ΑΡΙΣΤΟΤΕΛΕΙΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΟΝΙΚΗΣ

Τμημα ΗΜΜΥ. Τομεας Ηλεκτρονικης

Ασαφη Συστηματα (Υπολογιστικη Νοημοσυνη)

Εργασία 4

Επίλυση προβλήματος ταξινόμησης με χρήση μοντέλων TSK

Συντάκτης:
Χρήστος Χουτουρίδης [8997]
cchoutou@ece.auth.gr

Διδάσκοντες:
Θεοχάρης Ιωάννης
theochar@ece.auth.gr

Χαδουλός Χρήστος
christgc@auth.gr

26 Οκτωβρίου 2025

1. Εισαγωγη

Η παρούσα εργασία εστιάζει στην επίλυση προβλημάτων **ταξινόμησης** μέσω ασαφών συστημάτων τύπου **Takagi–Sugeno–Kang (TSK)**. Η ταξινόμηση αποτελεί μια από τις βασικότερες εφαρμογές της υπολογιστικής νοημοσύνης, καθώς συνδυάζει στοιχεία λογικής, στατιστικής και μηχανικής μάθησης για τη λήψη αποφάσεων με αβεβαιότητα. Τα μοντέλα TSK προσφέρουν ένα ευέλικτο πλαίσιο, στο οποίο η γνώση αναπαρίσταται με τη μορφή κανόνων τύπου *IF–THEN*, ενώ η εκπαίδευση των παραμέτρων μπορεί να γίνει με υβριδικούς αλγορίθμους (Least Squares + Gradient Descent) όπως ο anfis.

Η εργασία χωρίζεται σε δύο πειραματικά σενάρια:

- **Σενάριο 1**: Εφαρμογή σε ένα απλό, χαμηλής διαστασιμότητας dataset (Haberman), με στόχο τη σύγκριση των προσεγγίσεων *class-independent* και *class-dependent Subtractive Clustering*.
- **Σενάριο 2**: Εφαρμογή σε πιο σύνθετο πρόβλημα ταξινόμησης με δεδομένα υψηλής διαστασιμότητας, όπου αξιολογούνται τεχνικές προ-επεξεργασίας και επιλογής χαρακτηριστικών.

Μέσω των δύο σεναρίων εξετάζεται η διαδικασία εκπαίδευσης, αξιολόγησης και βελτιστοποίησης ενός ασαφούς συστήματος, καθώς και ο τρόπος με τον οποίο η παραμετροποίηση επηρεάζει την απόδοση, την ερμηνευσιμότητα και την υπολογιστική πολυπλοκότητα.

1.1. Παραδοτέα

Τα παραδοτέα της εργασίας αποτελούνται από:

- Την παρούσα αναφορά.
- Τον κατάλογο **source**, με τον κώδικα της Matlab.
- Το σύνδεσμο με το αποθετήριο που περιέχει τον κώδικα της Matlab καθώς και αυτόν της αναφοράς.

2. Υλοποιηση

Για την υλοποίηση ακολουθήθηκε μια κοινή ροή επεξεργασίας και ανάλυσης δεδομένων, με στόχο τη μέγιστη επαναχρησιμοποίηση κώδικα μεταξύ των δύο σεναρίων. Συγκεκριμένα, δημιουργήθηκαν ανεξάρτητες συναρτήσεις για:

1. Τον διαχωρισμό των δεδομένων (*split_ data*).
2. Την προ-επεξεργασία (*preprocess_ data*).
3. Την αξιολόγηση των αποτελεσμάτων (*evaluate_ classification*).
4. Και τη συστηματική εμφάνιση/αποθήκευση γραφημάτων (*plot_ results1, plot_ results2*).

Έτσι, η εκτέλεση κάθε σεναρίου πραγματοποιείται αυτόνομα μέσω των scripts `scenario1.m` και `scenario2.m`, τα οποία συντονίζουν τα παραπάνω στάδια. Η δομή αυτή εξασφαλίζει καθαρότερο κώδικα, ευκολότερη συντήρηση και πλήρη αναπαραγωγιμότητα των αποτελεσμάτων.

