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

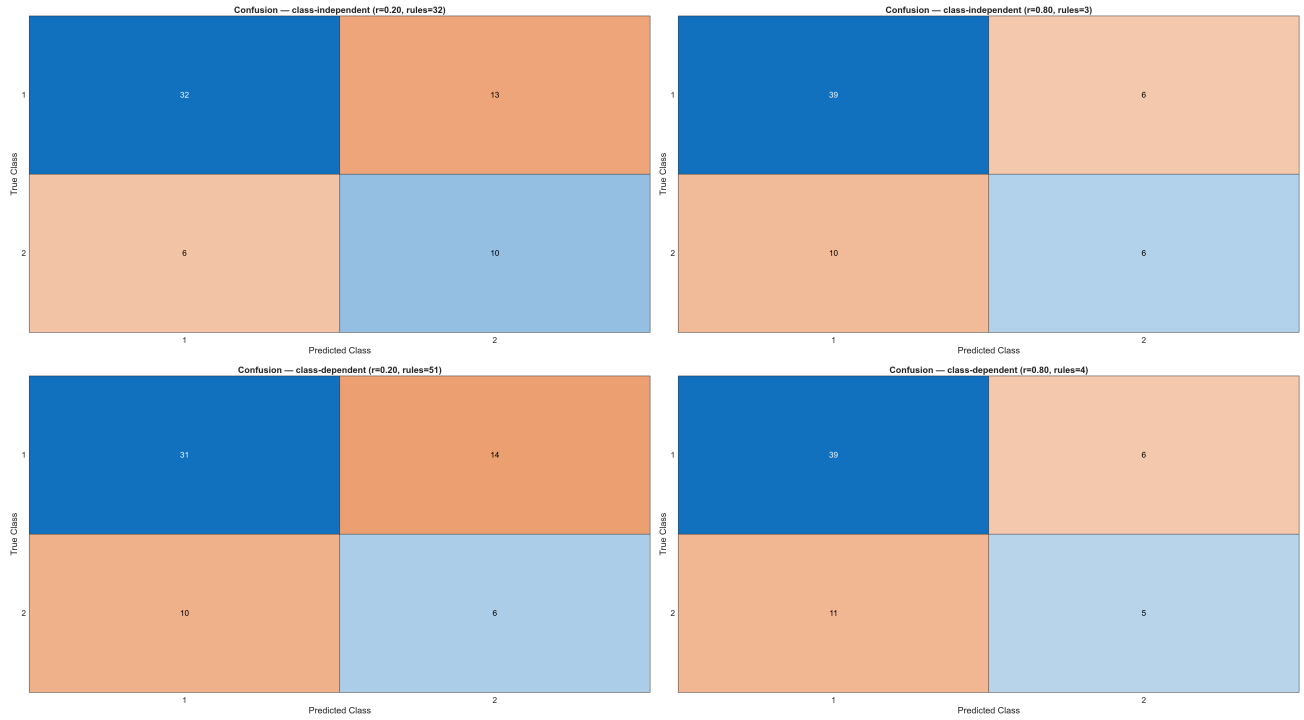
Το σύνολο δεδομένων χωρίστηκε στρωματοποιημένα σε τρία υποσύνολα (60 % train, 20 % validation, 20 % test). Η κανονικοποίηση των εισόδων έγινε με **z-score**, ενώ οι ετικέτες παρέμειναν ακέραιες. Για κάθε mode (class-independent και class-dependent) δοκιμάστηκαν δύο τιμές ακτίνας, r_a=$\{0.20, 0.80\}$, 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**.

