ΑΣΑΦΗ ΣΥΣΤΗΜΑΤΑ
ΥΠΟΛΟΓΙΣΤΙΚΗ ΝΟΗΜΟΣΥΝΗ
2020 - 2021

Εργασία 4 - Classification

Ονοματεπώνυμο: Παπαδάμ Στέφανος

ΑΕΜ: 8885

email: sgpapadam@ece.auth.gr

Ομάδα Εργασίας: -

Περιγραφή του προβλήματος

Η εργασία αυτή έχει σκοπό την διερεύνηση της ικανότητας των TSK μοντέλων στην επίλυση προβλημάτων ταξινόμησης μέσω ασαφών νευρωνικών μοντέλων. Για αυτό το λόγο επιλέγονται δύο σύνολα δεδομένων από τα οποία το πρώτο θα χρησιμοποιηθεί για μια διερεύνηση της εκπαίδευσης και αξιολόγησης των TSK μοντέλων, ενώ το δεύτερο το οποίο περιέχει περισσότερα χαρακτηριστικά και δείγματα θα χρησιμοποιηθεί για τον ίδιο σκοπό αλλά υλοποιώντας πιο σύνθετα βήματα όπως επιλογή χαρακτηριστικών και μεθόδους βελτιστοποίησης των μοντέλων μέσω διασταυρωμένης επικύρωσης.

Μέρος 1 - Haberman's Survival dataset

Περιγραφή dataset

Το πρώτο dataset το οποίο χρησιμοποιείται είναι το haberman's survival το οποίο περιλαμβάνει 306 δείγματα και 3 χαρακτηριστικά ενώ διαθέτει μία έξοδο που παίρνει τις τιμές 1 και 2. Τα χαρακτηριστικά με σειρά από τη στήλη 1 έως την 3 είναι τα εξής:

- ηλικία ασθενούς την χρονική στιγμή της εξέτασης.
- η χρονολογία που ο ασθενής πραγματοποίησε την επέμβαση.
- αριθμός θετικών axillary nodes που εντοπίστηκαν.

Η μοναδική έξοδος του dataset είναι η:

- κατάσταση επιβίωσης.
    - 1 σε περίπτωση που ο ασθενής έζησε τουλάχιστον 5 χρόνια.
    - 2 σε περίπτωση που ο ασθενής απεβίωσε μέσα σε 5 χρόνια.

Διαχωρισμός dataset

Αρχικά θα πρέπει να γίνει διαχωρισμός του dataset σε τρία σύνολα. Τα σύνολα αυτά είναι τα training, validation και check τα οποία χρησιμοποιούνται για εκπαίδευση, επικύρωση και έλεγχο αντίστοιχα. Τα μεγέθη τους είναι 60%, 20%, 20% του αρχικού συνόλου δεδομένων αντίστοιχα. Επειδή μας ζητείται το κάθε ένα από τα τρία σετ να περιέχει περίπου όμοια κατανεμημένα α δεδομένα, αρχικά διαχωρίζουμε τα δεδομένα στις 2 κλάσεις και αποθηκεύουμε στους πίνακες output1 και output2 τα δεδομένα των κλάσεων 1 και 2 αντίστοιχα. Έπειτα, με χρήση της dividerand χωρίζουμε το μήκος των output1 και output2 σε τρία διαστήματα (60%, 20%, 20%). Με βάση τα splits που προέκυψαν καθορίζουμε τα σύνολα training_data, validation_data και testing_data. Στη συνέχεια γίνεται ένα shufffle στα δεδομένα για να υπάρχει μια τυχαιότητα και ακολουθεί ο υπολογισμός του λόγου των δεδομένων της κλάσης 1 προς τα δεδομένα της κλάσης 2 που περιλαμβάνονται σε κάθε σετ αλλά και στο αρχικό σετ δεδομένων για να αποδειχθεί ότι είναι σχεδόν ίδιος. Το ποσοστό αυτό υπολογίζεται και στα 4 σετ δεδομένων (αρχικό, training, testing, validation) περίπου στο 36%.

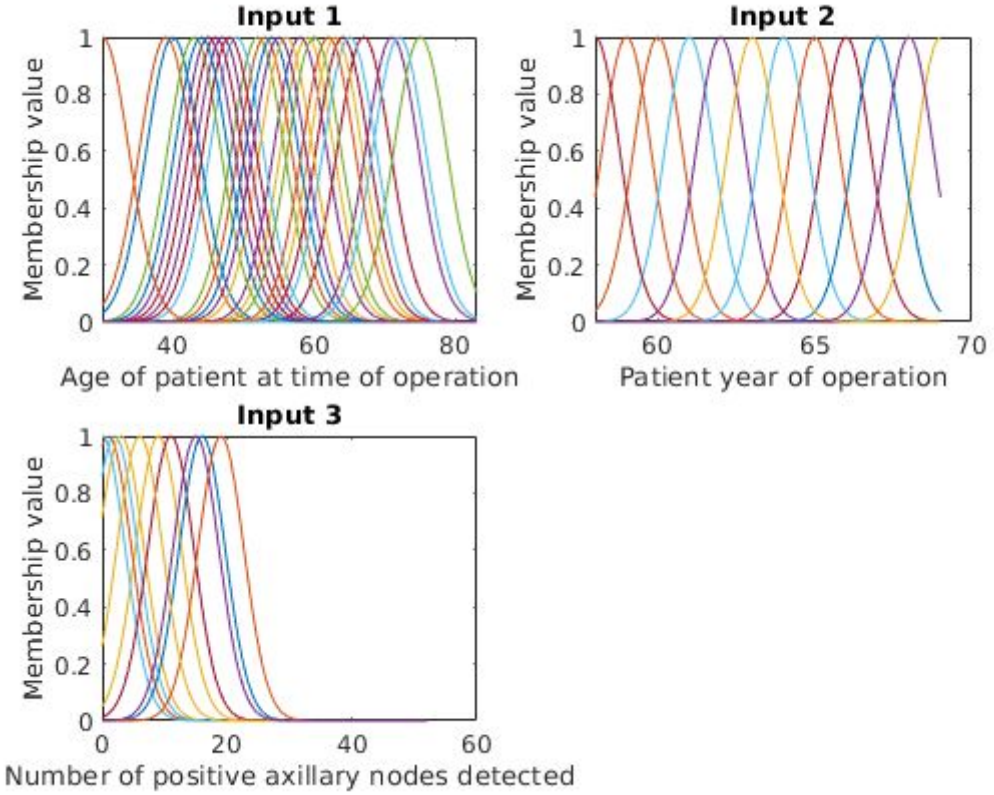
Εκπαίδευση TSK μοντέλων

Αφού πραγματοποιήθηκε ο διαχωρισμός στη συνέχεια υλοποιούμε την εκπαίδευση των τεσσάρων διαφορετικών μοντέλων που περιγράφονται στην εκφώνηση. Η διαμέριση του χώρου γίνεται με τη μέθοδο subtractive clustering. Στα δύο πρώτα μοντέλα το subtractive clustering εκτελείται για όλα τα δεδομένα του συνόλου εκπαίδευσης ενώ στα δύο επόμενα μοντέλα εκτελείται στα δεδομένα κάθε κλάσης ξεχωριστά. Τα δύο πρώτα μοντέλα υλοποιούνται με τη μέθοδο genfis2 ενώ για τα δύο επόμενα υλοποιείται η συνάρτηση classDep1 η οποία επιστρέφει το fis μοντέλο και βασίζεται στο αρχείο TSK_classification που μας παρέχετε. Στο πρόγραμμα μας αποθηκεύουμε κάθε μοντέλο στον πίνακα ‘m’ έτσι ώστε να μπορούμε να εκτελέσουμε την ίδια διαδικασία εκπαίδευσης τέσσερις φορές για κάθε μοντέλο. Οι εποχές που χρησιμοποιήθηκαν ήταν 100 και για την εκπαίδευση χρησιμοποιήθηκε η εντολή anfis ρυθμίζοντας τις παραμέτρους με την εντολή anfisOptions. Για την πρόβλεψη του αποτελέσματος χρησιμοποιείται η εντολή evalfis. Παρακάτω παρουσιάζονται τα αποτελέσματα των τεσσάρων μοντέλων.

