαμβάνει τις συμμεταβλητές diagtime, status με τις οποίες θα γίνει η ανάλυση παλινδρόμησης. Η παράμετρος κανονικοποίησης δίνεται μέσω της συνάρτησης *gamma.mu*. Για τα μοντέλα *Regression* (*SVCR (3.3.2))*, *Van Belle1* (*RANKSVMC (4.2.1))* και *Van Belle2 (MODEL1 (4.2.2))* χρειάζεται μόνο μία παράμετρος, ενώ δύο παράμετροι είναι αναγκαίοι για το μοντέλο *Hybrid* (*MODEL2 (4.2.3))*.

```
> survsvm.reg <- survivalsvm(Surv(diagtime, status) ~ ., subset =
train.index, data = veteran,type = "regression", gamma.mu = 1,
opt.meth = "quadprog", kernel = "add_kernel")
> print(survsvm.reg)

survivalsvm result

Call:

 survivalsvm(Surv(diagtime, status) ~ ., subset = train.index, data
= veteran, type = "regression", gamma.mu = 1, opt.meth =
"quadprog", kernel = "add_kernel")

Survival svm approach              : regression
```

```
Type of Kernel                        : add_kernel
Optimization solver used              : quadprog
Number of support vectors retained : 39
survivalsvm version                   : 0.0.5
```

Τώρα μπορούμε να κάνουμε τις προβλέψεις για τις παρατηρήσεις που μας δίνονται μέσω του *test.index*:

```
> pred.survsvm.reg <- predict(object = survsvm.reg,newdata =
veteran, subset = test.index)
```

Χρησιμοποιούμε τέλος την εντολή *print* για να εμφανίσουμε τα αποτελέσματά μας.

```
> print(pred.survsvm.reg)

survivalsvm prediction

Type of survivalsvm                    : regression
Type of kernel                         : add_kernel
Optimization solver used in model      : quadprog
predicted risk ranks                     : 13.89 14.95 11.12 15.6
10.7 ...
survivalsvm version                    : 0.0.5
```

Ακριβώς με την ίδια λογική που έγινε η παραπάνω εφαρμογή των μοντέλων της R θα γίνει και η εφαρμογή των υπολοίπων μοντέλων αλλάζοντας τα *type* και *kernels*.

Για παράδειγμα δίνεται παρακάτω ο κώδικας που εφαρμόζει το μοντέλο *SVCR* παίρνοντας το *linear kernel* στα δεδομένα μας:

```
> ## efarmogh tou svcr(regression) me lin_kernel
> survsvm.reg_lin <- survivalsvm(Surv(diagtime, status) ~ ., subset
= train.index, data = veteran,type = "regression", gamma.mu = 1,
opt.meth = "quadprog", kernel = "lin_kernel")
> print(survsvm.reg_lin)

survivalsvm result

Call:
```

```
 survivalsvm(Surv(diagtime, status) ~ ., subset = train.index, data
= veteran, type = "regression", gamma.mu = 1, opt.meth = "quadprog",
kernel = "lin_kernel")

Survival svm approach              : regression
Type of Kernel                     : lin_kernel
Optimization solver used           : quadprog
Number of support vectors retained : 5
survivalsvm version                : 0.0.5
> pred.survsvm.reg_lin <- predict(object = survsvm.reg_lin, newdata
= veteran, subset = test.index)
> print(pred.survsvm.reg_lin)


survivalsvm prediction

Type of survivalsvm                    : regression
Type of kernel                         : lin_kernel
Optimization solver used in model      : quadprog
predicted risk ranks                   : 70.17 -8.53 -18.27 -36.44
-48.32 ...
survivalsvm version                    : 0.0.5
```

Και ο κώδικας που εφαρμόζει στο μοντέλο SVCR τον RBF kernel:

```
> ## efarmogh tou svcr(regression) me rbf_kernel
> survsvm.reg_RBF <- survivalsvm(Surv(diagtime, status) ~ ., subset
= train.index, data = veteran,type = "regression", gamma.mu = 1,
opt.meth = "quadprog", kernel = "RBF_kernel")
> print(survsvm.reg_RBF)

survivalsvm result

Call:

 survivalsvm(Surv(diagtime, status) ~ ., subset = train.index, data
= veteran, type = "regression", gamma.mu = 1, opt.meth = "quadprog",
kernel = "RBF_kernel")

Survival svm approach              : regression
Type of Kernel                     : RBF_kernel
Optimization solver used           : quadprog
Number of support vectors retained : 93
survivalsvm version                : 0.0.5
```

```
> pred.survsvm.reg_RBF <- predict(object = survsvm.reg_RBF, newdata
= veteran, subset = test.index)
> print(pred.survsvm.reg_RBF)


survivalsvm prediction

Type of survivalsvm                        : regression
Type of kernel                             : RBF_kernel
Optimization solver used in model          : quadprog
predicted risk ranks                       : 8.95 8.95 8.95 8.95 8.95
...
survivalsvm version                        : 0.0.5
```

Έγινε, επίσης, η εφαρμογή των μοντέλων Vanbelle1, Vanbelle2 και Hybrid, τρεις φορές το καθένα με τρεις διαφορετικούς πυρήνες τους additive, linear και RBF.

5.2 Εφαρμογή των μοντέλων SVM σε πραγματικά δεδομένα και σύγκριση με άλλα μοντέλα Ανάλυσης Επιβίωσης

Για να αξιολογήσουμε τα μοντέλα SVM επιβίωσης και την εφαρμογή μας, τέσσερα δημόσια διαθέσιμα σύνολα δεδομένων επιβίωσης χρησιμοποιήθηκαν. Το πρώτο είναι τα δεδομένα από την δοκιμαστική μελέτη του καρκίνου των πνευμόνων των βετεράνων που παρουσιάστηκαν και στην αρχή (*Kalbfleisch και Prentice,* 2002), διαθέσιμη στην επιβίωση της συσκευασίας. Περιλαμβάνει 137 άτομα και 5 μεταβλητές. Δεύτερον, χρησιμοποιήθηκαν στοιχεία από την Ερμηνεία μιας Δοκιμαστικής Μεσοπρόθεσμης μελέτης (*Ermerson and Banks,* 1994) στην οποία συμμετείχαν 130 άτομα. Δύο αποτελέσματα επιβίωσης ήταν αυτά που μας ενδιέφεραν στην μελέτη, η πλήρης ίαση και ο θάνατος. Τα αντίστοιχα σύνολα δεδομένων φέρουν την ένδειξη *leuk_cr* και *leuk_death*. Παρουσιάζονται 10 μεταβλητές και στα δύο σύνολα δεδομένων. Τρίτον, χρησιμοποιήθηκαν τα δεδομένα από τη Γερμανική Ομάδα Μελέτης για τον Καρκίνο του Μαστού 2 (*Germany Breast Cancer Study Group 2*, *GBSG2*) (*Schumacher et al.*, 1994) που αποτελείται από 686 δείγματα και 8 μεταβλητές. Τέλος, χρησιμοποιήθηκαν τα δεδομένα από τη μελέτη του Καρκίνου του

πνεύμονα του Mayo (*Mayo Clinic Lung Cancer*, MCLC) (*Loprinzi et al.,* 1994) που περιλαμβάνει 168 άτομα και 8 μεταβλητές. Ο Πίνακας 5.1 παρέχει μια σύντομη περίληψη των συνόλων δεδομένων που χρησιμοποιήθηκαν.

