Εργασία 4
Επίλυση προβλήματος ταξινόμησης με χρήση μοντέλων TSK

Φοιτητής : Μπεκιάρης Θεοφάνης AEM:8200

1)Πρώτη εφαρμογή Wifi-localization Dataset

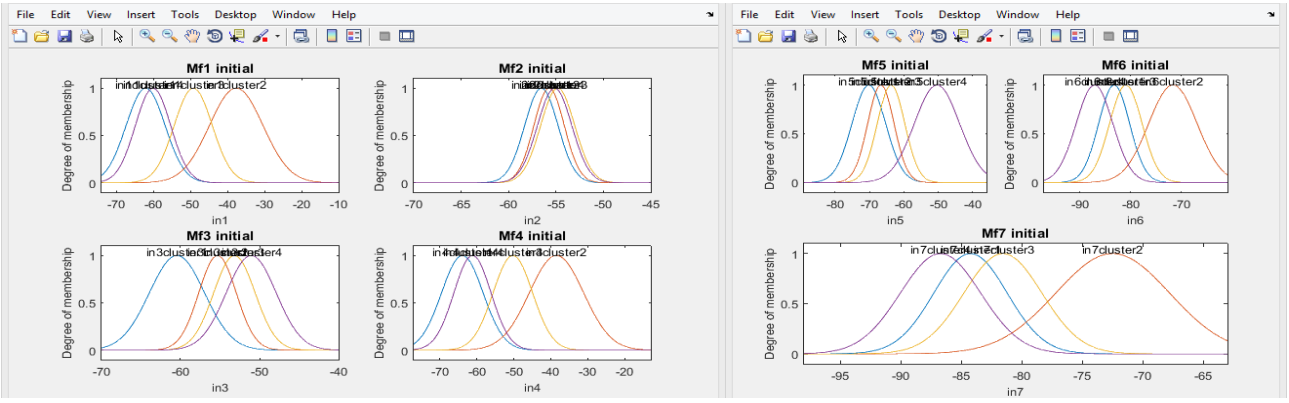
• Στο πρώτο κομμάτι της εργασίας αυτό που πρέπει να κάνουμε είναι αρχικά να διαχωρίσουμε τα δεδομένα σε 3 ομάδες,training data,evaluation data και check data,με ποσοστό 60%,20% και 20%. Πριν διαχωρίσουμε τα δεδομένα σε αυτές τις 3 ομάδες πρώτα θα τα ανακατεύσουμε με σκοπό η συχνότητα εμφάνισης δειγμάτων που ανήκουν σε μια συγκεκριμένη κλάση, σε κάθε ένα από τα τρία σύνολα διαμέρισης, να είναι όσο το δυνατόν πιο "όμοια" με την αντίστοιχη συχνότητα εμφάνισής τους στο αρχικό σύνολο δεδομένων.

• *Εκπαίδευση TSK μοντέλων με διαφορετικές παραμέτρους:* Στην συνέχεια η διαμέριση του χώρου εισόδου θα γίνει με τη μέθοδο του Fuzzy C-Means (FCM) και τα TSK μοντέλα που θα προκύψουν θα διαφέρουν ως προς την παράμετρο που καθορίζει τον αριθμό των κανόνων. Για αυτό τον σκοπό χρησιμοποιούμε την συνάρτηση genfis3 του Matlab. Ο αριθμός των κανόνων αντίστοιχα για κάθε μοντέλο είναι NR = [4,8,12,16] και μπαίνει ως όρισμα στην συνάρτηση **genfis3.**