Μοντέλο 1

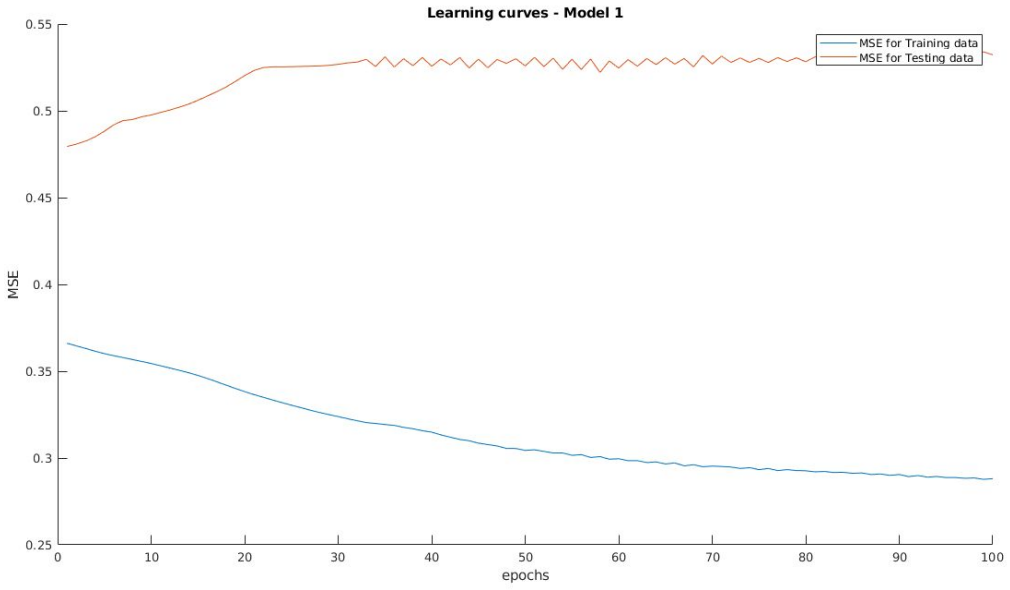
- Ακτίνα cluster: 0.2
- Ανεξάρτητη κλάσης
- κανόνες: 32

Συναρτήσεις Συμμετοχής



Εικόνα 1: Συναρτήσεις συμμετοχής των 3 εισόδων.

Καμπύλες μάθησης

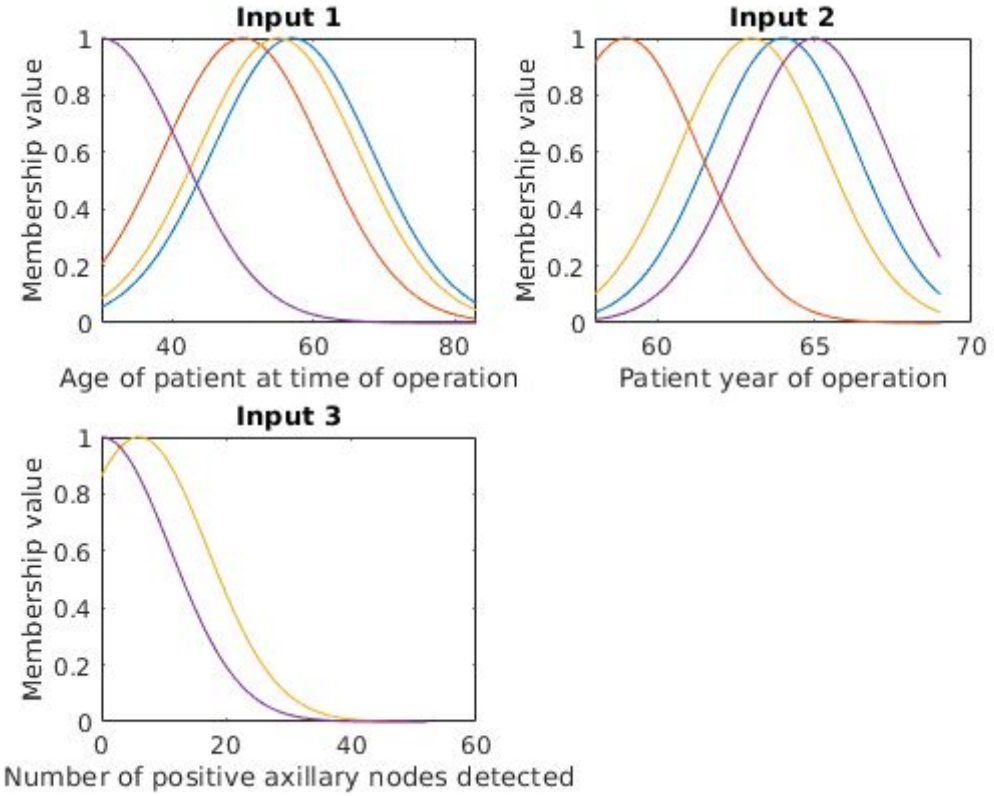


Εικόνα 2: Καμπύλες μάθησης μοντέλου 1.

Μοντέλο 2

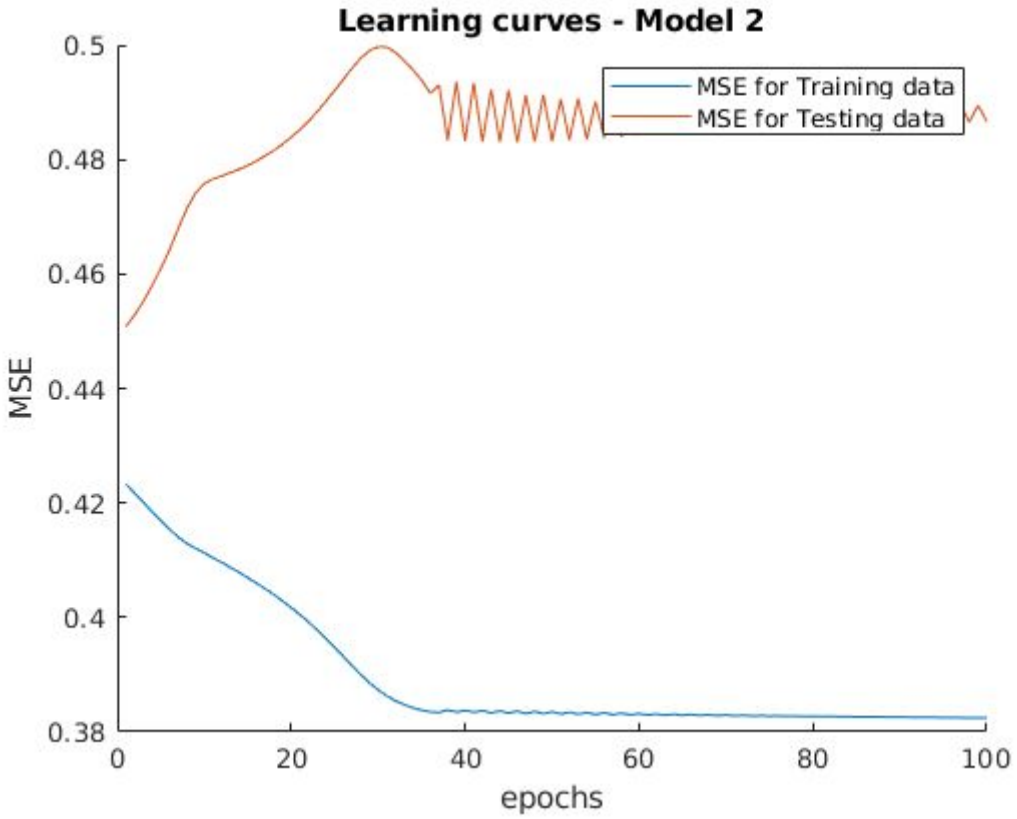
- Ακτίνα cluster: 0.6
- Ανεξάρτητη κλάσης
- κανόνες: 4

Συναρτήσεις Συμμετοχής



Εικόνα 3: Συναρτήσεις συμμετοχής των 3 εισόδων.