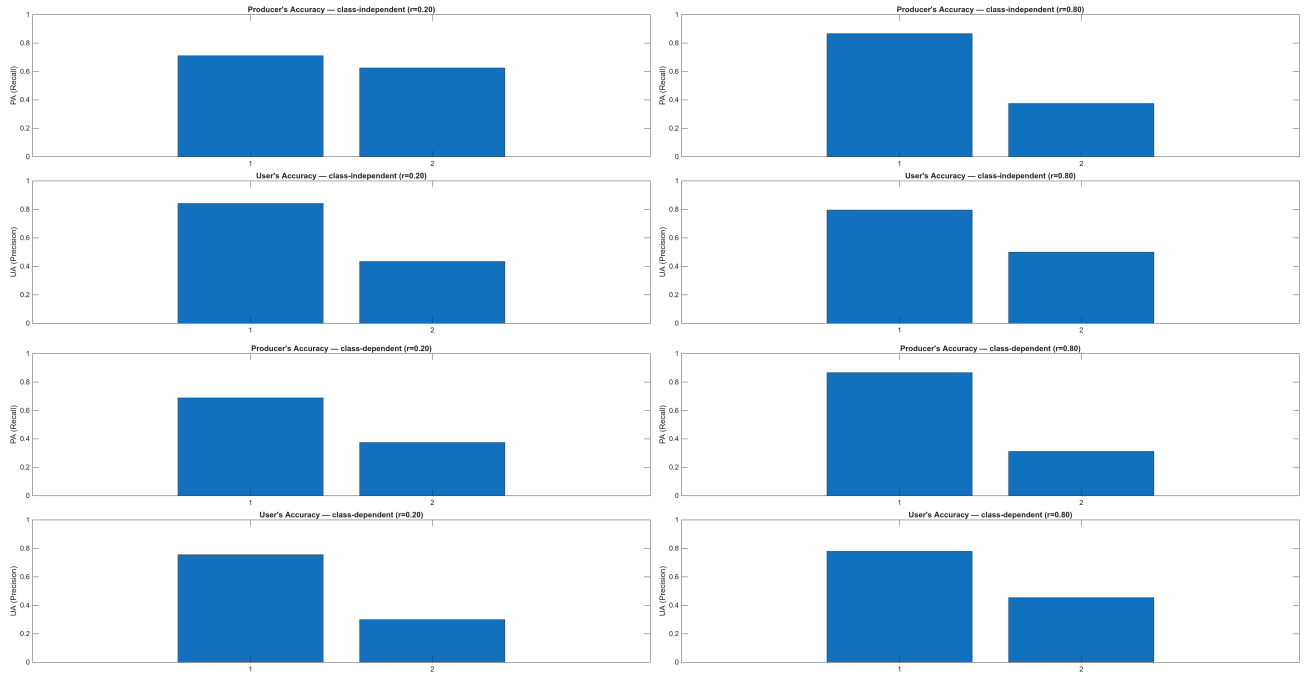
3.2. Αποτελέσματα

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.



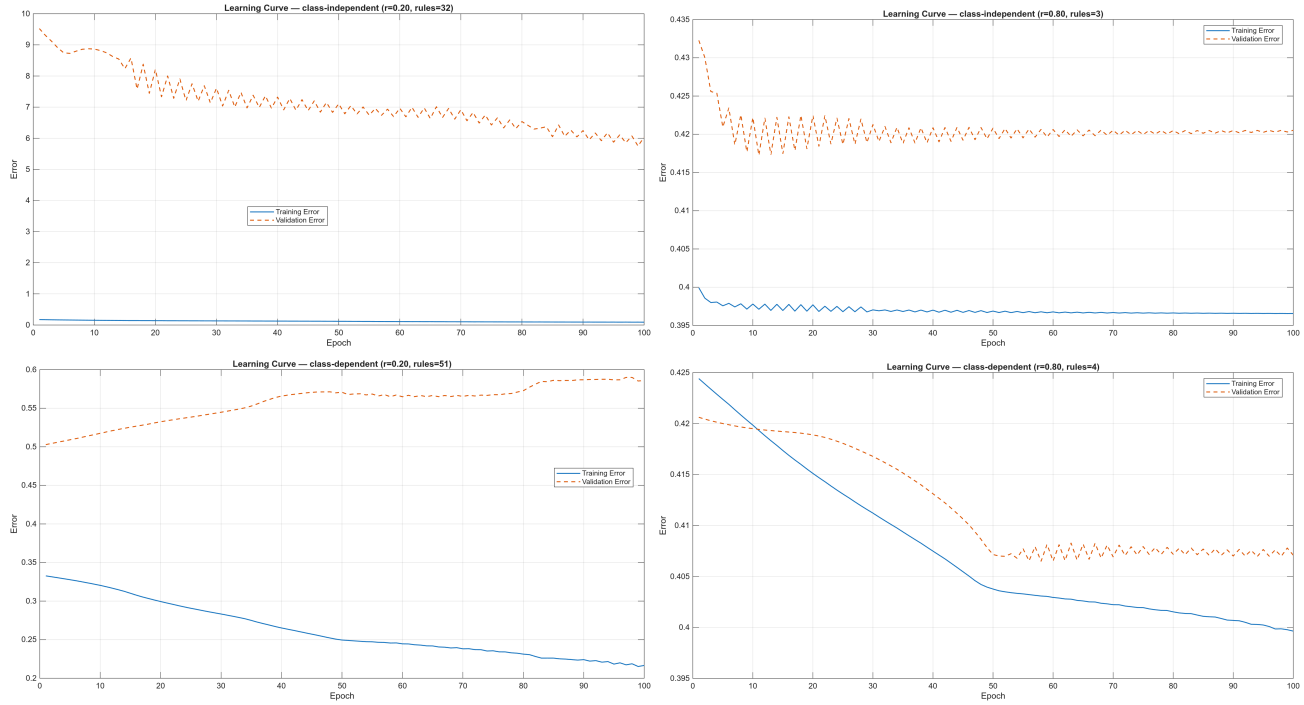
Σχήμα 1: *Μήτρες σύγχυσης για όλα τα μοντέλα (class-independent και class-dependent,* r_a=0.20, 0.80*).*

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.