Πίνακας 5.1 Τα σύνολα δεδομένων που χρησιμοποιήθηκαν για τη σύγκριση της απόδοσης της πρόβλεψης και του χρόνου υπολογισμού των μοντέλων. Τα δεδομένα leuk_cr και leuk_death διαφέρουν μόνο στο γεγονός που εξετάζεται. Στο leuk_cr, το γεγονός είναι η πλήρης ίαση, ενώ στο leuk_death το γεγονός που μας ενδιαφέρει είναι ο θάνατος. Οι δύο τελευταίες στήλες δείχνουν τα ονόματα των μεταβλητών για την κατάσταση αποκοπής των δεδομένων και του χρόνου επιβίωσης.

Data set	Sample size	#Covariates	Status	Survival time
veteran	137	5	status	time
leuk_cr	130	10	complete_rem	data_cr
leuk_death	130	10	status_last_fol_up	data_last_fol_up
GBSG2	686	8	cens	time
MCLC	168	8	status	time

Για κάθε σύνολο δεδομένων, προσαρμόσαμε τα τέσσερα μοντέλα SVM επιβίωσης *Regression (SSVR)* που αντιστοιχεί στο *SVCR* (3.3.2) της παρούσας εργασίας, του *Van Belle1* που αντιστοιχεί στο *RANKSVMC* (4.2.1), του *VanBelle2* που αντιστοιχεί στο *MODEL1* (4.2.2), και του *Hybrid* το οποίο αντιστοιχεί στο *MODEL2* (4.2.3) χρησιμοποιώντας τους πυρήνες *linear*, *additive* και *RBF*. Η *ranking* προσέγγιση των *SVM* στην Ανάλυση επιβίωσης που εφαρμόζονται στο πακέτο της R *survpack* εφαρμόζεται με γραμμικούς και *RBF* πυρήνες, τους δύο πυρήνες που προσφέρονται από το πακέτο. Το μοντέλο *Cox PH*, το τυχαίο δάσος επιβίωσης (*Random Survival Forest, RSF*) (*Ishwaran et al*., 2008) και το *Gradient Boosting* (*Gboost*) για την ανάλυση επιβίωσης *Ridgeway* (1999) χρησίμοποιθηκαν ως μοντέλα αναφοράς.

Για τα μοντέλα των τυχαίων δέντρων επιβίωσης χρησιμοποιήθηκε το πακέτο *randomForestSRC* (*Ishwaran και Kogalur*, 2016). Ο αριθμός των τυχαίων μεταβλητών για διαχωρισμό και ο ελάχιστος αριθμός γεγονότων στους τερματικούς κόμβους συντονίστηκαν κατά την κατασκευή των δέντρων επιβίωσης. Το *randomForestSRC* αναφέρεται σε αυτές τις παραμέτρους ως *mtry* και *nodesize*, αντίστοιχα. Για τα μοντέλα gradient boosting που εφαρμόζονται στο *mboost* (*Hothorn et al*., 2016), τοποθετήθηκε ένα *PH* μοντέλο ως το βασικό για εκπαίδευση. Ο αριθμός των επαναλήψεων *boosting* και ο

συντελεστής παλινδρόμησης ρυθμίστηκαν. Στο *mboost*, αυτές οι παράμετροι ονομάζονται *mstop* και *nu*, αντίστοιχα.

Ο συντονισμός διεξήχθη με 5 - 10 φορές εμφωλευμένη διασταύρωση επικύρωσης. Τα σύνολα δεδομένων χωρίστηκαν τυχαία σε 5 σχεδόν ισοδύναμα μικρότερα δείγματα και σε κάθε επανάληψη, μία από τις 5 ομάδες των δειγμάτων αυτών χρησιμοποιήθηκε ως δοκιμαστικό σύνολο δεδομένων. Στις υπόλοιπες ομάδες, τα μοντέλα εκπαιδεύτηκαν μετά τον συντονισμό των παραμέτρων του μοντέλου με 10 φορές διασταυρωμένη επικύρωση. Τα καλύτερα μοντέλα επιλέχθηκαν με βάση το *c-index*. Το πακέτο *mlr* (*Bischl et al.*, 2016) χρησιμοποιήθηκε για ρύθμιση παραμέτρων.

Τα πειράματα για να εκτελεστούν χρειάζονται υπολογιστικές πλατφόρμες υψηλής απόδοσης. Ο Πίνακας 5.2 συνοψίζει τον εμπειρικό μέσο χρόνο εκτέλεσης που απαιτείται από κάθε μοντέλο *SVM* επιβίωσης για να εκτελέσει μία διαδικασία αναδειγματοληψίας. Όπως απεικονίζεται, οι χρόνοι εκτέλεσης των διεργασιών επηρεάζονται από τη συνάρτηση του πυρήνα που χρησιμοποιείται. Σε σύγκριση με τις συναρτήσεις πυρήνα των *linear* και *additive*, οι οποίες απαιτούνται περίπου ίσους χρόνους εκτέλεσης, οι χρόνοι εκτέλεσης της συνάρτησης πυρήνα *RBF* είναι υψηλότεροι. Ο λόγος είναι ότι το *RBF* απαιτεί μια επιπλέον παράμετρο, η οποία πρέπει να βελτιστοποιηθεί μέσω συντονισμού. Η επίδραση του αριθμού παραμέτρων παρατηρείται επίσης στην προσέγγιση *hybrid* επιβίωσης *SVM*, η οποία απαιτούσε περισσότερο χρόνο εκτέλεσης από ό, τι οι άλλες προσεγγίσεις. Τέλος, ο χρόνος εκτέλεσης των μοντέλων *ranking vanbelle1* και *vanbelle2*, που εφαρμόστηκαν σε *survivalsvm*, ήταν σημαντικά μικρότερος από αυτόν της προσέγγισης *evers*, που εφαρμόζεται στο πακέτο *survpack*.

Πίνακας 5.2 Ο μέσος χρόνος υπολογισμού για κάθε μοντέλο SVM επιβίωσης για να εκτελέσει μία διαδικασία. Η διαδικασία αυτή περιλαμβάνει ρύθμιση των παραμέτρων κανονικοποίησης για τα μοντέλα ranking και παλινδρόμησης. Ο χρόνος υπολογισμού των μοντέλων που χρησιμοποιούν τη συνάρτηση πυρήνα RBF συγκριτικά με τις συναρτήσεις πυρήνα linear και additive, χρειάζονται μία ακόμα παράμετρο να υπολογιστεί, οπότε ο χρόνος υπολογισμού είναι υψηλότερος. Επιπλέον, εφόσον η προσέγγιση hybrid χρησιμοποιεί δύο παραμέτρους κανονικοποίησης, χρειάζεται επίσης περισσότερο χρόνο για να βρει τις καλύτερες παραμέτρους κανονικοποίησης. Η εφαρμογή της προσέγγισης evers στο survpack δεν περιλαμβάνει τη συνάρτηση πυρήνα additive. † Διακόπηκε μετά από 10 ημέρες υπολογισμού.