Καμπύλες μάθησης

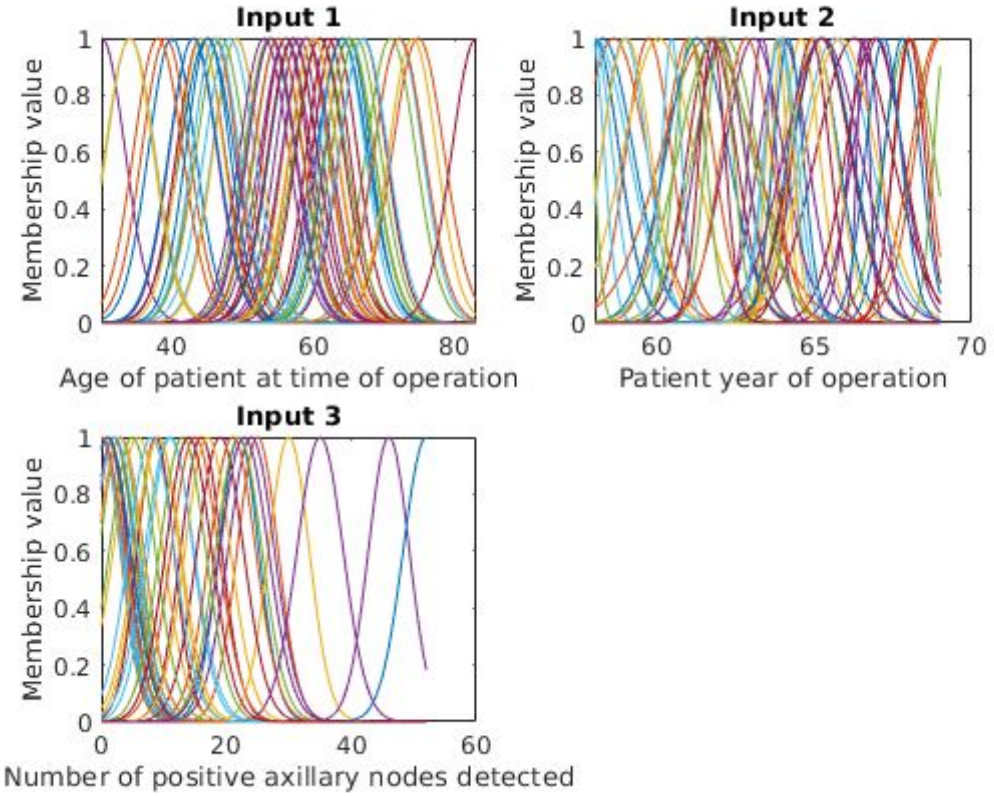


Εικόνα 4: Καμπύλες μάθησης μοντέλου 2.

Μοντέλο 3

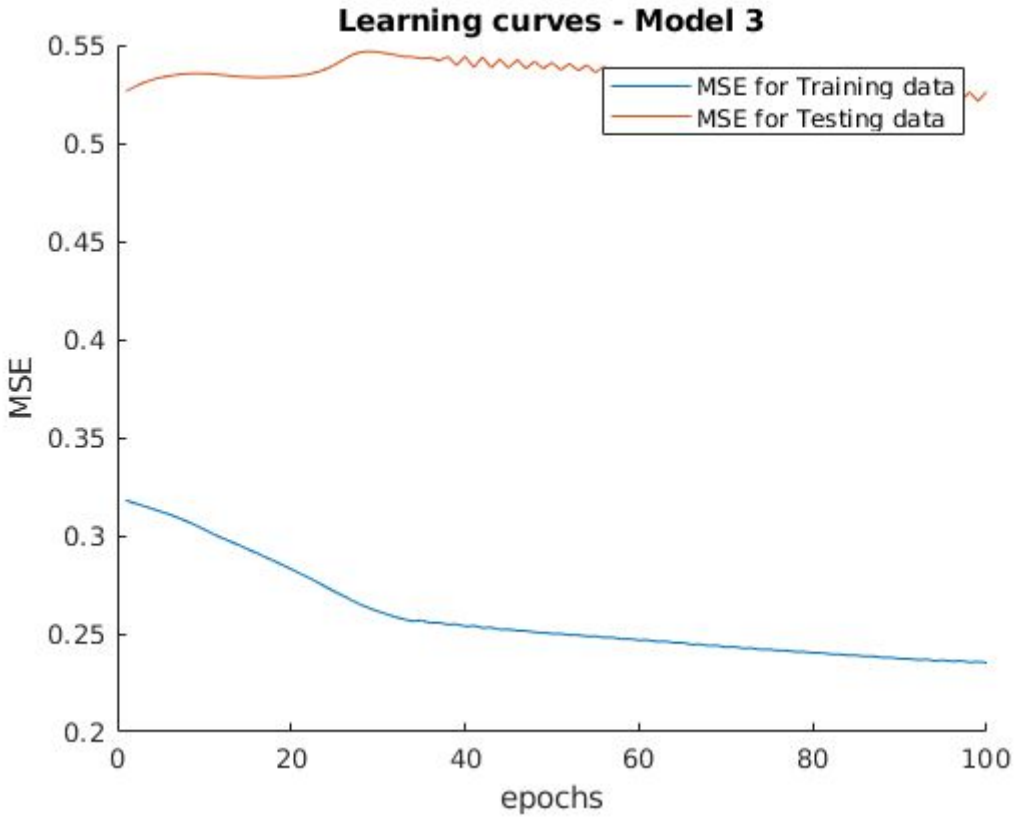
- Ακτίνα cluster: 0.2
- Εξαρτημένη από την κλάση
- κανόνες: 56

Συναρτήσεις Συμμετοχής



Εικόνα 5: Συναρτήσεις συμμετοχής των 3 εισόδων.

Καμπύλες μάθησης

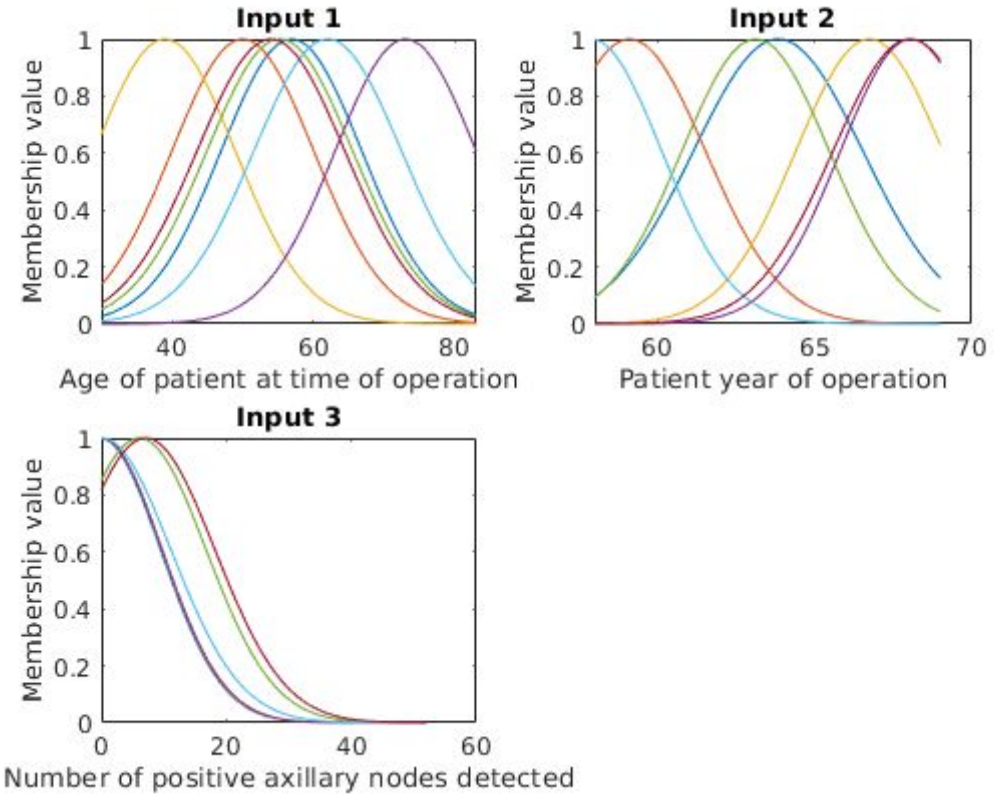


Εικόνα 6: Καμπύλες μάθησης μοντέλου 3.