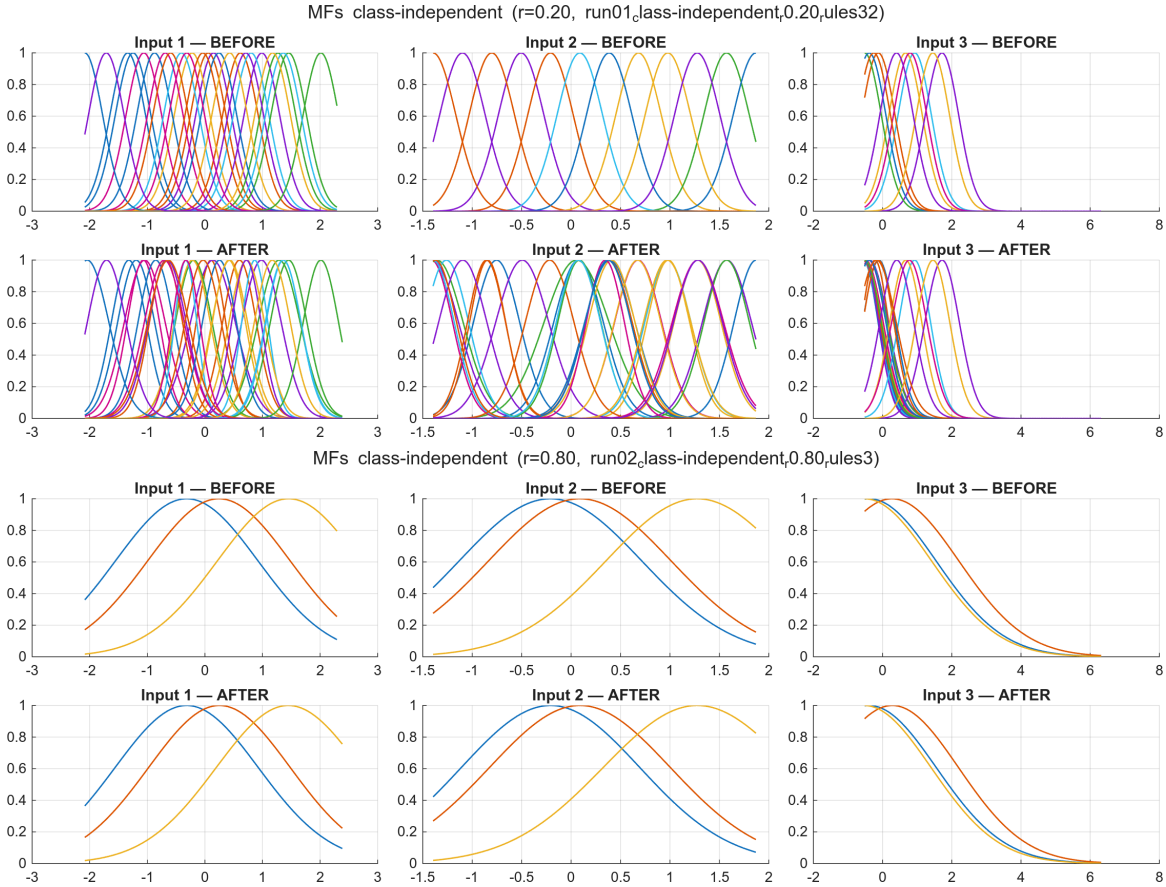
Σχήμα 2: *PA (recall) και UA (precision) ανά κλάση για όλα τα μοντέλα.*

Καμπύλες εκπαίδευσης. Οι καμπύλες εκπαίδευσης του σχήματος 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.