2.1. Διαχωρισμός δεδομένων — *split_ data()*

Η συνάρτηση *split_ data()* αναλαμβάνει να διαχωρίσει το αρχικό dataset σε τρία υποσύνολα: εκπαίδευσης, ελέγχου (validation) και δοκιμής (test). Η διαδικασία βασίζεται σε τυχαία δειγματοληψία με καθορισμένο seed, ώστε τα ίδια σύνολα να μπορούν να αναπαραχθούν σε επόμενες εκτελέσεις. Ο διαχωρισμός είναι στρωματοποιημένος, ώστε η αναλογία των κλάσεων να παραμένει ίδια σε όλα τα υποσύνολα.

2.2. Προ-επεξεργασία δεδομένων — *preprocess_ data()*

Η προ-επεξεργασία εφαρμόζεται για να βελτιωθεί η σταθερότητα της εκπαίδευσης και να εξισορροπηθούν οι τιμές των εισόδων. Στο πλαίσιο της εργασίας επιλέχθηκε κανονικοποίηση τύπου **z-score**, δηλαδή:

$$x_i' = \frac{x_i - \mu_i}{\sigma_i}$$

όπου μ_i και σ_i είναι ο μέσος και η τυπική απόκλιση κάθε χαρακτηριστικού. Η συνάρτηση *preprocess_ data()* εφαρμόζει αυτόματα την κανονικοποίηση στα train/validation/test σύνολα και επιστρέφει τα απαραίτητα στατιστικά για τυχόν επαναφορά στις αρχικές κλίμακες.

2.3. Αξιολόγηση αποτελεσμάτων — *evaluate_ classification()*

Η συνάρτηση *evaluate_ classification()* συγκρίνει τις προβλέψεις του μοντέλου με τις πραγματικές ετικέτες και υπολογίζει τις μετρικές αξιολόγησης (OA, PA, UA, κ). Επιπλέον, παράγει και επιστρέφει τη μήτρα σύγχυσης. Η υλοποίηση έγινε ώστε να υποστηρίζει οποιονδήποτε αριθμό κλάσεων, ενώ παρέχει και normalized εκδόσεις των μετρικών, διευκολύνοντας τη συγκριτική ανάλυση μεταξύ μοντέλων.

2.4. Απεικόνιση αποτελεσμάτων

Για τη γραφική παρουσίαση των αποτελεσμάτων δημιουργήθηκαν οι συναρτήσεις *plot_ results1()* και *plot_ - results2()*, οι οποίες παράγουν και αποθηκεύουν αυτόματα όλα τα ζητούμενα γραφήματα κάθε σεναρίου. Κάθε γράφημα αποθηκεύεται σε υποκατάλογο (`figures_scn1`, `figures_scn2`) με συνεπή ονοματολογία, ώστε να διευκολύνεται η ενσωμάτωσή τους στην αναφορά. Η σχεδίαση περιλαμβάνει μήτρες σύγχυσης, καμπύλες εκπαίδευσης, συναρτήσεις συμμετοχής πριν και μετά την εκπαίδευση, καθώς και συγκεντρωτικά διαγράμματα OA–κ και Rules–Accuracy.

Το επόμενο τμήμα παρουσιάζει το πρώτο σενάριο και τα αποτελέσματά του, εφαρμόζοντας τη μεθοδολογία που περιγράφηκε παραπάνω.

3. Σεναριο 1 — Εφαρμογη σε απλο Dataset

Το πρώτο σενάριο αφορά την επίλυση ενός προβλήματος **δυαδικής ταξινόμησης** χρησιμοποιώντας μοντέλα TSK με σταθερές (singleton) εξόδους. Ως dataset χρησιμοποιήθηκε το *Haberman's Survival Dataset*, το οποίο περιλαμβάνει 306 δείγματα με τρεις αριθμητικές εισόδους (ηλικία, έτος εγχείρησης, αριθμός λεμφαδένων) και μία δυαδική ετικέτα που δηλώνει αν ο ασθενής επέζησε για τουλάχιστον πέντε έτη μετά την εγχείρηση.