Μοντέλο 4

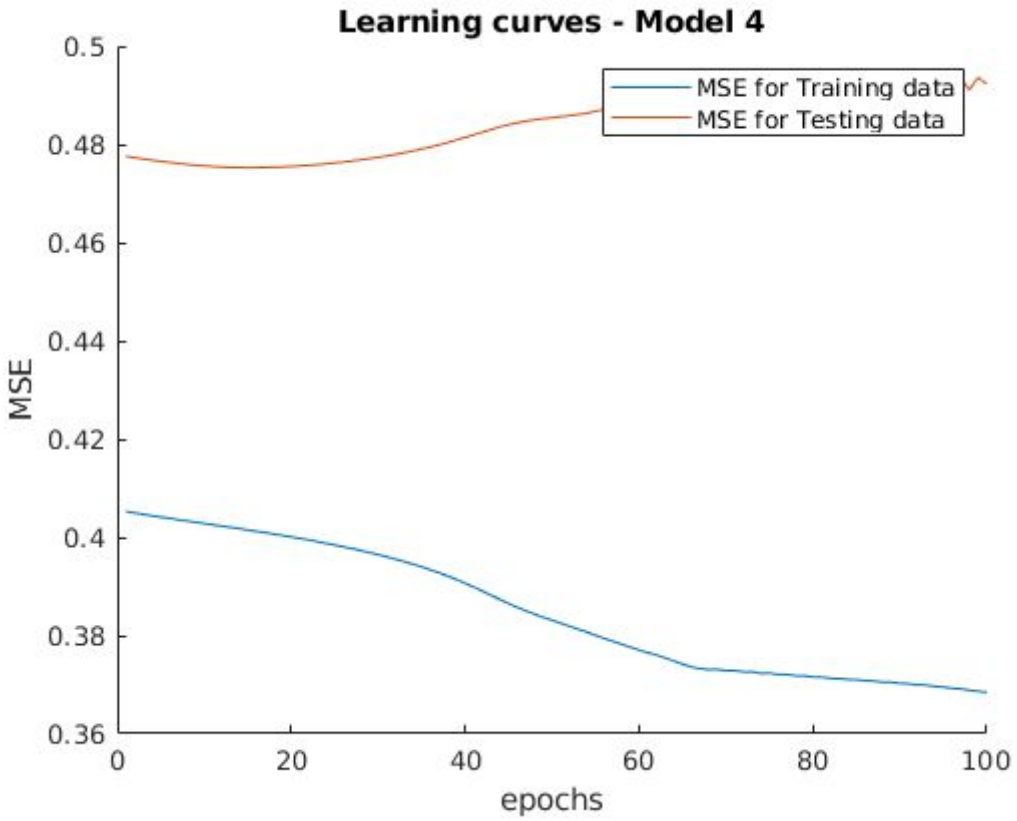
- Ακτίνα cluster: 0.6
- Εξαρτημένη από την κλάση
- κανόνες: 7

Συναρτήσεις Συμμετοχής



Εικόνα 7: Συναρτήσεις συμμετοχής των 3 εισόδων.

Καμπύλες μάθησης



Εικόνα 8: Καμπύλες μάθησης μοντέλου 4.

Μετρικές

Error Matrix 1	Actual data		
Predicted data		1	2
	1	42	9
	2	3	7

Πίνακας 1: Πίνακας σφαλμάτων μοντέλου 1.

Error Matrix 2	Actual data		
Predicted data		1	2
	1	44	13
	2	1	3

Πίνακας 2: Πίνακας σφαλμάτων μοντέλου 2.

Error Matrix 3	Actual data		
Predicted data		1	2
	1	36	8
	2	9	8

Πίνακας 3: Πίνακας σφαλμάτων μοντέλου 3.

Error Matrix 4	Actual data		
Predicted data		1	2
	1	43	12
	2	2	4

Πίνακας 4: Πίνακας σφαλμάτων μοντέλου 4.

Μοντέλο	Εξάρτηση από κλάσεις	Κανόνες	Ακτίνα	OA	PA	UA	k
1	Οχι	32	0.2	0.8033	[0.9333, 0.4375]	[0.8235, 0.7]	0.4218
2	Οχι	4	0.6	0.7705	[0.9778, 0.1875]	[0.7719, 0.75]	0.2179
3	Ναι	56	0.2	0.7213	[0.8, 0.5]	[0.8182, 0.4706]	0.2941
4	Ναι	7	0.6	0.7705	[0.9556, 0.25]	[0.7818, 0.6667]	0.2574

Πίνακας 5: Πίνακας μετρικών.