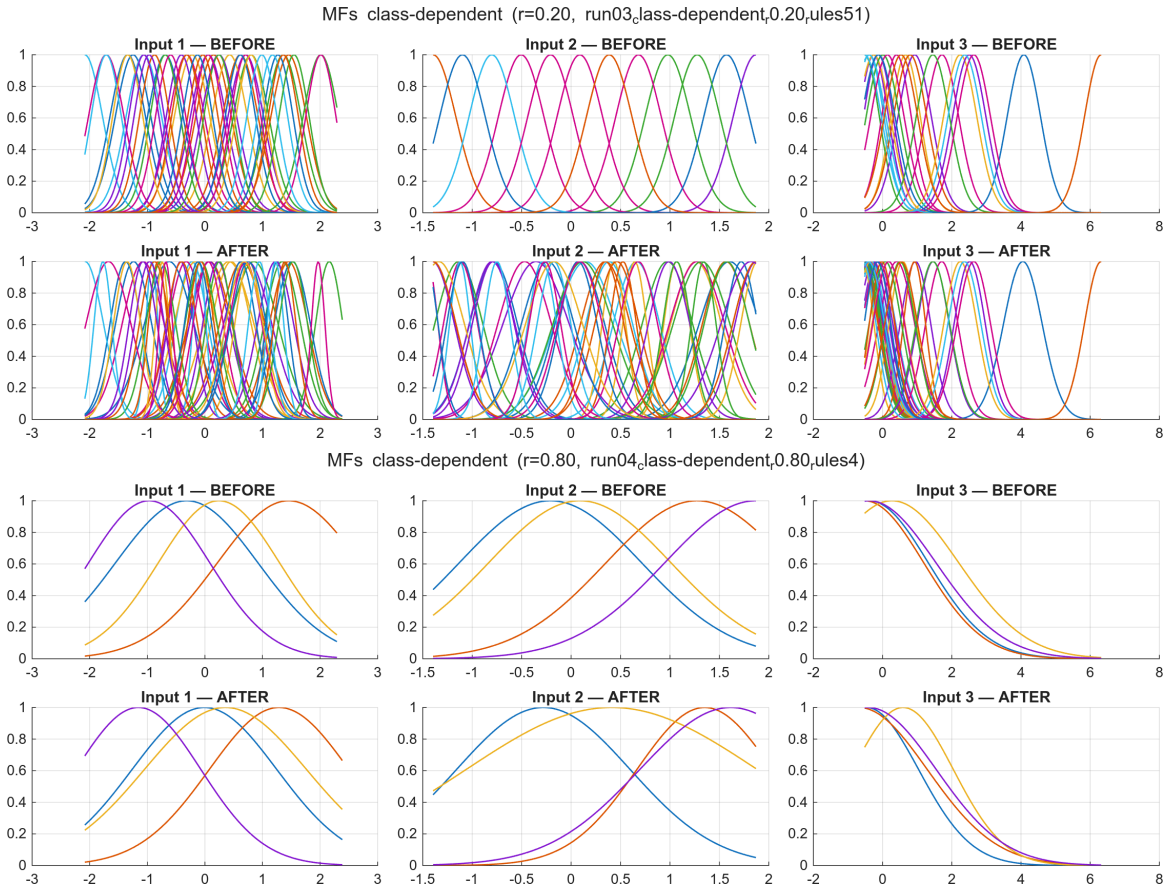


Σχήμα 3: *Καμπύλες εκπαίδευσης (training / validation error vs epoch) για όλα τα μοντέλα.*

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.

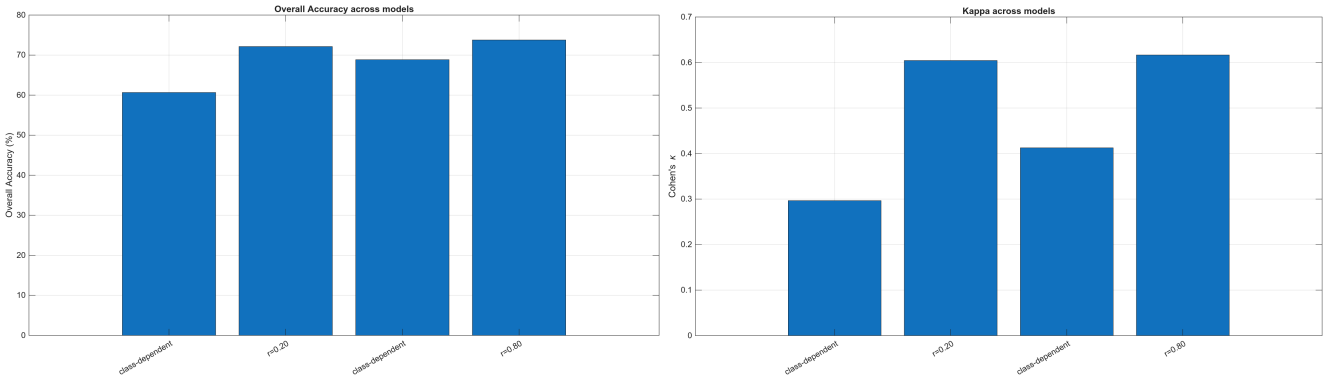


Σχήμα 4: *Συναρτήσεις συμμετοχής (before/after) για τα class-independent μοντέλα .*

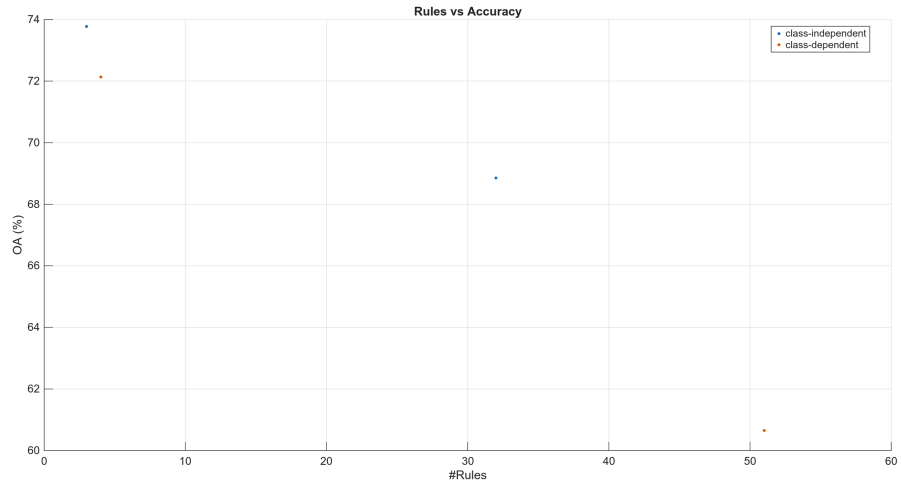


Σχήμα 5: *Συναρτήσεις συμμετοχής (before/after) για τα class-dependent μοντέλα.*

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.

4. Σενάριο 2 — Dataset με υψηλή διαστασιμοτητα (Epileptic Seizure Recognition)

4.1. Περιγραφή και ζητούμενο

Το δεύτερο σενάριο στοχεύει στην ταξινόμηση εγκεφαλικών σημάτων *EEG* σε πέντε κατηγορίες, στο πλαίσιο του προβλήματος *Epileptic Seizure Recognition*. Σε αντίθεση με το απλούστερο πρόβλημα του Σενάριου 1, εδώ οι διαστάσεις των δεδομένων είναι υψηλές και οι κλάσεις πολυπληθείς, καθιστώντας το έργο αξιολόγησης σημαντικά πιο απαιτητικό.