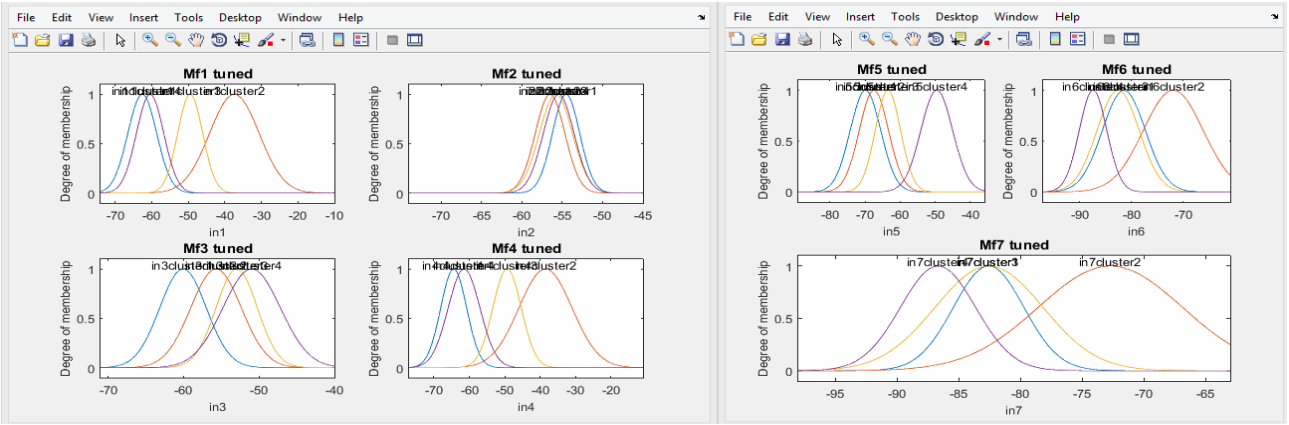
TSK model 1 με 4 κανόνες

Η εκπαίδευση των μοντέλων έχει γίνει για 300 εποχές. Παρακάτω παρουσιάζονται οι μορφές των ασαφών συνόλων πριν και μετά την διαδικασία εκπαίδευσης του μοντέλου για τις 7 εισόδους/χαρακτηριστικά.

Ασαφής σύνολα πριν την εκπαίδευση



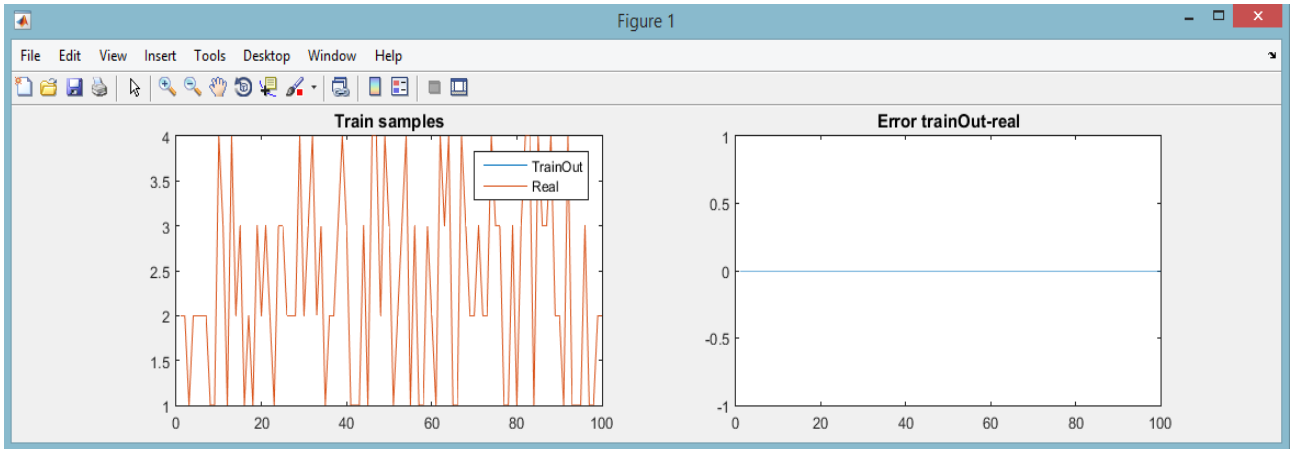
Ασαφή σύνολα μετά την εκπαίδευση του μοντέλου