Ο στόχος είναι η εκπαίδευση και αξιολόγηση μοντέλων TSK που προκύπτουν μέσω **Subtractive Clustering (SC)** τόσο στην *class-independent* όσο και στην *class-dependent* παραλλαγή, καθώς και η διερεύνηση της επίδρασης του παραμέτρου ακτίνας r_a στην πολυπλοκότητα και την απόδοση του συστήματος.

3.1. Πειραματική διαδικασία

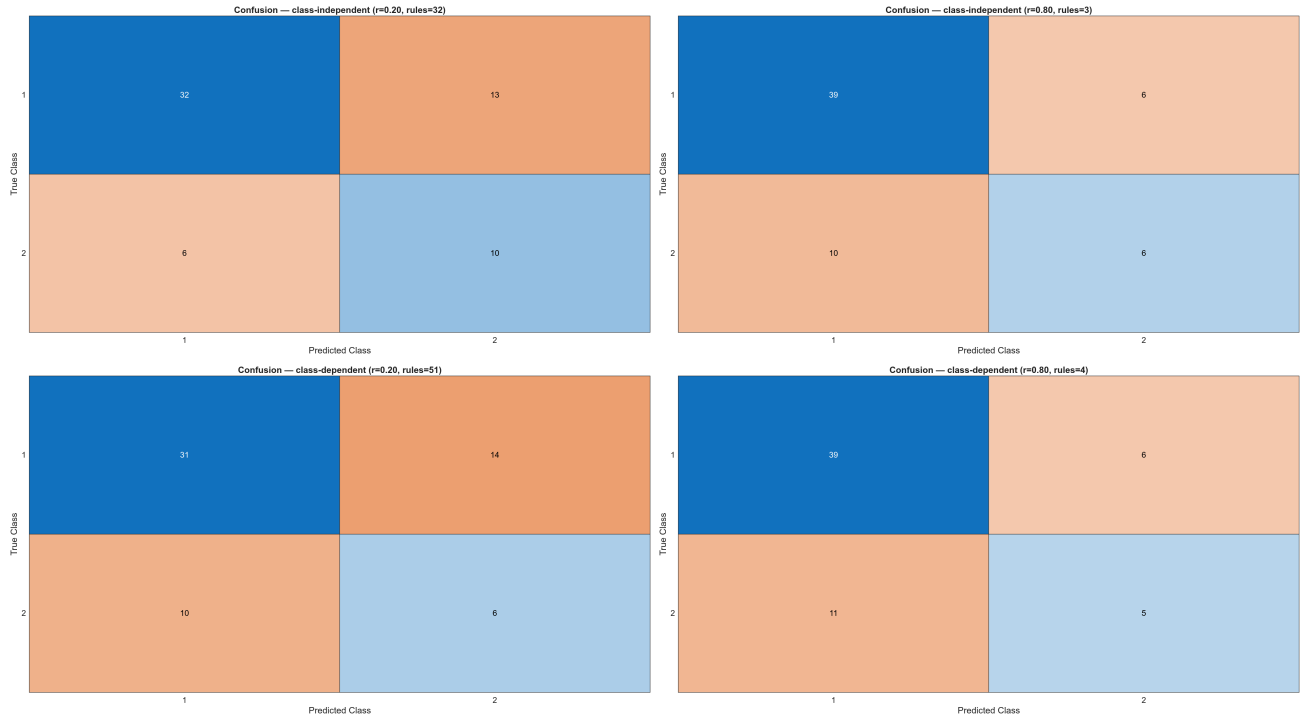
Το σύνολο δεδομένων χωρίστηκε στρωματοποιημένα σε τρία υποσύνολα (60 % train, 20 % validation, 20 % test). Η κανονικοποίηση των εισόδων έγινε με **z-score**, ενώ οι ετικέτες παρέμειναν ακέραιες. Για κάθε mode (class-independent και class-dependent) δοκιμάστηκαν δύο τιμές ακτίνας, r_a=$\{0.20, 0.80\}$, ώστε να προκύψουν τέσσερα μοντέλα συνολικά. Η εκπαίδευση πραγματοποιήθηκε με **anfis** για 100 εποχές και με validation έλεγχο για αποφυγή υπερεκπαίδευσης.