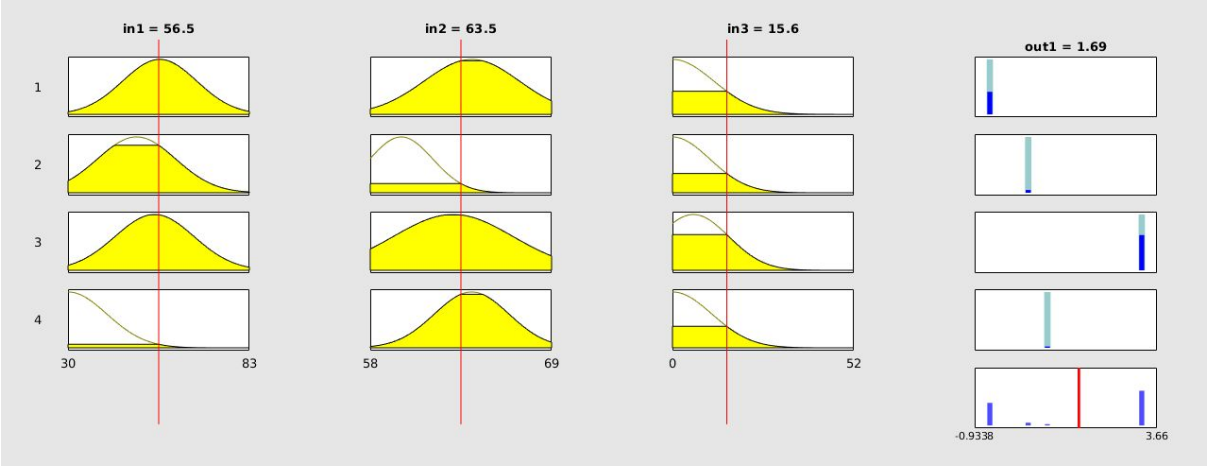
Σχολιασμός

Όπως παρατηρούμε από τον πίνακα 5 η συνολική ακρίβεια των ταξινομητών είναι καλή σχετικά καθώς τα μοντέλα παρουσιάζουν συνολική ακρίβεια μέσα στο εύρος 77% - 80% με μοναδική εξαίρεση το μοντέλο 3 το οποίο παρουσιάζει ακρίβεια γύρω στο 72%. Το dataset που δόθηκε γενικά έχει ένα σχετικά μικρό μέγεθος γεγονός που ίσως δικαιολογεί και την μετρια προς καλή απόδοση των ταξινομητών. Επίσης, το dataset διαθέτει πολύ περισσότερα δεδομένα που ανήκουν στην κλάση 1 από ότι δεδομένα που ανήκουν στην κλάση 2. Αυτό καθιστά δύσκολο το γεγονός να ταξινομηθούν σωστά τα δεδομένα της κλάσης 2. Αυτό παρατηρείται άλλωστε και από τους δείκτες PA και UA οι οποίοι καθορίζουν την απόδοση του ταξινομητή όσον αφορά την κάθε κλάση ξεχωριστά. Συγκεκριμένα, παρατηρούμε ότι για τον δείκτη PA η κλάση 1 παρουσιάζει πολύ καλύτερη ακρίβεια σε κάθε μοντέλο. Το ίδιο συμπέρασμα προκύπτει και από τον δείκτη UA καθώς σε όλες τις περιπτώσεις ο δείκτης της κλάσης 1 είναι μεγαλύτερος από αυτόν της κλάσης 2. Θα μπορούσαμε να πούμε μόνο ότι στα μοντέλα 2 και 4 οι δείκτες UA είναι κοντά για τις δύο κλάσεις. Παρατηρούμε ακόμη ότι έχουμε μεγαλύτερο αριθμό κανόνων με μικρότερη ακτίνα χωρίς βέβαια να φαίνεται να επηρεάζει σε μεγάλο βαθμό το μήκος της ακτίνας και κατ' επέκταση ο αριθμός των κανόνων τη συνολική ακρίβεια του μοντέλου. Σε ότι αφορά την εξάρτηση του subtractive clustering από τις κλάσεις παρατηρείται ότι όταν ο αλγόριθμος τρέχει ανεξάρτητα από τις κλάσεις τότε παρουσιάζει ελαφρώς καλύτερα αποτελέσματα αν κρίνουμε από την συνολική ακρίβεια. Το καλύτερο μοντέλο θα μπορούσαμε να πούμε ότι είναι το μοντέλο 1 με βάση την συνολική ακρίβεια. Τα μοντέλα 2 και 4 παρατηρούμε ότι έχουν ίδια ακρίβεια οπότε θα κρίνουμε το αποτέλεσμα με βάση τον δείκτη κ όπου παρατηρούμε ότι το μοντέλο 4 έχει τον υψηλότερο. Έπειτα ακολουθεί το μοντέλο 2 και τελειώνουμε με το μοντέλο 3. Ο πίνακας 6 παρουσιάζει με φθίνουσα σειρά απόδοσης τα 4 μοντέλα.

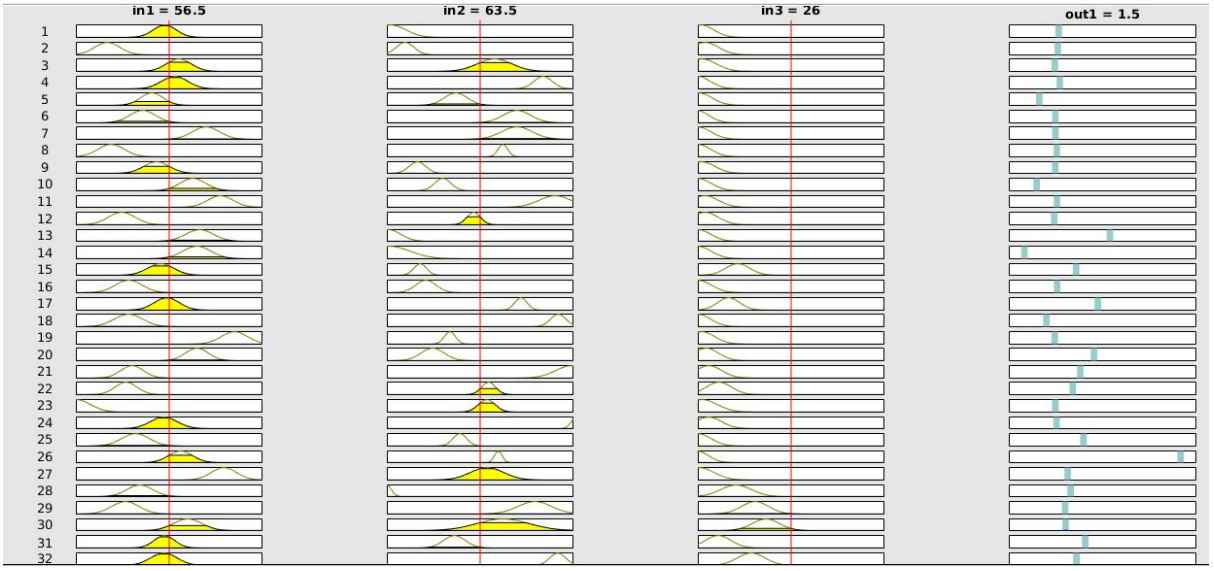
Μοντέλο 1
Μοντέλο 4
Μοντέλο 2
Μοντέλο 3

Πίνακας 6

Σε ότι αφορά την επικάλυψη των προβολών των ασαφών συνόλων θα εξετάσουμε τις δύο περιπτώσεις των ακτινών 0.2 και 0.6 των μοντέλων 1 και 2 αντίστοιχα. Αρχικά εξετάζουμε το μοντέλο 2 λόγω μικρότερου αριθμού κανόνων. Συγκεκριμένα, παρατηρούμε ότι το μοντέλο 2 έχει τέσσερις κανόνες οι οποίοι παρουσιάζουν επικάλυψη για τις τρεις εισόδους. Περισσότερο αισθητό το αποτέλεσμα γίνεται για την είσοδο 3 όπου τα σύνολα των κανόνων παρουσιάζουν μεγάλη επικάλυψη. Η επικάλυψη αυτή σημαίνει ότι μια μικρή αλλαγή στην είσοδο 3 δεν θα προκαλέσει σημαντική διαφορά στην έξοδο. Η εικόνα 9 παρουσιάζει τους κανόνες του μοντέλου 2 και την επικάλυψη τους. Στα ίδια συμπεράσματα καταλήγουμε και για το μοντέλο 1 των 32 κανόνων και ακτίνας 0.2. Όπως φαίνεται στην εικόνα 10 υπάρχει σημαντική επικάλυψη μεταξύ των κανόνων της εισόδου 3.



Εικόνα 9: Κανόνες μοντέλου 2.



Εικόνα 10: Κανόνες μοντέλου 1.

Μέρος 2 - Epileptic Seizure Recognition dataset

Περιγραφή - Διαχωρισμός dataset

Το dataset που επιλέγεται στο δεύτερο μέρος της εργασίας έχει υψηλότερο βαθμό διαστασιμότητας και αποτελείται από 11500 δείγματα και 179 χαρακτηριστικά. Επειδή ο αριθμός δειγμάτων και χαρακτηριστικών είναι πολύ μεγάλος δεν είναι δυνατόν να χρησιμοποιηθεί η μέθοδος του πρώτου μέρους γιατί ο αριθμός κανόνων θα αυξάνονταν σε πολύ μεγάλο βαθμό. Για να αποφευχθεί αυτό θα χρησιμοποιήσουμε τη μεθοδο της επιλογής χαρακτηριστικών και της διαμέρισης του χώρου οι οποίες βέβαια εισάγουν δύο ελεύθερες μεταβλητές στο πρόβλημα (αριθμός χαρακτηριστικών προς επιλογή και αριθμός ομάδων που δημιουργούνται) οι οποίες θα προσεγγιστούν με τη μέθοδο αναζήτησης πλέγματος. Ο διαχωρισμός των δεδομένων γίνεται όπως έγινε και στο πρώτο μέρος της εργασίας με χρήσης της dividerand ακολουθούμενη από ένα shuffle στα τρια σετ δεδομένων έτσι ώστε να εισάγουμε μια τυχαιότητα.Το dataset χωρίζεται κατά 60%, 20%, 20% για τα training, validation και testing σετ αντίστοιχα.