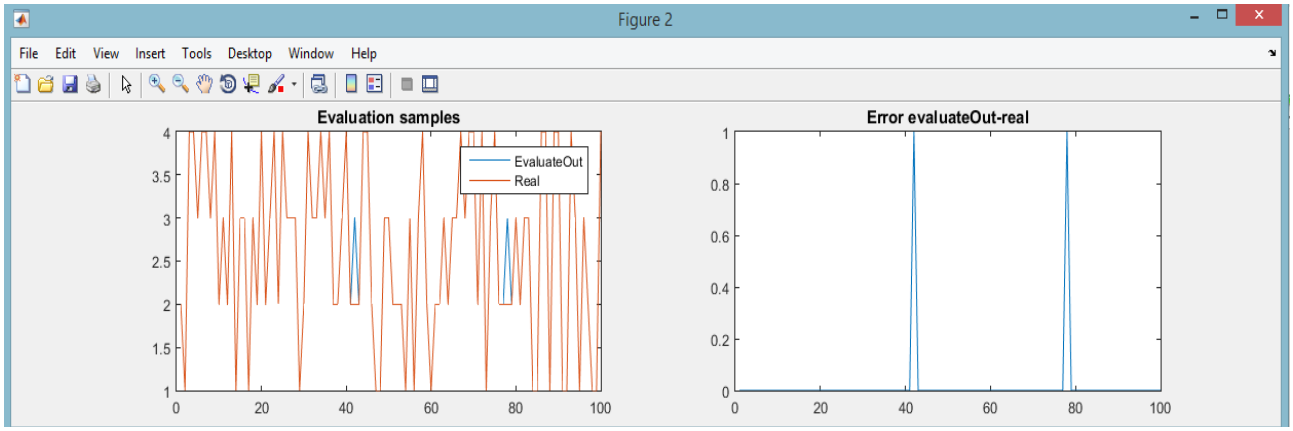
Στα παραπάνω διαγράμματα δεν παρατηρούνται πολύ μεγάλες αλλαγές στα ασαφή σύνολα. Κάποια από αυτά δεν έχουν αισθητές διαφορές, από την άλλη για κάποια βλέπουμε ότι κατά την εκπαίδευση έχουν τροποποιηθεί αρκετά. Παρόλο που έχουμε μικρές τροποποιήσεις η απόδοση του μοντέλου όπως θα φανεί και παρακάτω είναι πολύ καλή.

Τα αποτελέσματα από τις εξόδους ή αλλιώς οι προβλέψεις του εκπαιδευμένου μοντέλου σε σχέση με τις πραγματικές τιμές για κάθε κατηγορία δεδομένων φαίνεται παρακάτω. Επιπλέον στα δεξιά παρατίθεται και το σφάλμα της πραγματικής τιμή από την τιμή της πρόβλεψης. Τα διαγράμματα έχουν γίνει για διάστημα 100 τιμών για την καλύτερη απεικόνιση των δεδομένων.