Data set	Kernel	Mean runtime in minutes				
		vanbelle1	vanbelle2	SSVR	hybrid	evers
veteran	linear	0.45	0.44	1.33	16.12	6.57
	additive	0.58	0.60	1.41	15.97	
	RBF	4.79	5.06	14.02	144.03	43.33
leuk_cr	linear	0.87	0.90	1.38	26.24	3.56
	additive	0.98	0.99	1.54	30.75	
	RBF	2.98	3.21	8.57	238.97	21.13
leuk_death	linear	0.29	0.30	0.96	28.08	4.82
	additive	0.33	0.36	0.96	30.84	
	RBF	2.99	3.10	8.52	269.54	20.04
GBSG2	linear	2.90	3.01	41.89	1064.53	1005.39
	additive	3.65	4.21	65.57	374.68	
	RBF	30.91	35.23	597.77	NA†	NA†
MCLC	linear	0.47	0.46	1.83	48.07	14.92
	additive	0.64	0.62	2.10	17.18	
	RBF	4.86	5.11	17.98	585.94	81.34

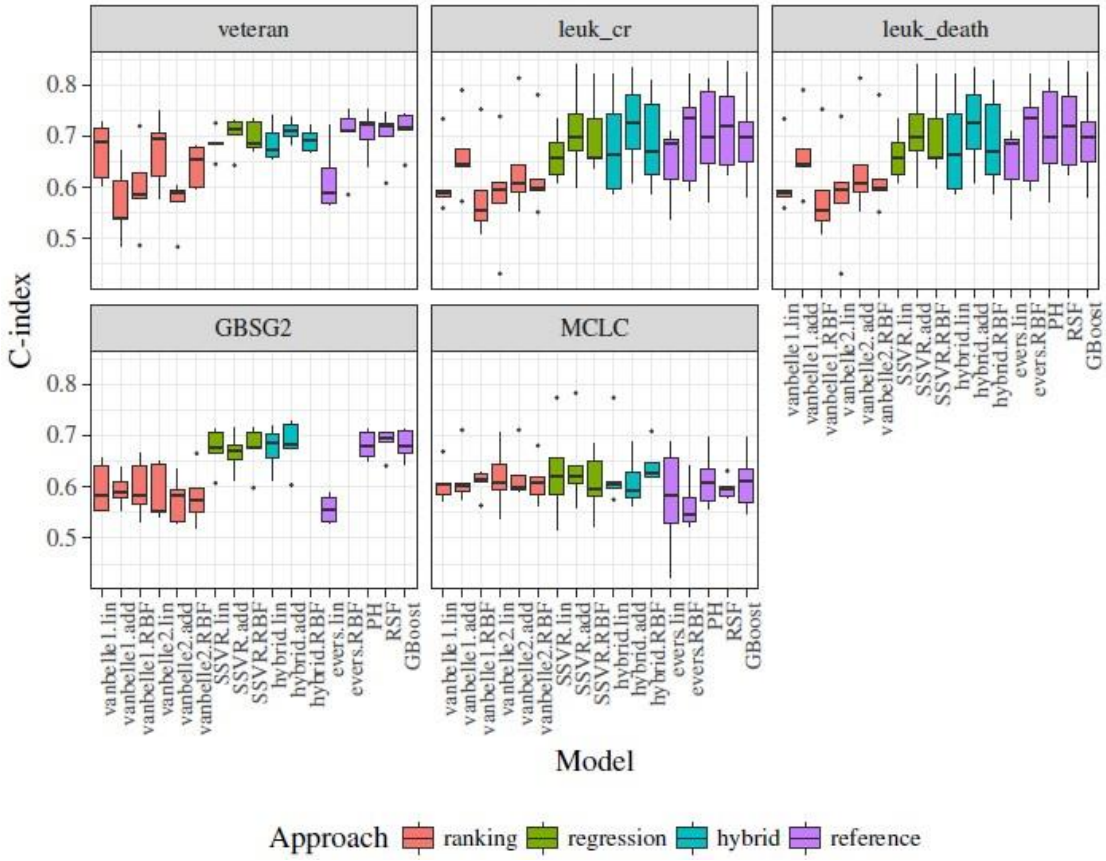
Ο Πίνακας 5.3 και το Σχήμα 5.1 παρουσιάζουν τις εκτιμήσεις απόδοσης των μοντέλων που συγκρίνουμε με βάση το δείκτη σύγκρισης. Από τα 17 μοντέλα, το καλύτερο μοντέλο έλαβε την κατάταξη 1, το χειρότερο μοντέλο την κατάταξη 17. Στην περίπτωση που υπήρχαν ισοπαλίες, σε όλα τα ισόπαλα μοντέλα δόθηκε η μέση τιμή των αντίστοιχων αριθμών που ήταν στην σειρά της κατάταξης. Το μοντέλο *hybrid* με την συνάρτηση πυρήνα *additive* στα δεδομένα που αφορούσαν το *veteran*, έχει εμφανίσει την καλύτερη προσαρμογή απ' όλα τα μοντέλα SVM επιβίωσης, με ελάχιστες μόνο διαφορές από το μοντέλο *PH* εφαρμόζοντας την προσέγγιση *SSVR*. Τα *ranking* μοντέλα (*vanbelle1* και *vanbelle2*) είχαν τις χειρότερες προσαρμογές από τα άλλα μοντέλα. Οι διαφορές μεταξύ των μοντέλων αυτών ήταν μικρές. Για το σύνολο δεδομένων

leuk_cr, η προσέγγιση *evers* με τον πυρήνα *RBF* ήταν η καλύτερη προσέγγιση *SVM*, ακολουθούμενη από την προσέγγιση *hybrid* με τον πυρήνα *additive*. Τα καλύτερα μοντέλα για τα δεδομένα αναφοράς ήταν το μοντέλο *PH* και *GBoost*. Για το *leuk_death*, η προσέγγιση *hybrid* με τον πυρήνα *additive* είχε την ίδια τιμή με το μοντέλο των τυχαίων δέντρων επιβίωσης. Το μοντέλο *RSF* είχε επίσης καλύτερη απόδοση από τα μοντέλα αναφοράς που εφαρμόστηκαν στο σύνολο δεδομένων *GBSG2*, ενώ τα *evers* παρουσίασαν την χειρότερη απόδοση. Το μοντέλο *hybrid* ήταν ακόμα το καλύτερο μοντέλο *SVM* επιβίωσης, με σχεδόν τα ίδια αποτελέσματα για τους πυρήνες *linear* και *additive*. Δεν μπόρεσαν να βρεθούν καθόλου αποτελέσματα για τον πυρήνα *RBF* στις 10 ημέρες υπολογισμού. Για το σύνολο δεδομένων *MCLC*, το μοντέλο SVM επιβίωσης *hybrid* σε συνδυασμό με τον πυρήνα πυρήνα *RBF* ήταν το καλύτερο, ενώ το μοντέλο *PH* ήταν το καλύτερο μοντέλο αναφοράς. Οι διαφορές ήταν μικρές, εκτός από τους *evers*, οι οποίοι παρουσίασαν χειρότερα. Συνοπτικά, υπάρχουν μόνο μικρές διαφορές μεταξύ των καλύτερων μοντέλων *SVM* επιβίωσης και τα καλύτερων μοντέλων αναφοράς. Ωστόσο, οι διαφορές μεταξύ των διαφορετικών προσεγγίσεων *SVM* και των πυρήνων που χρησιμοποιούνταν κάθε φορά ήταν σημαντικές.