Οι μετρικές αξιολόγησης που χρησιμοποιήθηκαν ήταν:

- **Overall Accuracy (OA)**,
- **Producer's Accuracy (PA)** και **User's Accuracy (UA)** ανά κλάση,
- και ο **συντελεστής συμφωνίας κ του Cohen**.

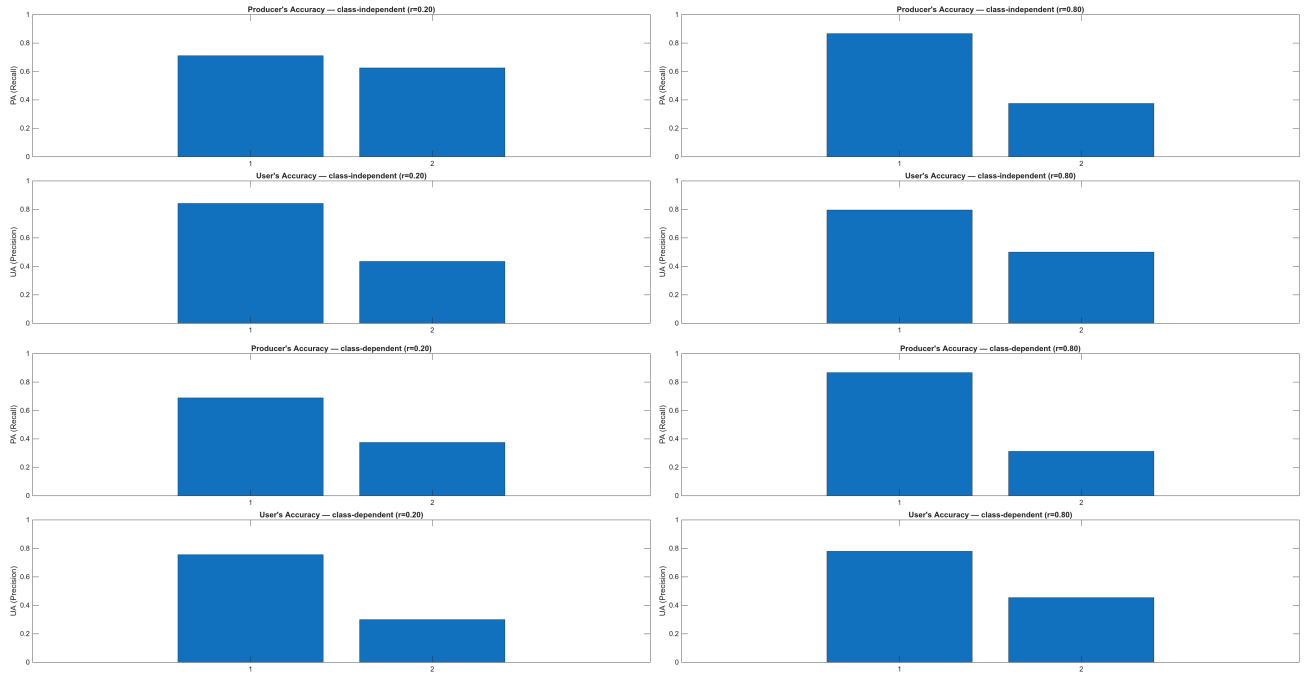
3.2. Αποτελέσματα

Μήτρες σύγχυσης. Το σχήμα 1 παρουσιάζει τις μήτρες σύγχυσης και για τα τέσσερα μοντέλα. Και στις δύο προσεγγίσεις, η απόδοση για την κλάση 1 είναι σταθερά υψηλότερη, γεγονός που σχετίζεται με την ανισορροπία του dataset (˜73 % δείγματα κλάσης 1). Η αύξηση της ακτίνας μειώνει τον αριθμό κανόνων και οδηγεί σε απλούστερα μοντέλα χωρίς σημαντική απώλεια ακρίβειας.