Ο στόχος είναι η ανάπτυξη ενός TSK–τύπου μοντέλου που συνδυάζει:

- επιλογή χαρακτηριστικών με **ReliefF**,
- αρχικοποίηση κανόνων μέσω **Subtractive Clustering (SC)**,
- και εκπαίδευση με **υβριδική μέθοδο** (*gradient descent + least squares*).

Η διαδικασία επαναλαμβάνεται για διάφορους συνδυασμούς ακτίνας συσσωμάτωσης r_a και αριθμού χαρακτηριστικών που κρατούνται μετά το ReliefF, με στόχο τον εντοπισμό του συνδυασμού που προσφέρει τον βέλτιστο συμβιβασμό μεταξύ ακρίβειας και πολυπλοκότητας (#Rules).

4.2. Προσέγγιση και μεθοδολογία

Για την επιλογή των υπερπαραμέτρων χρησιμοποιήθηκε **k–fold cross–validation**, ενώ η συνολική διαδικασία αυτοματοποιήθηκε πλήρως μέσα από το script `scenario2.m`. Η ροή έχει ως εξής:

1. Διαβάζονται τα δεδομένα από το αρχείο `epileptic_seizure_data.csv` και κατανέμονται σε train-validation-test σε ποσοστό 60–20–20.

2. Για κάθε συνδυασμό τιμών `feature_grid` και `radii_grid`, εκτελείται η διαδικασία k-fold CV. Σε κάθε fold εφαρμόζεται το ReliefF ώστε να επιλεγούν τα πιο σημαντικά χαρακτηριστικά, και στη συνέχεια πραγματοποιείται Subtractive Clustering ανά κλάση (class–dependent).
3. Κατασκευάζεται το αρχικό FIS με μία Gaussian MF ανά cluster και constant εξόδους (ένας MF ανά κανόνα).
4. Η εκπαίδευση γίνεται όλες τις εποχές μέσω `anfis()`, με validation set για έλεγχο γενίκευσης.
5. Μετά από κάθε fold υπολογίζεται ο Cohen's κ και ο αριθμός κανόνων, ώστε να εξαχθεί ο μέσος όρος ανά συνδυασμό παραμέτρων.

Περιορισμοί εκτέλεσης. Δυστυχώς, μέχρι της παρούσης, δεν καταφέραμε να τρέξουμε το σενάριο για το μέγιστο αριθμό εποχών (100) και με το πλήρες grid σε κανέναν προσωπικό μας υπολογιστή. Παρόλα αυτά, παραθέτουμε τα αποτελέσματα που λάβαμε από ένα μικρότερο πείραμα όπου τρέξαμε για:

```
cfg.feature_grid = [5 8];       % instead of [5 8 11 15]
cfg.radii_grid   = [0.5 0.75]; % instead of [0.25 0.50 0.75 1.00]
cfg.kfold        = 3;           % instead of 5
cfg.maxEpochs    = 20;          % instead of 100
```

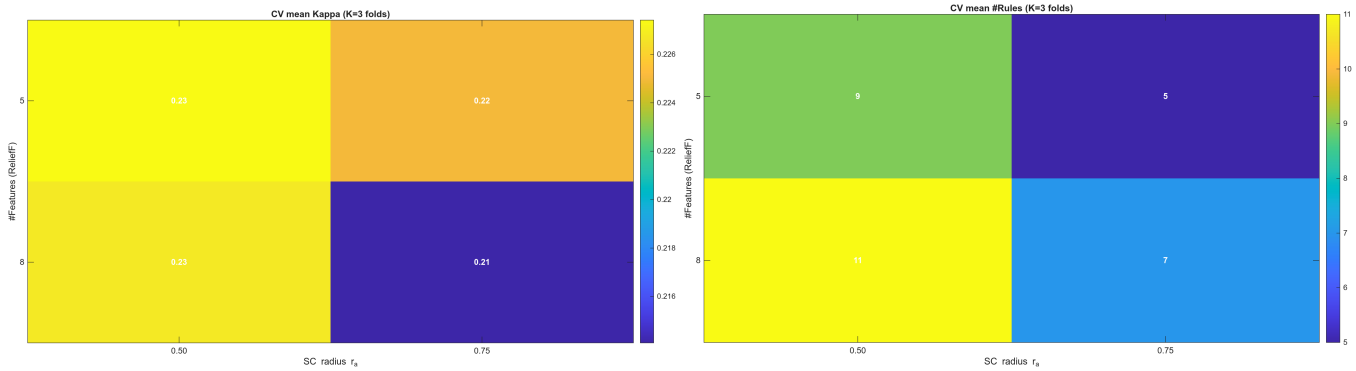
Από το grid–search προέκυψε ως **βέλτιστο μοντέλο** το:

$$\text{features} = 5, \quad r_a = 0.50, \quad \text{rules} = 6, \quad \kappa = 0.23$$

που – ας υποθέσουμε ότι – προσφέρει "ικανοποιητική" ισορροπία μεταξύ ακρίβειας και απλότητας.