Πίνακας 5.3 Επίδοση Πρόβλεψης των πέντε *SVM* επιβίωσης (*vanbelle1, vanbelle2, SSVR, hybrid* και *evers*) και τριών μεθόδων αναφοράς (*PH*, *RSF* και *GBoost*) σε 5 σύνολα δεδομένων. Ο *c-index* υπολογίστηκε για κάθε μέθοδο χρησιμοποιώντας 5-10 φορές εμφωλευμένη διασταύρωση επικύρωσης. Φαίνονται οι μέσες τιμές και οι τυπικές αποκλίσεις (*Standard deviations, SD*). Ο *c-index* και η κατάταξη παρουσιάζεται για κάθε μοντέλο. Τιμή *Rank* 1 δόθηκε στο μεντέλο με την καλύτερη προσαρμογή, ενώ τιμή *Rank* 17 δόθηκε σε αυτό με τη χειρότερη προσαρμογή. Στην περίπτωση που υπήρχαν ισοπαλίες, σε όλα τα ισόπαλα μοντέλα δόθηκε η μέση τιμή των αντίστοιχων αριθμών που ήταν στην σειρά της κατάταξης. Ο βελτιστοποιητής *quadprog* χρησιμοποιήθηκε στο πακέτο *survivalsvm*. Τα καλύτερα μοντέλα *SVM* και τα καλύτερα μοντέλα αναφοράς έχουν υπογραμμιστεί με ανοιχτό και σκούρο γκρι χρώμα, αντίστοιχα. Η εκτέλεση της προσέγγισης *evers* στο *survpack* δεν περιλαμβάνει την συνάρτηση πυρήνα *additive*.

† Διακόπηκε μετά από 10 ημέρες υπολογισμού.

		veteran		leuk_cr		leuk_death		GBSG2		MCLC	
Method	**Kernel**	***C*-index (SD)**	**Rank**	***C*-index (SD)**	**Rank**	***C*-index (SD)**	**Rank**	***C*-index (SD)**	**Rank**	***C*-index (SD)**	**Rank**
vanbelle1	linear	0.68 (0.05)	11	0.61 (0.05)	13	0.62 (0.06)	15	0.60 (0.05)	9	0.61 (0.05)	9.5
	additive	0.57 (0.07)	17	0.63 (0.06)	12	0.66 (0.08)	11	0.59 (0.03)	11	0.62 (0.05)	5.5
	RBF	0.60 (0.10)	14	0.53 (0.19)	17	0.59 (0.12)	17	0.60 (0.07)	10	0.61 (0.03)	7
vanbelle2	linear	0.59 (0.15)	16	0.59 (0.06)	14.5	0.61 (0.08)	16	0.59 (0.06)	12	0.61 (0.06)	12
	additive	0.59 (0.07)	15	0.57 (0.03)	16	0.64 (0.10)	13	0.58 (0.04)	13	0.62 (0.05)	5.5
	RBF	0.64 (0.05)	12	0.59 (0.06)	14.5	0.63 (0.11)	14	0.58 (0.07)	14	0.61 (0.05)	9.5
SSVR	linear	0.69 (0.03)	8.5	0.66 (0.05)	10	0.68 (0.08)	10	0.67 (0.04)	6.5	0.63 (0.09)	3
	additive	0.71 (0.05)	3	0.70 (0.03)	7	0.71 (0.09)	3	0.67 (0.04)	6.5	0.64 (0.08)	2
	RBF	0.70 (0.04)	4	0.68 (0.07)	9	0.70 (0.09)	4	0.67 (0.06)	8	0.61 (0.08)	14
hybrid	linear	0.69 (0.04)	10	0.71 (0.05)	5	0.68 (0.05)	9	0.68 (0.04)	4	0.62 (0.03)	4
	additive	0.71 (0.02)	1	0.72 (0.02)	4	0.72 (0.09)	1.5	0.68 (0.05)	5	0.61 (0.05)	9.5
	RBF	0.69 (0.03)	8.5	0.71 (0.06)	6	0.69 (0.11)	8	NA†		0.64 (0.05)	1
evers	linear	0.62 (0.08)	13	0.64 (0.11)	11	0.65 (0.09)	12	0.56 (0.03)	15	0.58 (0.13)	16
	RBF	0.70 (0.08)	6	0.73 (0.06)	3	0.70 (0.12)	6	NA†		0.56 (0.06)	17
PH Model		0.71 (0.04)	2	0.73 (0.03)	1.5	0.70 (0.10)	5	0.68 (0.03)	2.5	0.61 (0.05)	9.5
RSF		0.70 (0.05)	5	0.70 (0.06)	8	0.72 (0.09)	1.5	0.69 (0.03)	1	0.60 (0.02)	15
GBoost		0.70 (0.09)	7	0.73 (0.03)	1.5	0.69 (0.09)	7	0.68 (0.03)	2.5	0.61 (0.06)	13



Σχήμα 5.1 Η απόδοση πρόβλεψης για τα πέντε μοντέλα *SVM* επιβίωσης (*vanbelle1*, *vanbelle2*, *SSVR*, *hybrid* και *evers*) και τρεις μεθόδους αναφοράς (*PH*, *RSF* και *GBoost*) σε 5 σύνολα δεδομένων. Ο δείκτης C υπολογίστηκε για κάθε μέθοδο με τη χρήση εγκάρσιας επικύρωσης 510. Χρησιμοποιήθηκε το *quadprog optimizer* στο πακέτο *survivalsvm*. Τα διαγράμματα δημιουργήθηκαν χρησιμοποιώντας το *ggplot2* (*Wickham*, 2009) και *tikzDevice* (*Sharpsteen et al.*., 2016).