Περιγραφή διαδικασίας

Αρχικά επιλέγονται οι τιμές για τον αριθμό των χαρακτηριστικών (features number) και οι τιμές για την ακτίνα των clusters (cluster radius). Οι δύο πίνακες με τις τιμές τους είναι οι εξής:

- features_number = [6, 10, 15, 20]
- cluster_radius = [0.15, 0.3, 0.45, 0.6]

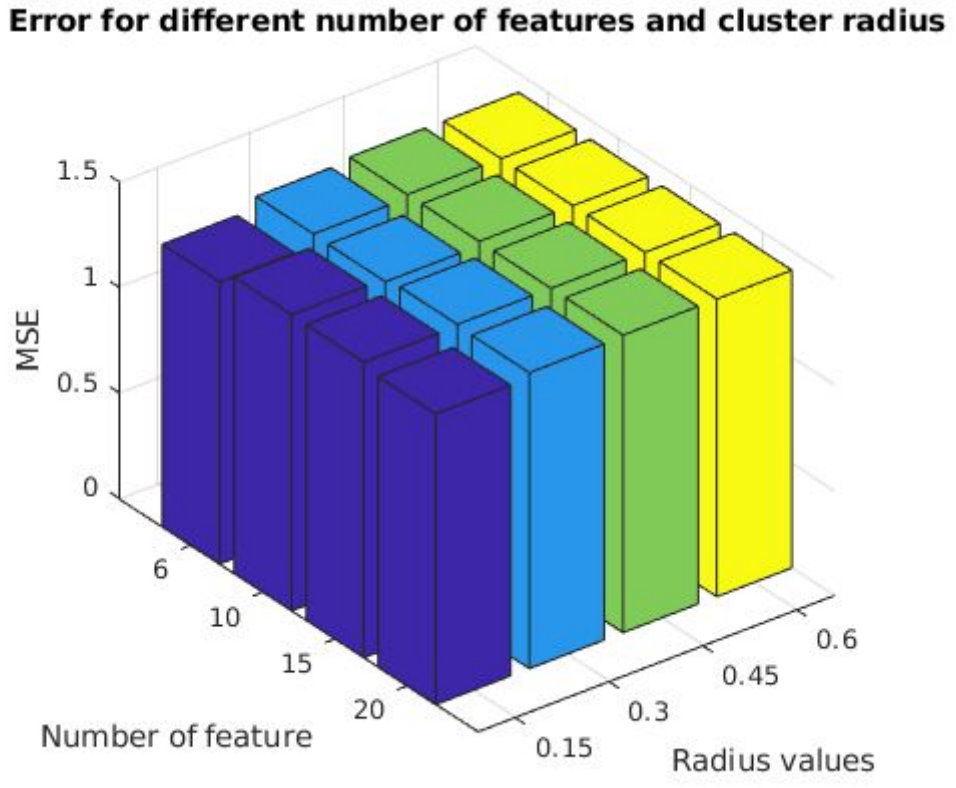
Η αξιολόγηση μέσω διασταυρωμένης επικύρωσης γίνεται με την συνάρτηση cvpartition του MATLAB επιλέγοντας 5 folds. Για την αξιολόγηση τον χαρακτηριστικών χρησιμοποιείται ο

αλγόριθμος relief χρησιμοποιώντας την έτοιμη συνάρτηση του MATLAB η οποία επιστρέφει στον πίνακα idx τα πιο σημαντικά χαρακτηριστικά και στον πίνακα weights τα αντίστοιχα βάρη. Έπειτα για τη βέλτιστη επιλογή παραμέτρων ξεκινάει μια σειρά από τρία loops. Η πρώτη επανάληψη τρέχει για κάθε αριθμό από features, η δεύτερη τρέχει για κάθε ακτίνα των clusters και η τρίτη για κάθε fold. Μέσα στην τριπλή loop εκπαιδεύουμε ένα fis model για κάθε fold του αλγορίθμου. Το fis μοντέλο δημιουργείται μέσω της συνάρτησης genfis2. Η εκπαίδευση πραγματοποιείται για κάθε fold του αλγορίθμου και τρέχει για 1 εποχή λόγω του μεγάλου χρόνου εκπαίδευσης. Στη συνέχεια αποθηκεύουμε για κάθε feature και radius το μέσο σφάλμα. Όταν τελειώσει η τριπλή επανάληψη επιλέγουμε το βέλτιστο μοντέλο με βάση το ελάχιστο σφάλμα. Στο τέλος εκπαιδεύουμε το βέλτιστο μοντέλο και υπολογίζουμε τις αντίστοιχες μετρικές. Για την απεικόνιση των διαφόρων διαγραμμάτων χρησιμοποιούμε το script “plot_script_classification.m”.

Διαγράμματα

Παρακάτω στην Εικόνα 11 παρουσιάζεται το διάγραμμα μπάρας του σφάλματος σε συνάρτηση με τον αριθμό χαρακτηριστικών και των διάφορων ακτινών. Επίσης, στην εικόνα 12 παρουσιάζεται ο αριθμός των κανόνων σε συνάρτηση με τις ίδιες ποσότητες.

Διάγραμμα σφαλμάτων



Εικόνα 11: Διάγραμμα μπάρας για το μέσο τετραγωνικό σφάλμα σε συνάρτηση με τον αριθμό χαρακτηριστικών και τις ακτίνες.

Στον πίνακα 6 παρουσιάζονται τα αποτελέσματα των μέσων τετραγωνικών σφαλμάτων για τους διάφορους συνδυασμούς χαρακτηριστικών και ακτινών.

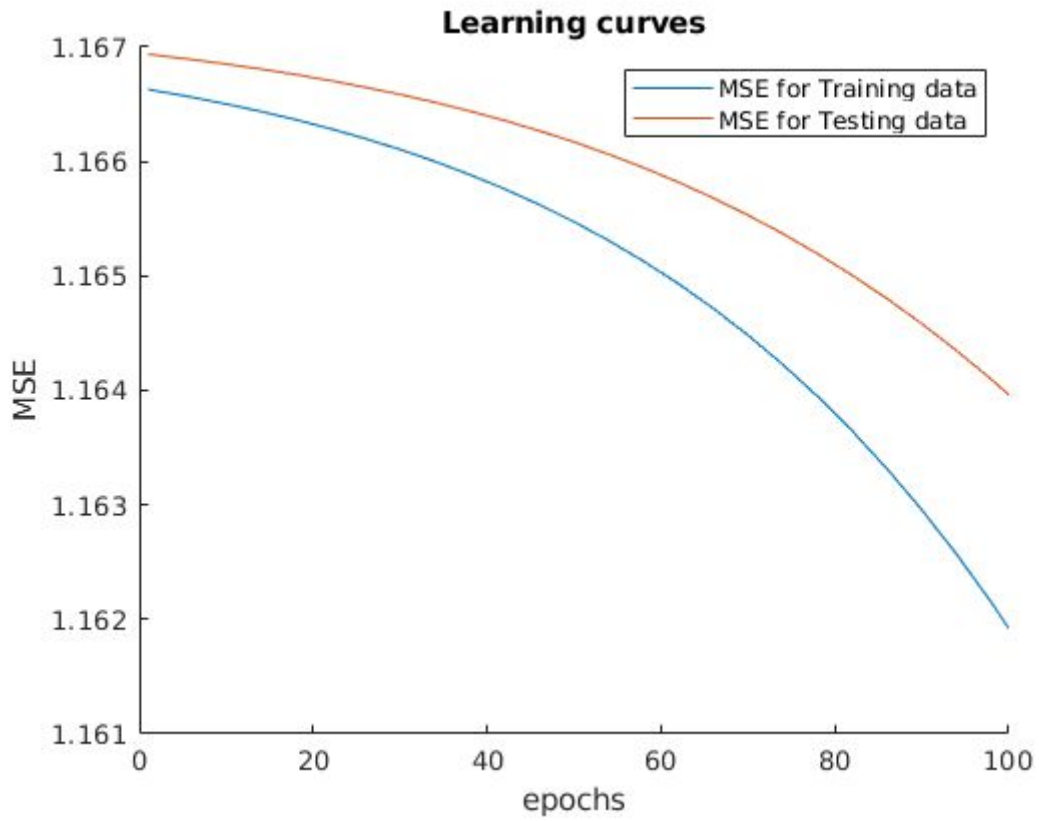
Radius Features	0.15	0.3	0.45	0.6
6	1.3376	1.3934	1.4099	1.4097
10	1.4039	1.3894	1.4100	1.4090
15	1.4011	1.4094	1.4098	1.4100
20	1.3781	1.4029	1.4103	1.4110

Πίνακας 6: Αποτελέσματα των μέσων τετραγωνικών σφαλμάτων για τους διάφορους συνδυασμούς χαρακτηριστικών και ακτινών.