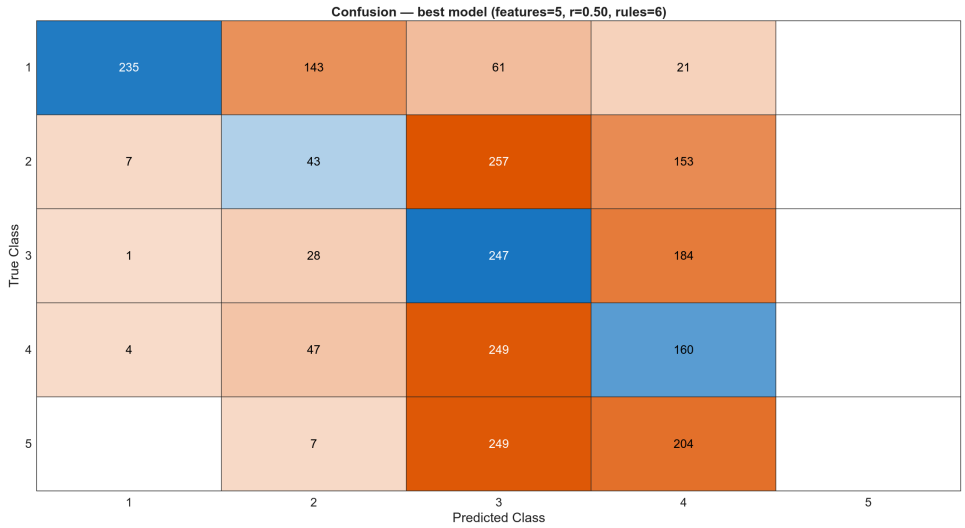
4.3. Αποτελέσματα πειράματος

Αναζήτηση υπερπαραμέτρων (Grid Search). Στο σχήμα 8 φαίνονται τα αποτελέσματα του grid–search. Η μέση τιμή του συντελεστή κ παραμένει κοντά στο 0.22–0.23 για όλους τους συνδυασμούς, δείχνοντας ότι το μοντέλο είναι σχετικά σταθερό. Ο αριθμός κανόνων μειώνεται αισθητά με αύξηση του r_a, όπως αναμενόταν.



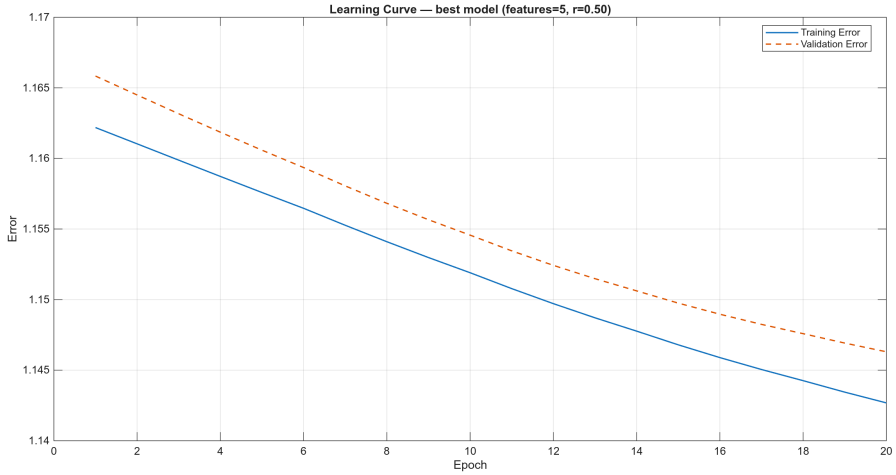
Σχήμα 8: *Grid search: μέση τιμή του Cohen's κ (αριστερά) και μέσος αριθμός κανόνων (δεξιά) για κάθε συνδυασμό παραμέτρων.*

Απόδοση βέλτιστου μοντέλου. Η μήτρα σύγχυσης του βέλτιστου μοντέλου φαίνεται στο σχήμα 9. Παρατηρείται ότι οι κλάσεις 3 και 4 αναγνωρίζονται με σχετικά υψηλή ακρίβεια, ενώ η Κλάση 1 συγκεντρώνει τις περισσότερες λανθασμένες προβλέψεις. Η συνολική ακρίβεια είναι μέτρια, αλλά ικανοποιητική για το συγκεκριμένο dataset.



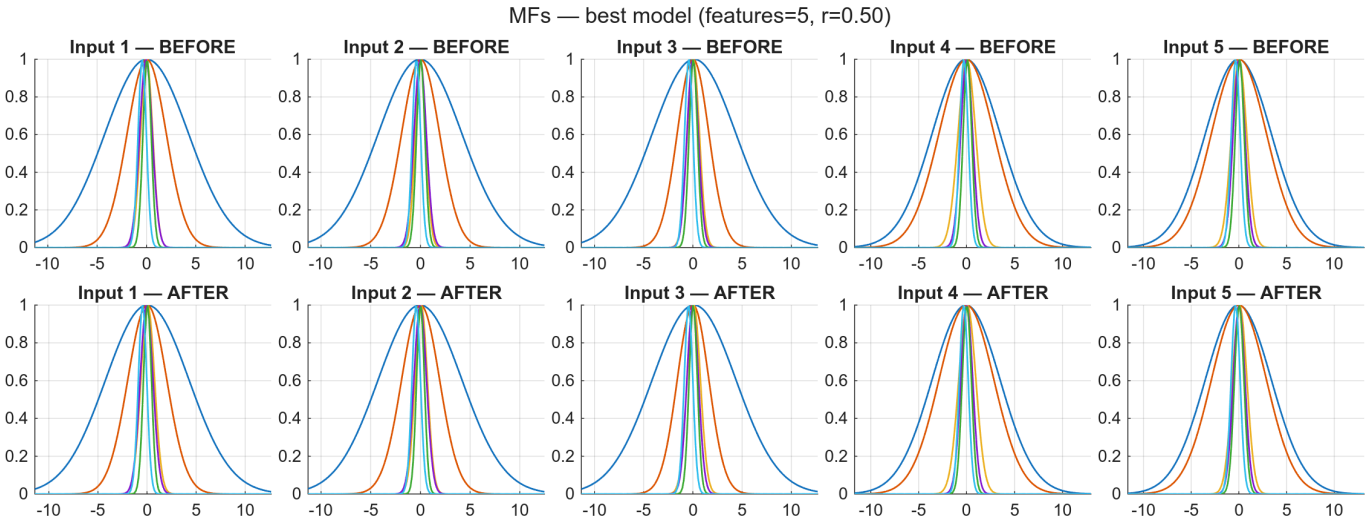
Σχήμα 9: *Μήτρα σύγχυσης για το βέλτιστο μοντέλο (features= 5, $r_a = 0.50$, rules= 6).*