5.3 Παρατηρήσεις και συμπεράσματα από την εφαρμογή της *R* στα δεδομένα

Παρουσιάσαμε το πακέτο *R survivalsvm* για την προσαρμογή μοντέλων *SVM* επιβίωσης. Τρεις προσεγγίσεις είναι διαθέσιμες στο πακέτο. Πρώτον, στην προσέγγιση παλινδρόμησης, το κλασικό *SVR* επεκτάθηκε για αποτελέσματα με αποκομμένες παρατηρήσεις. Δεύτερον, στην προσέγγιση ranking, οι κατατάξεις μεταξύ των προβλέψεων μοντέλου και παρατηρούμενων χρόνων επιβίωσης μεγιστοποιούνται βάσει του δείκτη σύγκρισης (*c-index*). Η τρίτη προσέγγιση, που ονομάζεται *hybrid*, συνδυάζει τις δύο πρώτες προσεγγίσεις σε ένα ενιαίο μοντέλο. Εφαρμόσαμε αυτές τις τρεις προσεγγίσεις

στο πακέτο *survivalsvm* και χρησιμοποιήσαμε 5 σύνολα δεδομένων για να συγκρίνουμε την απόδοση πρόβλεψης και το χρόνο εκτέλεσης τους με την εφαρμογή των μοντέλων *Cox PH*, τυχαίων δέντρων επιβίωσης (*RSF*) και *Gradient Boosting*. Επιπλέον, συμπεριλάβαμε ένα εφαρμογή μιας παραλλαγής της προσέγγισης *ranking* (*Evers* και *Messow*, 2008) στη σύγκριση. Από τα μοντέλα *SVM* επιβίωσης, η προσέγγιση *hybrid* γενικά αποδίδει καλύτερα από την άποψη πρόβλεψης αλλά ήταν πιο αργή στον υπολογισμό. Παρατηρήσαμε μόνο μικρές διαφορές μεταξύ των καλύτερων μοντέλων *SVM* και των καλύτερων μοντέλων αναφοράς και ήταν εφικτό να καθοριστεί ένας σαφής νικητής.

Συγκρίνοντας τα *ranking* μοντέλα τα μοντέλα παλινδρόμησης, η προσέγγιση των *evers* απαιτούσε πάντα περισσότερο χρόνο υπολογισμού από τις προσεγγίσεις που εφαρμόστηκαν στο *survivalsvm*. Το μοντέλο *hybrid* ήταν η καλύτερη προσέγγιση των *SVM* επιβίωσης, όμως ο χρόνος εκτέλεσης αυξήθηκε σημαντικά. Αυτό οφείλεται στο γεγονός ότι η hybrid προσέγγιση χρειάζεται δύο παραμέτρους κανονικοποίησης, ενώ οι προσεγγίσεις *ranking* και παλινδρόμησης απαιτούν μόνο μία παράμετρο.

Οι συναρτήσεις του πυρήνα που παρουσιάζουν τις καλύτερες επιδόσεις κάθε φορά εξαρτώνται από το σύνολο των δεδομένων στο οποίο εφαρμόζονται και το επιλεγμένο μοντέλο *SVM*. Για τις *ranking* προσεγγίσεις, οι διαφορές ήταν μεγαλύτερες από αυτές των μοντέλων παλινδρόμησης και *hybrid*. Για την προσέγγιση *hybrid*, οι πυρήνες *additive* και *RBF* παρουσίασαν τα καλύτερα αποτελέσματα. Ωστόσο, οι χρόνοι εκτέλεσης για τον πυρήνα *RBF* ήταν σημαντικά μεγαλύτεροι. Και πάλι, αυτό οφείλεται στη ρύθμιση μιας επιπλέον παραμέτρου.

Η εφαρμογή μας χρησιμοποιεί τετραγωνικό προγραμματισμό και ένα εσωτερικό βελτιστοποιητή σημείου για την επίλυση του τετραγωνικού προβλήματος βελτιστοποίησης που προέρχεται από το πρωταρχικό πρόβλημα βελτιστοποίησης διανυσμάτων φορέα υποστήριξης. Όταν χρησιμοποιείται ο τετραγωνικός προγραμματισμός για μη θετικά ημι-ορισμένο πίνακα πυρήνα, αυτός ο πίνακας είναι τροποποιείται ελαφρά στον πλησιέστερο θετικά ημι-ορισμένο πίνακα. Καλώντας τον εσωτερικό βελτιστοποιητή σημείου δεν γίνεται καμία τροποποίηση στον αρχικό πίνακα, αλλά είναι υπολογιστικά πιο αργό, αφού το λογισμικό εφαρμόζεται πλήρως στον *R*. Οι *Pölsterl et al*. (2015)

πρότειναν έναν γρήγορο αλγόριθμο για την εκπαίδευση *SVM* επιβίωσης στον αρχικό χώρο. Αυτός ο αλγόριθμος είναι γρήγορος σε μικρές διαστάσεις, αλλά για προβλήματα μεγάλων διαστάσεων οι συγγραφείς προτείνουν τη μείωση των διαστάσεων πριν από την εφαρμογή ενός αλγορίθμου *SVM*. Ωστόσο, μερικοί ειδικοί και γρήγοροι αλγόριθμοι, όπως η διαδοχική ελάχιστη βελτιστοποίηση (*Sequential Minimal Optimization, SMO*) (*Platt*, 1998), η οποία είναι διαθέσιμη για κλασικά προβλήματα βελτιστοποίησης *SVM*, τα οποία αποδείχθηκαν πιο ακριβή (*Horn et al.*, 2016). Η εφαρμογή για τα *SVM* επιβίωσης θα μπορούσε ενδεχομένως να βελτιωθεί με επέκταση της *SMO* διαδικασίας βελτιστοποίησης.

Έχοντας περιορισμούς μόνο στον πλησιέστερο γείτονα στην προσέγγιση κατάταξης, όπως διατυπώνεται στο *vanbelle1* και *vanbelle2*, μπορεί να βελτιωθεί σημαντικά η υπολογιστική απόδοση, αλλά μπορεί όμως να μειωθεί και η απόδοση πρόβλεψης. Κατ 'αρχήν, ο αριθμός των πλησιέστερων γειτόνων δεν περιορίζεται στις επιλογές του *Evers* and *Messow* (2008) και *Van Belle et al.* (2008). Δεδομένου ότι ο βέλτιστος αριθμός γειτόνων εξαρτάται από το σύνολο δεδομένων και τη διαθεσιμότητα των υπολογιστικών πόρων, μπορεί να συμπεριληφθεί ως μία ακόμη παράμετρος ρύθμισης.

Εν κατακλείδι, δείξαμε ότι τα *SVM* είναι μια χρήσιμη εναλλακτική λύση για την πρόβλεψη επιβίωσης. Το πακέτο *survivalsvm* παρέχει μια γρήγορη και εύκολη στη χρήση εφαρμογή των διαθέσιμων προσεγγίσεων των *SVM* επιβίωσης. Τα αποτελέσματά μας δείχνουν ότι η επιλογή του μοντέλου *SVM* και της συνάρτησης πυρήνα είναι πολύ σημαντικός. Εκτός από την απόδοση πρόβλεψης, ο χρόνος υπολογισμού είναι μια σημαντική πτυχή για τα μεγάλα σύνολα δεδομένων. Συμπερασματικά, προτείνεται να διεξάγονται πειράματα αναφοράς χρησιμοποιώντας διάφορες προσεγγίσεις και διαθέσιμους πυρήνες πριν γίνει ανάλυση ενός συνόλου δεδομένων.

Κεφάλαιο 6- Σύνοψη

Στην παρούσα διπλωματική παρουσιάστηκαν Μέθοδοι Διανυσμάτων υποστήριξης στην Ανάλυση Επιβίωσης. Τα αποκομμένα δεδομένα που παρουσιάζονται συχνά σε αυτόν τον τομέα είναι το πιο σημαντικό στοιχείο που πρέπει να ληφθεί υπόψη στην προσπάθεια εκτίμησης του χρόνου επιβίωσης. Παρουσιάστηκαν, λοιπόν στην εργασία αυτή μοντέλα που λαμβάνουν υπόψη αυτή την ιδιαιτερότητα των δεδομένων επιβίωσης και έγινε, επίσης, προσπάθεια εφαρμογής αυτών των μοντέλων σε πραγματικά δεδομένα.