Σχολιασμός Αποτελεσμάτων Πίνακα

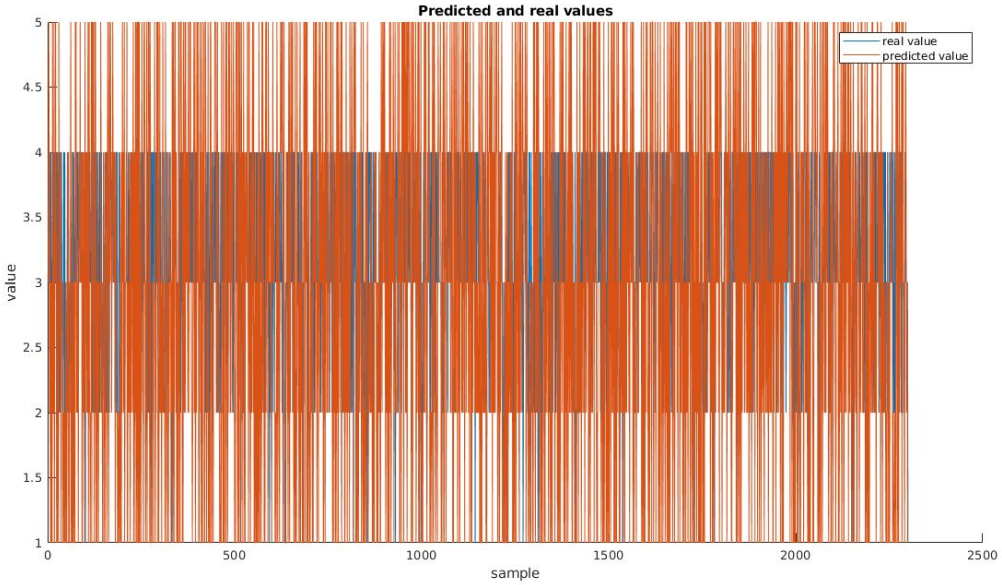
Από τον πίνακα των σφαλμάτων γίνεται αντιληπτό ότι για μικρότερη ακτίνα κρατώντας σταθερό τον αριθμό των features το σφάλμα μειώνεται. Εξαίρεση μπορούμε να πούμε ότι αποτελεί η περίπτωση που έχουμε 10 features όπου το σφάλμα δεν ακολουθεί ανοδική πορεία καθώς αυξάνεται η ακτίνα αλλά ταλαντώνεται. Από την άλλη πλευρά, αν κρατήσουμε σταθερή την ακτίνα δεν παρατηρείται κάποιο μοτίβο ως προς τη μεταβολή του σφάλματος καθώς αυξάνονται τα features. Η καλύτερη περίπτωση παρατηρείται για 6 features και ακτίνα 0.15 (πορτοκαλί κελί) όπου το σφάλμα είναι ελάχιστο. Επίσης, καλές περιπτώσεις είναι τα εξής ζεύγη (features, radius): (20, 0.15), (6, 0.3), (10, 0.3). Ωστόσο, η περίπτωση (6, 0.15) είναι αισθητά καλύτερη οπότε και επιλέγεται ως η βέλτιστη. Το μοντέλο που εκπαιδεύτηκε διαθέτει έξι εισόδους και έξι κανόνες. Για την εκπαίδευση του βέλτιστου μοντέλου χρησιμοποιήθηκε η συνάρτηση classDep2 η οποία έχει αντίστοιχη λειτουργία με την συνάρτηση classDep1 του πρώτου μέρους της εργασίας. Ο λόγος που έγινε αυτό είναι διότι είχε ελαφρώς καλύτερα αποτελέσματα από την genfis2.