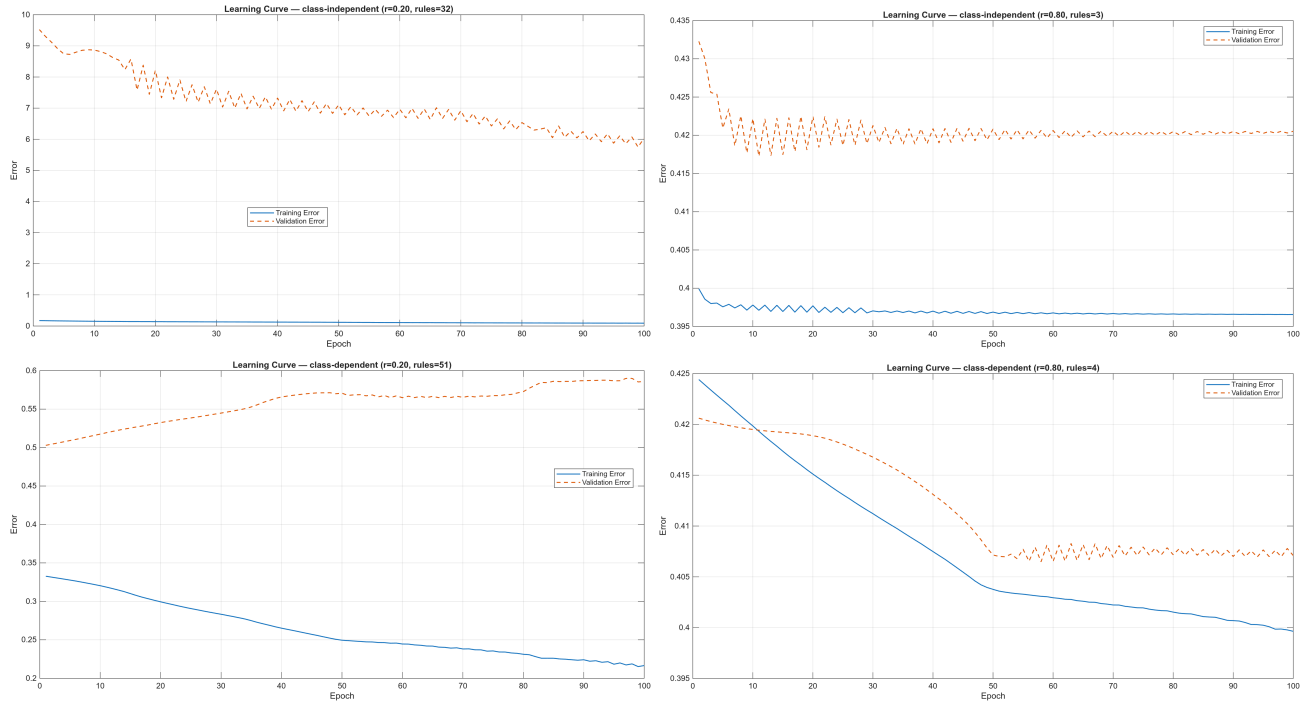
Σχήμα 1: *Μήτρες σύγχυσης για όλα τα μοντέλα (class-independent και class-dependent,* r_a=0.20, 0.80*).*

Μετρικές PA/UA. Το σχήμα 2 απεικονίζει την ακρίβεια παραγωγού (PA) και χρήστη (UA) ανά κλάση για κάθε μοντέλο. Η ανισορροπία των δεδομένων οδηγεί σε χαμηλότερες τιμές PA/UA για τη μειοψηφούσα κλάση 2, ωστόσο η συνολική τάση παραμένει σταθερή μεταξύ των modes.