Συγκεκριμένα στο 1ο κεφάλαιο έγινε μια εισαγωγή στην Ανάλυση επιβίωσης, με παρουσίαση των βασικών συναρτήσεων του χρόνου επιβίωσης και βασικών μεθόδων για την εκτίμησή τους. Περιεγράφηκαν τα αποκομμένα δεδομένα και έγινε ταξινόμηση των μεθόδων της Ανάλυσης Επιβίωσης. Ακόμα παρουσιάστηκε και ένα μέτρο αξιολόγησης της απόδοσης, ο δείκτης σύγκρισης.

Στο 2ο έγινε μια παρουσίαση των βασικών μεθόδων Μηχανικής Μάθησης. Παρουσιάστηκε επίσης η γενική ιδέα των Μηχανών Διανυσμάτων Υποστήριξης και η εφαρμογή της σε γραμμικά διαχωρίσιμα και μη, προβλήματα.

Στο 3ο γίνεται περιγραφή της Ανάλυσης Παλινδρόμησης μέσω Μηχανών Διανυσμάτων Υποστήριξης και παρουσίαση των βασικών μοντέλων που χρησιμοποιούνται στην Ανάλυση Επιβίωσης, λαμβάνοντας υπόψη την ιδιαιτερότητα των αποκομμένων δεδομένων.

Στο 4ο εισάγεται η Ταξινόμηση μέσω Μηχανών Διανυσμάτων Υποστήριξης. Παρουσιάζονται τέσσερα μοντέλα με βάση περιορισμούς κατάταξης.

Στο 5ο έγινε εφαρμογή σε πραγματικά δεδομένα κάποιων από τα μοντέλα που παρουσιάστηκαν και παρατηρήσαμε τα διαφορετικά αποτελέσματα που είχαμε με την εφαρμογή των διαφορετικών μοντέλων και τη χρήση διαφορετικών πυρήνων, μέσω της γλώσσας προγραμματισμού *R*.

Τέλος, στο 6ο κεφάλαιο παρουσιάστηκε μια σύνοψη της εργασίας.

Βιβλιογραφία

1. B Baesens, T Van Gestel, M Stepanova, D Van den Poel & J Vanthienen (2005) Neural network survival analysis for personal loan data, *Journal of the Operational Research Society*, 56:9, 1089-1098
2. Bellazzi, R. and Zupan, B.. (2008). Predictive data mining in clinical medicine: current issues and guidelines. *International journal of medical informatics* 77, 2, 81–97.
3. S original by Berwin A. Turlach R port by Andreas Weingessel <Andreas.Weingessel@ci.tuwien.ac.at> (2013). quadprog: *Functions to solve Quadratic Programming Problems.*. R package version 1.5-5. https://CRAN.R-project.org/package=quadprog
4. Biganzoli, E. , Boracchi, P. , Mariani, L. and Marubini, E. (1998), Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Statist. Med*., 17: 1169-1186.
5. B. Bischl, M. Lang, L. Kotthoff, J. Schiffner, J. Richter, Z. Jones, and G. Casalicchio. Mlr*: Machine Learning in R*, 2016. URL https://CRAN.R-project.org/package=mlr. R package version 2.9.
6. Hans W. Borchers (2018). *pracma: Practical Numerical Math Functions*. R package version 2.1.8. https://CRAN.R-project.org/package=pracma
7. Bou-Hamad, I., Larocque D., Ben-Ameur, H. A review of survival trees. *Statistics Surveys* 5 (2011), 44—71
8. Boyd S., Vandenberghe L., *Convex optimization*, Cambridge University Press, Cambridge, 2004.
9. Breiman, L. (1996). Bagging predictors. *Machine learning* 24, 2, 123–140.
10. Breiman, L. (2001). Random forests. *Machine learning* 45, 1, 5–32.
11. Brown, S. F., Branford, A. J. and Moran, W. (1997). On the use of artificial neural networks for the analysis of survival data. *IEEE Transactions on Neural Networks* 8, 5 (1997), 1071–1077.
12. Bühlmann, Peter; Hothorn, Torsten. Boosting Algorithms: Regularization, Prediction and Model Fitting. *Statistical Science* 22 (2007), no. 4, 477--505..

13. Burges, C. J. C. (1998). "*A Tutorial on Support Vector Machines for Pattern Recognition*". Kluwer Academic Publishers, Boston.
14. Cesaire J. K. Fouodo (2018). survivalsvm: *Survival Support Vector Analysis*. R package version 0.0.5. https://CRAN.R-project.org/package=survivalsvm
15. Césaire J. K. Fouodo, Inke R. König, Claus Weihs, Andreas Ziegler and Marvin N. Wright , *The R Journal* (2018) 10:1, pages 412-423.
16. Ciampi, A. , Bush, R. S., Gospodarowicz, M. and Till, J. E. (1981), An approach to classifying prognostic factors related to survival experience for non-Hodgkin's lymphoma patients: Based on a series of 982 patients: 1967–1975. *Cancer*, 47: 621-627.
17. Ciampi A., Chang CH., Hogg S., McKinney S. (1987) *Recursive Partition: A Versatile Method for Exploratory-Data Analysis in Biostatistics*. In: MacNeill I.B., Umphrey G.J., Donner A., Jandhyala V.K. (eds) Biostatistics. The University of Western Ontario Series in Philosophy of Science (A Series of Books in Philosophy of Science, Methodology, Epistemology, Logic, History of Science and Related Fields), vol 38. Springer, Dordrecht
18. Ciampi, Antonio & Thiffault, Johanne & Nakache, Jean-Pierre & Asselain, Bernard. (1986). Stratification by stepwise regression, correspondence analysis and recursive partition: A comparison of three methods of analysis for survival data with covariates. *Computational Statistics & Data Analysis*. 4. 185-204. 10.1016/0167-9473(86)90033-2.
19. Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273-297.
20. Cox, D. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological), 34(2)*, 187-220.
21. Daemen, A., De Moor, B. (2009). "*Development of a kernel function for clinical data*", in: Proceedings of the 31th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS), IEEE, Piscataway, 2009, pp. 5913–5917.
22. Davis, R. B. and Anderson, J. R. (1989), Exponential survival trees. *Statistics in Medicine*, 8: 947-961.