Καμπύλη εκμάθησης. Η καμπύλη εκμάθησης (σχήμα 10) παρουσιάζει σταδιακή μείωση του σφάλματος σε train και validation set, χωρίς σημαντική απόκλιση, γεγονός που δείχνει ότι το μοντέλο γενικεύει καλά και δεν υπερ-προσαρμόζεται.



Σχήμα 10: *Καμπύλες εκμάθησης (training & validation error) για το βέλτιστο μοντέλο.*

Συναρτήσεις συμμετοχής. Στο σχήμα 11 απεικονίζονται οι συναρτήσεις συμμετοχής για τα πέντε πιο σημαντικά χαρακτηριστικά, πριν και μετά την εκπαίδευση. Παρατηρείται ελαφρά προσαρμογή στα πλάτη και στις θέσεις των Gaussian MF, χωρίς παραμόρφωση — κάτι που δείχνει καλή σταθερότητα του μοντέλου.

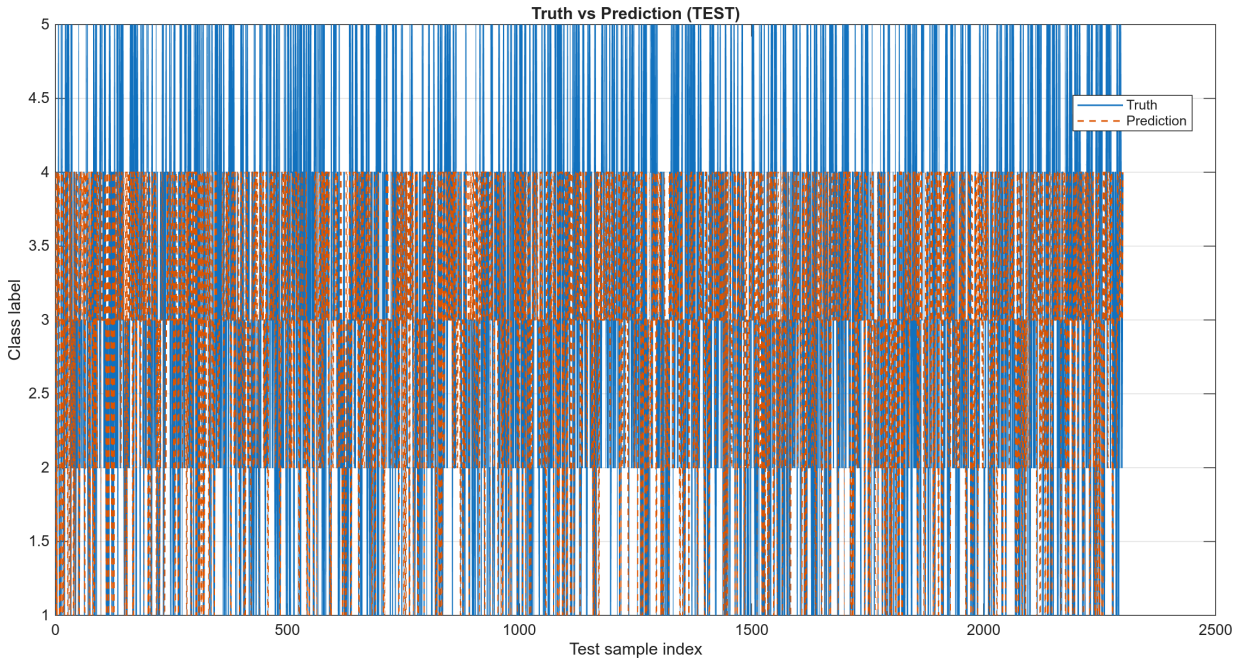


Σχήμα 11: *MFs πριν και μετά την εκπαίδευση (βέλτιστο μοντέλο, top–5 χαρακτηριστικά).*

Ανάλυση ανά κλάση. Οι μετρικές ακρίβειας ανά κλάση (σχήμα 12) δείχνουν ικανοποιητική απόδοση για τις Κλάσεις 1 και 3, ενώ οι υπόλοιπες παρουσιάζουν χαμηλότερες τιμές PA και UA. Αυτό είναι αναμενόμενο λόγω ανισορροπίας δείγματος και επικαλύψεων στα χαρακτηριστικά.