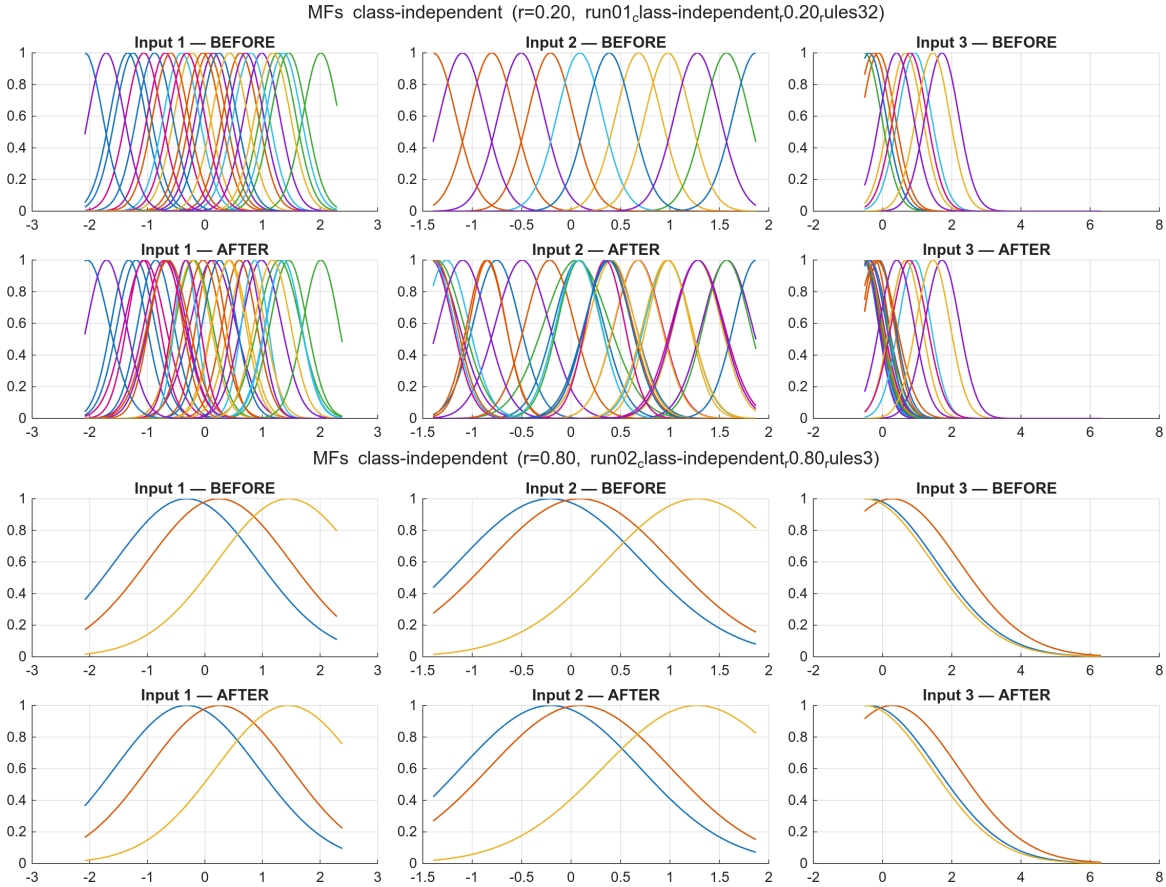
Σχήμα 2: *PA (recall) και UA (precision) ανά κλάση για όλα τα μοντέλα.*

Καμπύλες εκπαίδευσης. Οι καμπύλες εκπαίδευσης της Εικόνας 3 δείχνουν τη σύγκλιση του ANFIS για κάθε περίπτωση. Για μικρό r_a, το training error μηδενίζεται γρήγορα (υπερεκπαίδευση), ενώ για μεγάλο r_a η εκπαίδευση είναι πιο ομαλή και το validation error σταθεροποιείται χαμηλότερα, δείχνοντας καλύτερη γενίκευση.