23. Dietterich T.G. (2000) *Ensemble Methods in Machine Learning*. In: Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science, vol 1857. Springer, Berlin, Heidelberg
24. Z. Ding. "The application of support vector machine in survival analysis," *2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC),* Dengleng, 2011, pp. 6816-6819.
25. Douglas Bates and Martin Maechler (2018). *Matrix: Sparse and Dense Matrix Classes and Methods*. R package version 1.2-15. https://CRAN.R-project.org/package=Matrix.
26. Du and S. Dua, Cancer prognosis using support vector regression in imaging modality. *World Journal Of Clinical Oncology*, 2011. 2(1): 44.
27. S. Ermerson and P. Banks. Interpretation of a leukemia trial stopped early. In N. Lange, editor, *Case Studies in Biometry*, chapter 14, pages 275 – 99. John Wiley & Sons, 1994.
28. Evers, L., & Messow, C. (2008). Sparse kernel methods for high-dimensional survival data. *Bioinformatics*, 24 14, 1632-8.
29. Faraggi, D. and Simon, R. (1995). A neural network model for survival data. *Statistics in medicine* 14, 1 (1995), 73–82.
30. Fard, M. J., Wang, P., Chawla, S. and Reddy, C. K. (2016). A bayesian perspective on early stage event prediction in longitudinal data. *IEEE Transactions on Knowledge and Data Engineering* 28, 12 (2016), 3126–3139.
31. J.G. Silva Fonseca, T. Oozeki, T. Takashima, G. Koshimizu, Y. Uchida, et al., Use of support vector regression and numerically predicted cloudiness to forecast power output of a photovoltaic power plant in Kitakyushu, Japan. Progress in photovoltaics: *Research and applications*, 2012. 20(7): pp. 874-882.
32. Friedman, N., Geiger, D. and Goldszmidt, M. (1997). Bayesian network classifiers. *Machine learning* 29, 2, 131–163.
33. Goldfarb D., Idnani A. (1982) *Dual and primal-dual methods for solving strictly convex quadratic programs*. In: Hennart J.P. (eds) Numerical

Analysis. Lecture Notes in Mathematics, vol 909. Springer, Berlin, Heidelberg

34. Gordon, L. and Olshen, R. A. (1985). Tree-structured survival analysis. *Cancer treatment reports* 69, 10, 1065–1069
35. Harrell, F. E., Califf, R.M., Pryor,D.B., Lee, K.L. and Rosati, R.A. (1982). Evaluating the yield of medical tests". *Journal of the American Medical Association* 247, 18, 2543–2546.
36. Harrell, F.E., Lee, K.L., Califf, R. M., Pryor, D.B, and Rosati, R.A. (1984). Regression modelling strategies for improved prognostic prediction. *Statistics in medicine* 3, 2 (1984), 143–152.
37. Heagerty, P.J. and Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics* 61, 1, 92–105.
38. Herbrich R., Graepel T., Obermayer K., (2000). Large margin rank boundaries for ordinal regression, *Advances in Large Margin Classifiers* 115–132
39. D. Horn, A. Demircioglu, B. Bischl, T. Glasmachers, and C. Weihs. A comparative study on large scale kernelized support vector machines. *Advances in Data Analysis and Classification*, pages 1–17, 2016. URL https://dx.doi.org/10.1007/s11634-016-0265-7.
40. T. Hothorn, P. Buehlmann, T. Kneib, M. Schmid, and B. Hofner. mboost: *Model-Based Boosting*, 2016. URL http://CRAN.R-project.org/package=mboost. R package version 2.6-0.
41. Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A.A., & Laan, M.J. (2006). Survival ensembles. *Biostatistics*, 7 3, 355-73.
42. Hothorn, T. , Lausen, B. , Benner, A. and Radespiel-Tröger, M. (2004), Bagging survival trees. *Statistics in medicine*, 23: 77-91..
43. Ishwaran, Hemant; Kogalur, Udaya B.; Blackstone, Eugene H.; Lauer, Michael S. Random survival forests. *Annals of applied statistics*. 2 (2008), no. 3, 841-860.
44. Ishwaran, H. , Kogalur, U. B., Chen, X. and Minn, A. J. (2011), Random survival forests for high-dimensional data. *Statistical Analysis and Data Mining*, 4: 115-132..

45. H. Ishwaran and U. B. Kogalur. *Random Forests for Survival, Regression and Classification* (RF-SRC), 2016. URL https://cran.r-project.org/package=randomForestSRC. R package version 2.3.0.

46. Καμπουρλάζος, Β., Παπακώστας, Γ., 2015. *Εισαγωγή στην υπολογιστική νοημοσύνη*. [ηλεκτρ. βιβλ.] Αθήνα:Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών.

47. Kalbfleisch, J.D. and Prentice, R.L., (2002). *The Statistical Analysis of Failure Time Data*. Wiley series in probability and statistics, New York.

48. Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab - An S4 Package for Kernel Methods in R. *Journal of Statistical Software*, 11(9), 1 – 20.

49. Χ. Καρώνη. (2009), *Μοντέλα αξιοπιστίας και επιβίωσης*, Εκδόσεις ΣΥΜΕΩΝ.

50. Kazem, E. Sharifi, F.K. Hussain, M. Saberi, and O.K. Hussain, Support vector regression with chaos-based firefly algorithm for stock market price forecasting. *Applied Soft Computing*, 2013. 13(2): pp. 947-958.

51. F. M. Khan and V. B. Zubek, "Support Vector Regression for Censored Data (SVRc): A Novel Tool for Survival Analysis," *2008 Eighth IEEE International Conference on Data Mining*, Pisa, 2008, pp. 863-868.

52. Klein, J. P. and Moeschberger, M. L. (2005). *Survival Analysis Techniques for Censored and Truncated Data*. 2nd edition *Springer-Verlag*, New York.

53. Kononenko, I. (1993). Inductive and Bayesian learning in medical diagnosis. *Applied Artificial Intelligence an International Journal* 7, 4 317–337.

54. LeBlanc, M., & Crowley, J. (1992). Relative Risk Trees for Censored Survival Data. *Biometrics*, 48(2), 411-425.

55. Li, Y., Wang, J., Ye, J. and Reddy, C. K., (2016b). A multi-task learning formulation for survival analysis. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 1715–1724.

56. Y. Li, L. Wang, J. Wang, J. Ye and C. K. Reddy, "Transfer Learning for Survival Analysis via Efficient L2,1-Norm Regularized Cox Regression,"

2016 IEEE 16th International Conference on Data Mining (ICDM), Barcelona, 2016a, pp. 231-240.

57. Liang, Y., Xu, Q.-S., Li, H.-D., Cao, D.-S. (2011). *Support vector machines and their application in chemistry and biotechnology*. CRC-Press, Taylor and Francis Group, Boca Raton.
58. Liestbl, K. , Andersen, P. K. and Andersen, U. (1994), Survival analysis and neural nets. *Statistics in medicine* 13, 12 1189-1200.
59. Lisboa, P. JG, Wong, H., Harris, P. and Swindell, R. (2003). A bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer. *Artificial intelligence in medicine* 28, 1, 1–25.
60. C. L. Loprinzi, J. A. Laurie, H. S. Wieand, J. E. Krook, P. J. Novotny, J. W. Kugler, J. Bartel, M. Law, M. Bateman, and N. E. Klatt. Prospective evaluation of prognostic variables from patient-completed questionnaires. North Central Cancer Treatment Group. Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology, 12(3): 601–607, 1994. URL https://dx.doi.org/10.1200/JCO.1994.12.3.601.
61. David J C Mackay (1995) Probable networks and plausible predictions — a review of practical Bayesian methods for supervised neural networks, Network: *Computation in Neural Systems*, 6:3, 469-505.
62. Mahjub, H., Goli, S., Faradmal, J., Soltanian, A.-R., Performance Evaluation of Support Vector Regression Models for Survival Analysis: A Simulation Study. *International Journal of Advanced Computer Science and Applications(IJACSA),* 7(6), 2016.
63. Mariani, L., Coradini, D., Biganzoli, E., Boracchi, P., Marubini, E., Pilotti, S., Salvadori, B., Silvestrini, R., Veronesi, U., Zucali, R. and others. Prognostic factors for metachronous contralateral breast cancer: a comparison of the linear Cox regression model and its artificial neural network extension. *Breast cancer research and treatment* 44, 2 (1997), 167–178.
64. Mercer, J. (1909). Functions of Positive and Negative Type, and their Connection with the Theory of Integral Equations. *Philosophical*

Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character,209, 415-446.