Καμπύλες μάθησης



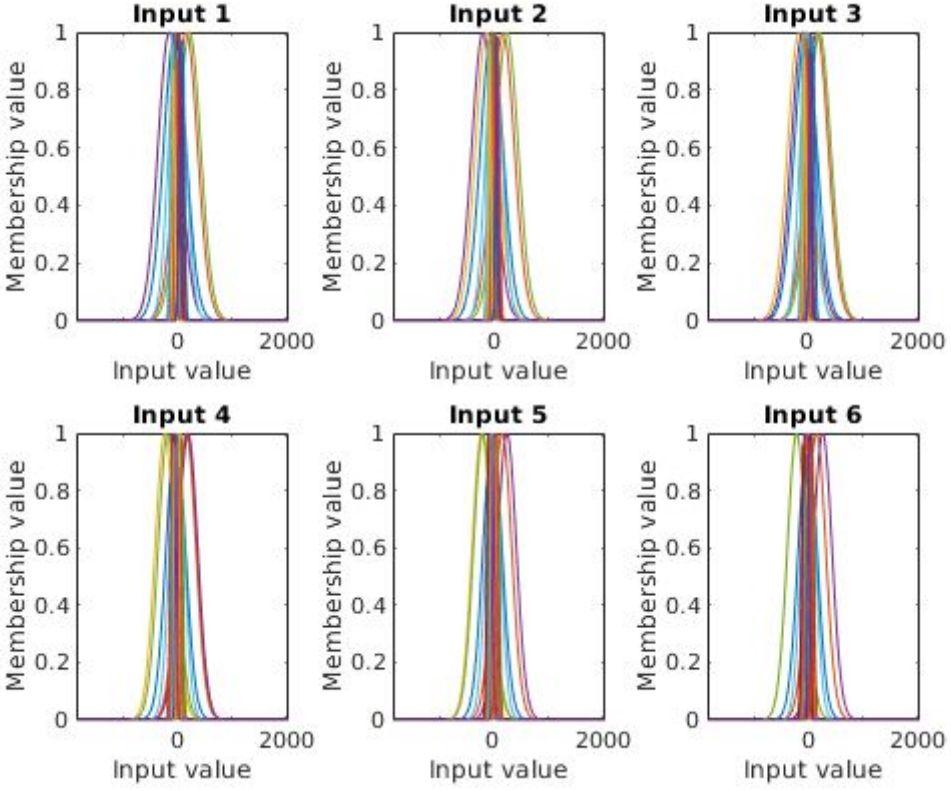
Εικόνα 12: Καμπύλη μάθησης

Διάγραμμα πραγματικών και προβλέψιμων τιμών

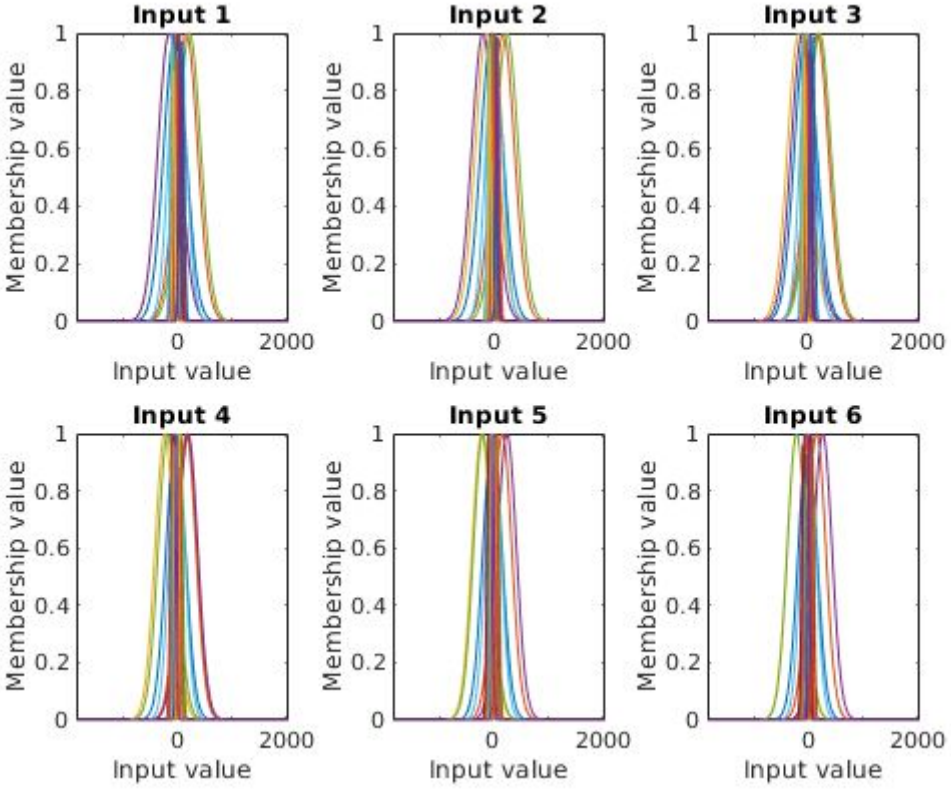


Εικόνα 13: Διάγραμμα πραγματικών και προβλέψιμων τιμών.

Αρχικά Σύνολα



Τελικά Σύνολα



Μετρικές

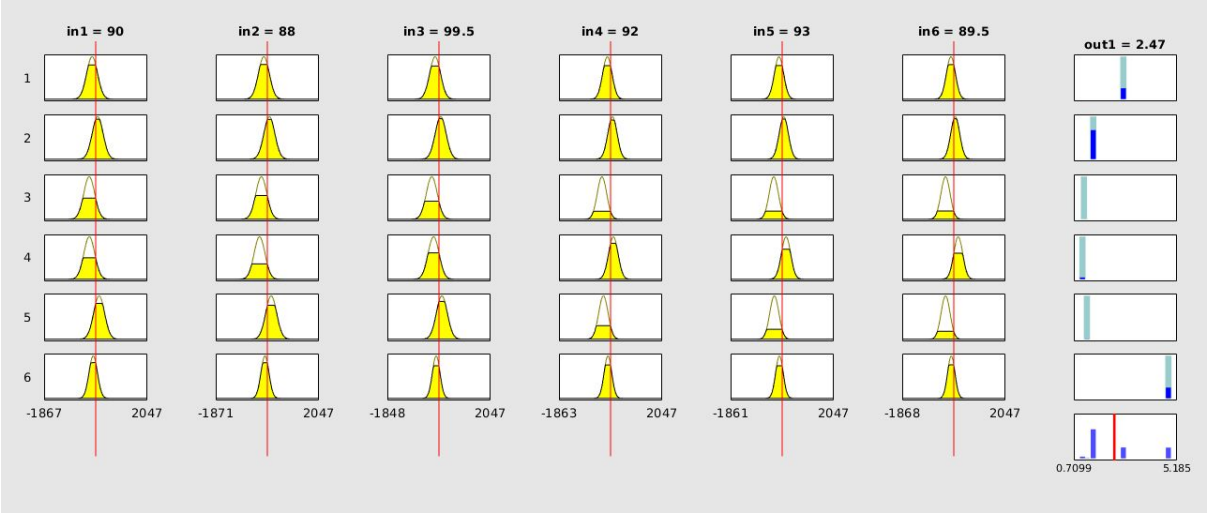
Error Matrix	Actual data					
Predicted data		**1**	**2**	**3**	**4**	**5**
	1	227	11	1	6	0
	2	152	41	37	39	13
	3	94	195	193	242	187
	4	14	211	213	167	257
	5	0	0	0	0	0

Μοντέλο	Βέλτιστο
OA	0.2730
PA	[0.4661,0.0895, 0.4347, 0.3678, 0]
UA	[0.9265, 0.1454, 0.2119, 0.1937, -]
k	0.0942

Παρατηρούμε στον πίνακα με τις μετρικές αξιολόγησης ότι η συνολική ακρίβεια είναι 27%. Το ποσοστό αυτό δεν είναι καθόλου ικανοποιητικό για ακρίβεια αλλά ήταν το καλύτερο που προέκυπτε. Σε ότι αφορά την ακρίβεια που παρουσιάζεται σε κάθε κλάση παρατηρείται ότι η κλάση 1 είναι αυτή που παρουσιάζει πολύ καλύτερα αποτελέσματα ταξινόμησης από τις υπόλοιπες 4. Αυτό είναι ένα καλό αποτέλεσμα εφόσον η κλάση 1 είναι και η σημαντικότερη διότι αυτή δείχνει αν υπάρχει επιληπτική κρίση. Τα χειρότερα αποτελέσματα παρατηρούνται στην κλάση 5 όπου βλέπουμε ότι δεν έχει προβλεφθεί κανένα αποτέλεσμα που να ανήκει σε αυτή την κλάση. Οι κλάσεις 2,3,4 παρατηρούμε ότι ανήκουν κάπου στη μέση σχετικά με την ακρίβεια παρουσιάζοντας μέτρια επίδοση. Οπότε, παρα το κακό ποσοστό συνολικής ακρίβειας μπορούμε να πούμε ότι το μοντέλο μας ικανοποιεί ως ενα βαθμο εφόσον ταξινομεί αρκετά καλά την κλάση 1.

Επικάλυψη κανόνων

Σε ότι αφορά την επικάλυψη των προβολών των ασαφών συνόλων παρατηρούμε ότι το μοντέλο έχει έξι κανόνες οι οποίοι παρουσιάζουν επικάλυψη για τις εισόδους. Περισσότερο αισθητό το αποτέλεσμα γίνεται για τις τρεις πρώτες εισόδους όπου τα σύνολα των κανόνων παρουσιάζουν μεγάλη επικάλυψη. Η επικάλυψη αυτή σημαίνει ότι μια μικρή αλλαγή στις εισόδους δεν θα προκαλέσει σημαντική διαφορά στην έξοδο. Η εικόνα 14 παρουσιάζει τους κανόνες του μοντέλου και την επικάλυψη τους. Το συμπέρασμα δηλαδή που προκύπτει είναι ότι κάποιοι κανόνες θα μπορούσαν να παραληφθούν διότι δεν επηρεάζουν σημαντικά το αποτέλεσμα στην έξοδο.



Εικόνα 14: Κανόνες μοντέλου.

Σύγκριση με grid partitioning

Σχετικά με την απόδοση του συστήματος παρατηρούμε ότι είναι αρκετά ικανοποιητική καθώς χρησιμοποιούμε σχετικά μικρό αριθμό χαρακτηριστικών και κανόνων. Αν είχαμε επιλέξει τη μέθοδο του grid partitioning με 178 χαρακτηριστικά θα έπρεπε να χρησιμοποιηθούν 2^{178} και 3^{178} κανόνες για 2 και 3 εισόδους αντίστοιχα σε αντίθεση με τη δική μας περίπτωση που χρησιμοποιούνται μόνο 6 κανόνες. Σίγουρα, η ακρίβεια θα ήταν μεγαλύτερη αλλά ο χρόνος εκτέλεσης θα ήταν απαγορευτικός. Συμπεραίνουμε λοιπόν ότι δεδομένου του μικρού αριθμού κανόνων που χρησιμοποιούμε τα αποτελέσματα είναι αρκετά ικανοποιητικά ως προς την ακρίβεια και τον χρόνο εκτέλεσης.