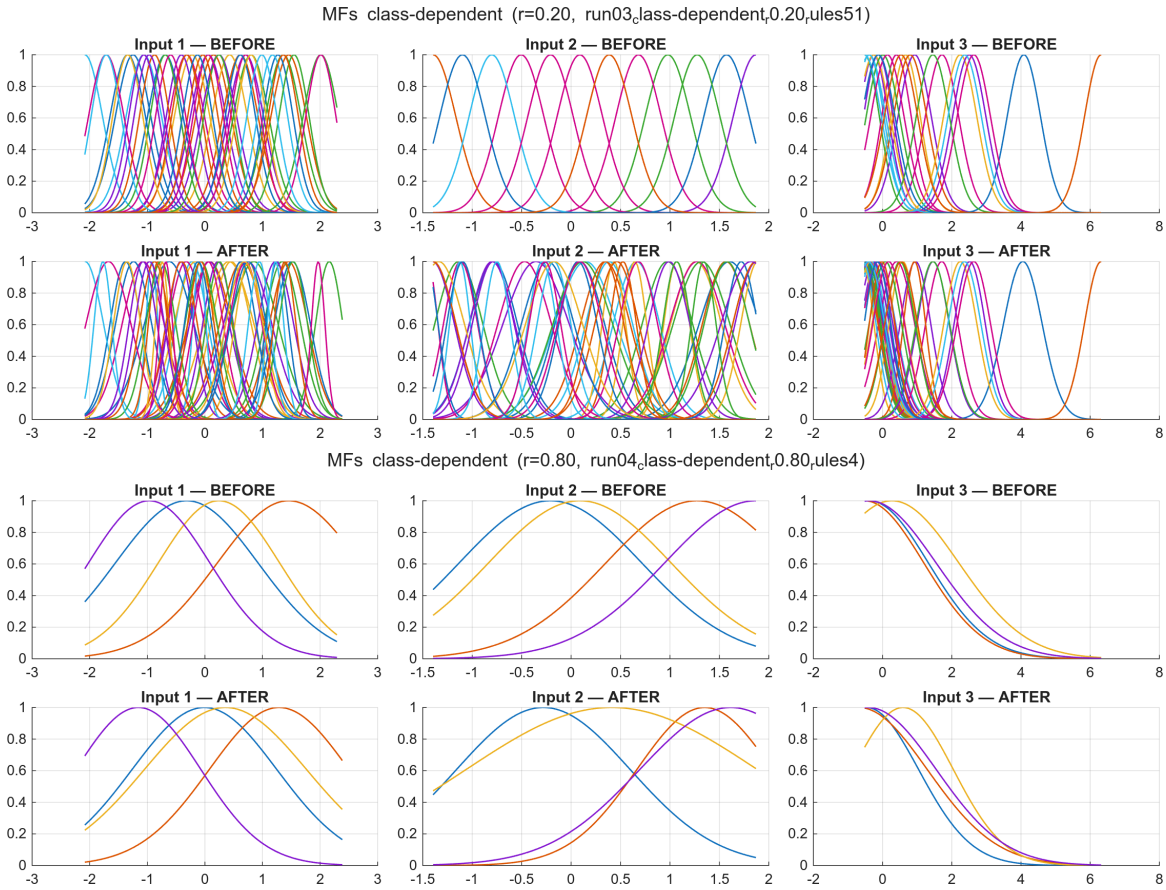


Σχήμα 3: *Καμπύλες εκπαίδευσης (training / validation error vs epoch) για όλα τα μοντέλα.*

Συναρτήσεις συμμετοχής. Τα σχήματα 4 και 5 απεικονίζει τις MFs πριν και μετά την εκπαίδευση για όλα τα μοντέλα. Για μικρό r_a, παρατηρείται υπερβολική επικάλυψη και πλήθος κανόνων· για μεγάλο r_a λιγότερες MFs και πιο ομαλές μεταβάσεις. Η μεταβολή των MFs επιβεβαιώνει τη σωστή προσαρμογή των παραμέτρων στο σύνολο εκπαίδευσης.