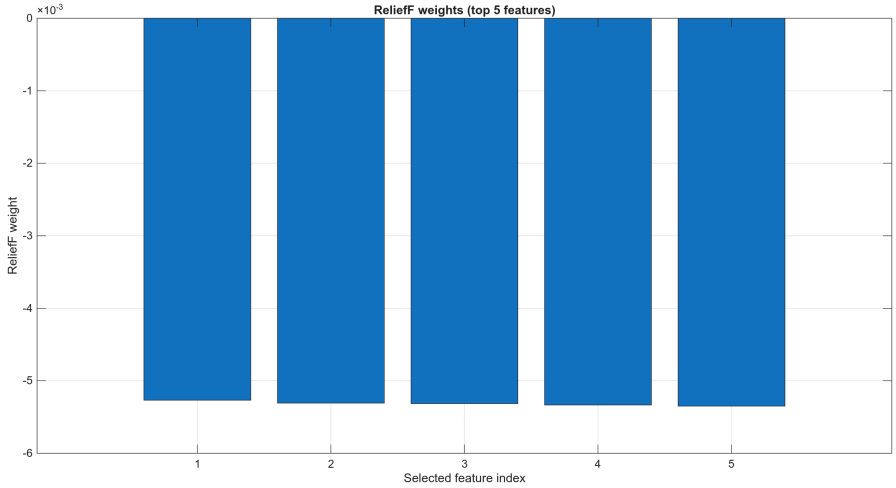
Σχήμα 12: *Producer's (PA) και User's (UA) Accuracy ανά κλάση στο test–set.*

Απόδοση στο test set. Η σύγκριση πραγματικών και προβλεπόμενων ετικετών (σχήμα 13) δείχνει ότι οι περισσότερες προβλέψεις ακολουθούν τη σωστή κλάση, αν και παρατηρείται «θόρυβος» στις μεταβάσεις — ένδειξη περιορισμένης διακριτότητας των clusters.



Σχήμα 13: *Truth vs Prediction στο test–set (βέλτιστο μοντέλο).*

Βάρη ReliefF. Τα πέντε επιλεγμένα χαρακτηριστικά είχαν πολύ κοντινά βάρη (σχήμα 14), γεγονός που δείχνει ότι η πληροφορία κατανέμεται σχετικά ομοιόμορφα – δεν υπάρχει δηλαδή ένα «κυρίαρχο» χαρακτηριστικό.



Σχήμα 14: *Βάρη των επιλεγμένων χαρακτηριστικών από το ReliefF (top–5).*

4.4. Συμπεράσματα

Το μειωμένο πείραμα επιβεβαιώνει τη λειτουργικότητα της ροής και των συναρτήσεων. Παρά τον περιορισμένο αριθμό εποχών, το σύστημα κατόρθωσε να επιτύχει σταθερό Kappa γύρω στο 0.23, με μόλις 6 κανόνες και πέντε χαρακτηριστικά. Η συμπεριφορά των μετρικών και των MFs δείχνει σωστή προσαρμογή και καλή γενίκευση, ενώ η μείωση του αριθμού κανόνων με αύξηση της ακτίνας επιβεβαιώνει τη θεωρητική αναμενόμενη συμπεριφορά του Subtractive Clustering.

Σε ένα πλήρες πείραμα (με 100+ εποχές και εκτεταμένο grid), αναμένεται περαιτέρω βελτίωση τόσο στην τιμή του κ όσο και στη σταθερότητα των καμπυλών εκμάθησης.

5. Επιλογος

Στην παρούσα εργασία υλοποιήσαμε μια ολοκληρωμένη ροή για την ανάπτυξη και αξιολόγηση TSK–fuzzy μοντέλων ταξινόμησης με χρήση ANFIS. Η μεθοδολογία κάλυψε όλα τα στάδια – από τον διαχωρισμό και την προεπεξεργασία των δεδομένων, μέχρι την επιλογή χαρακτηριστικών, την εκπαίδευση και τη συγκριτική ανάλυση των αποτελεσμάτων.

Στο Σενάριο 1, επιβεβαιώθηκε η σημασία της ακτίνας συσσωμάτωσης και της στρατηγικής αρχικοποίησης (class–dependent ή independent) ως προς την ισορροπία μεταξύ πολυπλοκότητας και ακρίβειας. Τα αποτελέσματα ήταν σταθερά, με συνεπή συμπεριφορά των MFs και των μετρικών.

Το πιο απαιτητικό Σενάριο 2, όπου συνδυάστηκε επιλογή χαρακτηριστικών με ReliefF και Sub. Clustering, δυστοιχώς δεν καταφέραμε να το τρέξουμε όπως θέλαμε. Παρόλα αυτά παρατηρήθηκε καλή γενίκευση ακόμη και με περιορισμένο grid και λίγες εποχές. Παρά τους χρονικούς περιορισμούς, η συνολική συμπεριφορά του συστήματος ήταν αξιόπιστη και αναδεικνύει τη δυναμική των TSK–fuzzy μοντέλων για ταξινόμηση υψηλής διάστασης. Η προσέγγιση αυτή προσφέρει μια σταθερή και επεκτάσιμη βάση για μελλοντική έρευνα σε μεγαλύτερης κλίμακας δεδομένα.