65. S. J. Pan and Q. Yang, "A Survey on Transfer Learning," in *IEEE Transactions on Knowledge and Data Engineering*, 22, 10, pp. 1345-1359, 2010.
66. Pencina, M. J. and D'Agostino, R. B. (2004), Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in medicine*, 23: 2109-2123.
67. Pepe, M. S. (2003). "*The statistical evaluation of medical tests for classification and prediction*". Oxford University Press, USA.
68. J. Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical report, Microsoft Research, 1998. URL https://www.microsoft.com/enus/research/publication/sequential-minimal-optimization-a-fast-algorithm-fortraining- support-vector-machines. Date of access: April 4, 2017.
69. S. Pölsterl, N. Navab, and A. Katouzian. Fast training of support vector machines for survival analysis. In A. Appice, P. P. Rodrigues, V. Santos Costa, J. Gama, A. Jorge, and C. Soares, editors, *Machine Learning and Knowledge Discovery in Databases*, ECML PKDD 2015, pages 243–259. Springer-Verlag, 2015. URL https://dx.doi.org/10.1007/978-3-319-23525-7_15.
70. R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
71. Raftery, A.E. (1995) Bayesian Model Selection in Social Research, *Sociological Methodology*, 25: 111-163.
72. Raftery, A., Madigan, D. and Volinsky, C. T. (1995). Accounting for model uncertainty in survival analysis improves predictive performance. *Bayesian Statistics* 5, 323–349.
73. Ravdin, P. M. and Clark, G. M. (1992). A practical application of neural network analysis for predicting outcome of individual breast cancer patients. *Breast Cancer Research and Treatment*, 22:285-293.
74. G. Ridgeway. The state of boosting*. Computing Science and Statistics*, 31:172–181, 1999.

75. Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain*. Psychological Review*, 65(6), 386-408.
76. Safavian, S. R. and Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics* 21, 3, 660–674
77. M. Schumacher, B. G., H. Bojar, K. Huebner, M. Olschewski, W. Sauerbrei, C. Schmoor, C. Beyerle, R. L. A. Neumann, and H. F. Rauschecker. Randomized 2 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. *Journal of Clinical Oncology*, 12(10):2086 – 2093, 1994. URL https://dx.doi.org/10.1200/JCO.1994.12.10.2086.
78. Segal, M. (1988). Regression Trees for Censored Data. *Biometrics*, 44(1), 35-47.
79. C. Sharpsteen, C. Bracken, K. Müller, and Y. Xie. tikzDevice: *R Graphics Output in LaTeX Format*, 2016. URL https://CRAN.R-project.org/package=tikzDevice. R package version 0.10-1.
80. P.K. Shivaswamy, W. Chu and M. Jansche. A support vector approach to censored targets. in Data Mining, 2007. *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, Omaha, NE, 2007, pp. 655-660.
81. Statnikov, A., Hardin, D., Guyon, I., & Aliferis, C. F. (2009). A gentle introduction to support vector machines in biomedicine. In *AMIA Annual Symposium* (pp. 1–207). San Francisco.
82. Steck, H., Krishnapuram, H. B., Dehing-Oberije, C., Lambin, P., and Raykar, V.C. (2008). On ranking in survival analysis: Bounds on the concordance index. In *Advances in neural information processing systems*. 1209–1216.
83. V. Van Belle, K. Pelckmans, J.A. Suykens and S. Van Huffel, Additive survival least-squares support vector machines. *Statistics in Medicine*, 2010. 29(2): pp. 296-308.
84. V. Van Belle, K. Pelckmans, J. A. K. Suykens, and S. Van Huffel, "Support Vector Machines For Survival Analysis*," in Proceedings of the third international conference on Computational Intelligence in Medicine*

and Healthcare (CIMED) (E. Ifeachor and A. Anastasiou, eds.), pp. 1–8, 2007.

85. Van Belle V., Pelckmans K., Suykens J.A.K., Van Huffel S., "Survival SVM: a Practical Scalable Algorithm", in *Proc. of the 16th European Symposium on Artificial Neural Networks (ESANN2008)*, Bruges, Belgium, Apr. 2008, pp. 89-94.
86. Van Belle V., Pelckmans K., Suykens J.A.K., Van Huffel S., "Learning Transformation Models for Ranking and Survival Analysis", *Journal of Machine Learning Research*, vol. 12, Mar. 2011, pp. 819-862.
87. V. Van Belle, K. Pelckmans, S. Van Huffel, J. A. K. Suykens; Improved performance on high-dimensional survival data by application of Survival-SVM, *Bioinformatics*, 27, 1, 2011, 87–94.
88. Van Belle V., Pelckmans K., Van Huffel S., Suykens J.A.K., "Support vector methods for survival analysis: a comparison between ranking and regression approaches", *Artificial Intelligence in Medicine*, vol. 53, no. 2, Oct. 2011, pp. 107-118.
89. V. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, New York, NY, 1998.
90. Vinzamuri, Bhanukiran, Yan Li, and Chandan K. Reddy. "Active Learning based Survival Regression for Censored Data." *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM). ACM*, 2014.
91. J. Wang, L. Li, D. Niu and Z. Tan, An annual load forecasting model based on support vector regression with differential evolution algorithm. *Applied Energy*, 2012. 94: pp. 65-70.
92. P. Wang, Y. Li and C.K. Reddy, Machine Learning for Survival Analysis: A Survey, ACM Comput. Surv. Article 1(1) (2017), 38.
93. H.Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, 2009. ISBN 978-0-387-98140-6. URL http://ggplot2.org. [p420]
94. H. Yijun, Y. Shu, H. Minghuang, Y. Xiaowei, Q. Fang, et al., Application of support vector machine to predict 5-year survival status of patients with Nasopharyngeal Carcinoma after treatment. The Chinese-German *Journal of Clinical Oncology*, 2006. 5(1): pp. 8-12.

95. Zupan B., Demšar J., Kattan M.W., Beck J.R., Bratko I. (1999) Machine Learning for Survival Analysis: A Case Study on Recurrence of Prostate Cancer. In: Horn W., Shahar Y., Lindberg G., Andreassen S., Wyatt J. (eds) *Artificial Intelligence in Medicine. AIMDM* 1999. Lecture Notes in Computer Science, vol 1620. Springer, Berlin, Heidelberg.