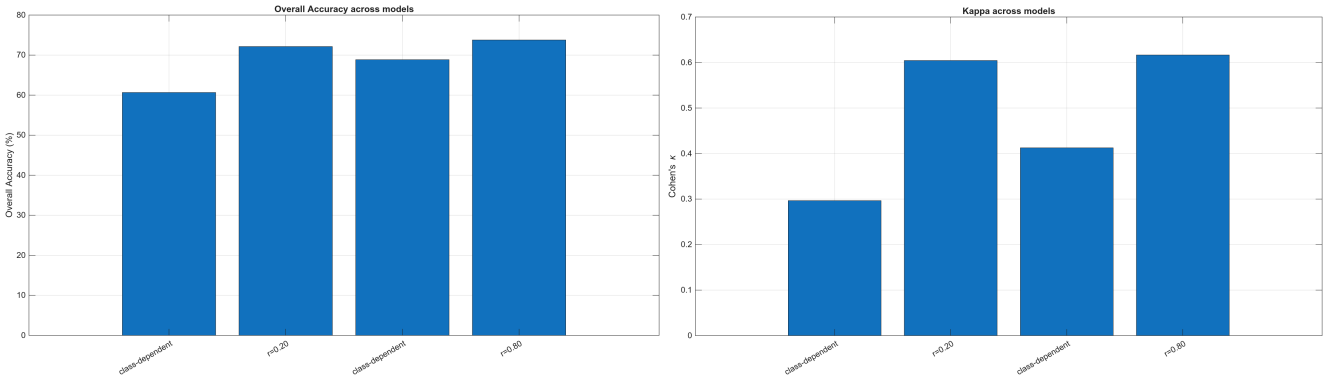


Σχήμα 4: *Συναρτήσεις συμμετοχής (before/after) για τα class-independent μοντέλα .*

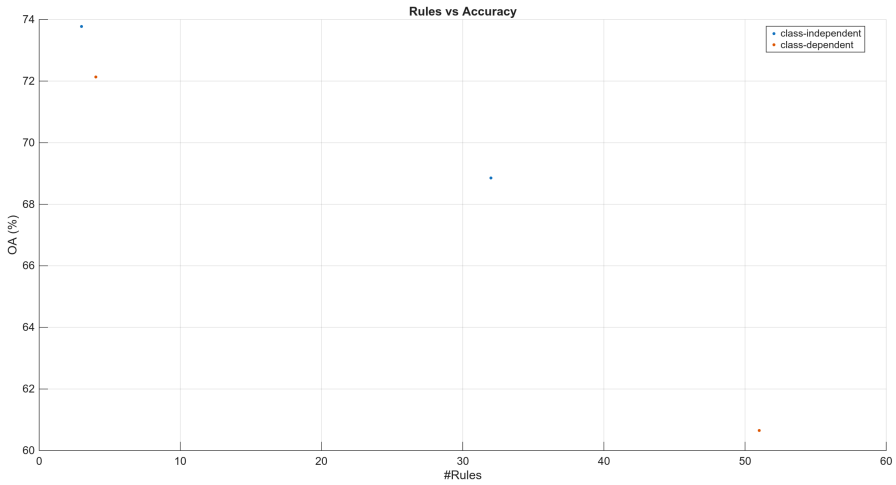


Σχήμα 5: *Συναρτήσεις συμμετοχής (before/after) για τα class-dependent μοντέλα.*

Συνολικές μετρικές και συσχετίσεις. Το σχήμα 6 παρουσιάζει την OA και κ για κάθε μοντέλο, ενώ το σχήμα 7 δείχνει τη σχέση αριθμού κανόνων – ακρίβειας. Παρατηρείται ότι η μέγιστη OA ($\approx 73\%$) και ο υψηλότερος κ προκύπτουν στα μοντέλα με r_a=0.80 (3–4 κανόνες), επιβεβαιώνοντας ότι η απλούστευση δεν μειώνει την απόδοση.



Σχήμα 6: *Συνολική OA και συντελεστής κ για όλα τα μοντέλα.*



Σχήμα 7: *Σχέση αριθμού κανόνων και Overall Accuracy.*

3.3. Συμπεράσματα

Συνοψίζοντας:

- Οι TSK μοντελοποιήσεις μέσω SC παρείχαν σταθερή ταξινόμηση ακόμη και σε ανισόρροπα δεδομένα.
- Η αύξηση του αριθμού κανόνων δεν βελτίωσε σημαντικά την απόδοση αλλά μείωσε την ερμηνευσιμότητα.
- Η class-dependent εκδοχή δεν υπερείχε σημαντικά, υποδεικνύοντας μικρή διακριτότητα των κλάσεων.
- Οι καμπύλες εκμάθησης και οι MFs δείχνουν ομαλή εκπαίδευση χωρίς αστάθειες.
- Το **class-independent μοντέλο με** r_a=0.80 παρουσιάζει την καλύτερη ισορροπία μεταξύ ακρίβειας, απλότητας και γενίκευσης.

3.4. Επίλογος

Αναπτύξαμε