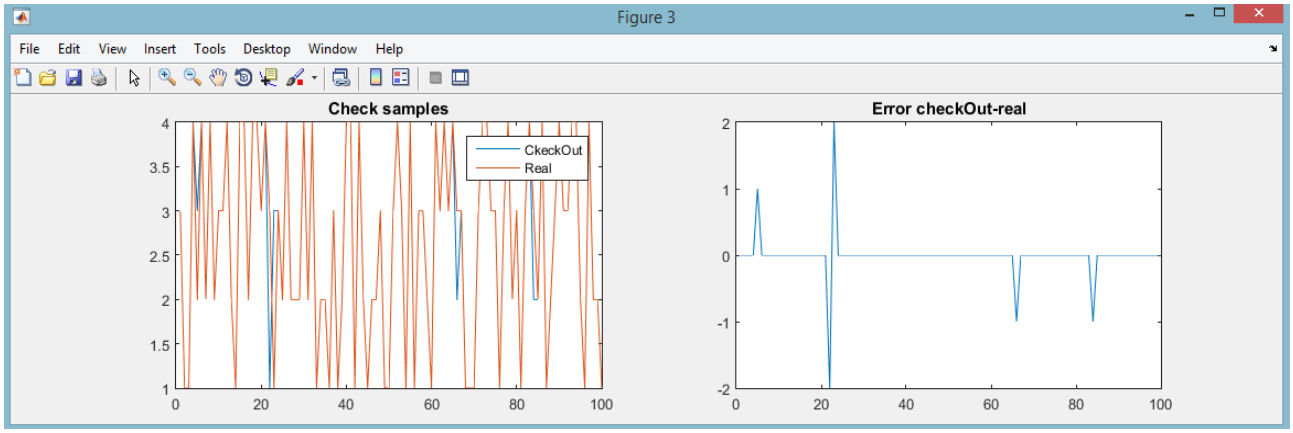
Δεδομένα εκπαίδευσης



Δεδομένα επικύρωσης

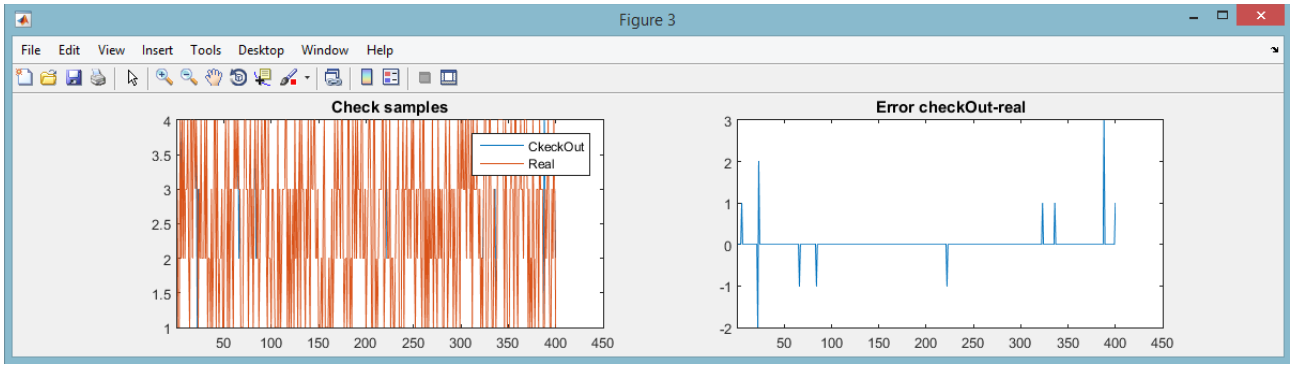


Δεδομένα ελέγχου



Να σημειώσουμε ότι η έξοδος είναι κατηγορική, δηλαδή δηλώνει ένα σύνολο/κατηγορία και όχι μια πραγματική τιμή. Ανάλογα με τις τιμές της εισόδου ο μοντέλο πρέπει να μαντέψει σε πια κατηγορία ανήκουν τα δεδομένα. Επομένως η έξοδος δεν παίρνει συνεχές τιμές,άλλα διακριτές όπως φαίνεται και στα παραπάνω διαγράμματα και η οποίες δηλώνουν μία κατηγορία. Από τα δεδομένα που προκύπτουν παρατηρούμε ότι μοντέλο λειτουργεί με μεγάλο ποσοστό επιτυχίας, στις 100 προβλέψεις βλέπουμε ότι για τα training δεδομένα δεν έχουμε καμία λάθος πρόβλεψη, για τα evaluation δεδομένα 2 λάθος και για τα check 5 λάθος, δηλαδή έχουμε ποσοστά επιτυχίας 100%,98% και 95%.

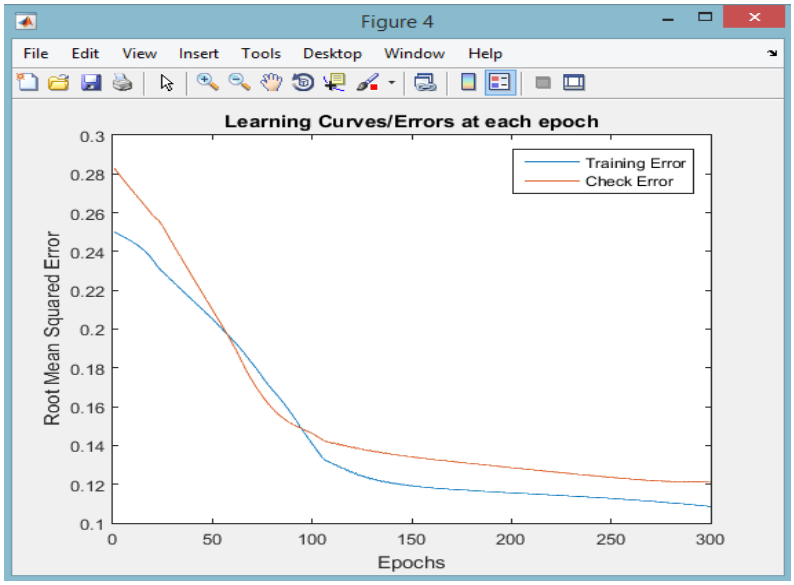
Απεικόνιση ολόκληρου του συνόλου πρόβλεψης των check δεδομένων



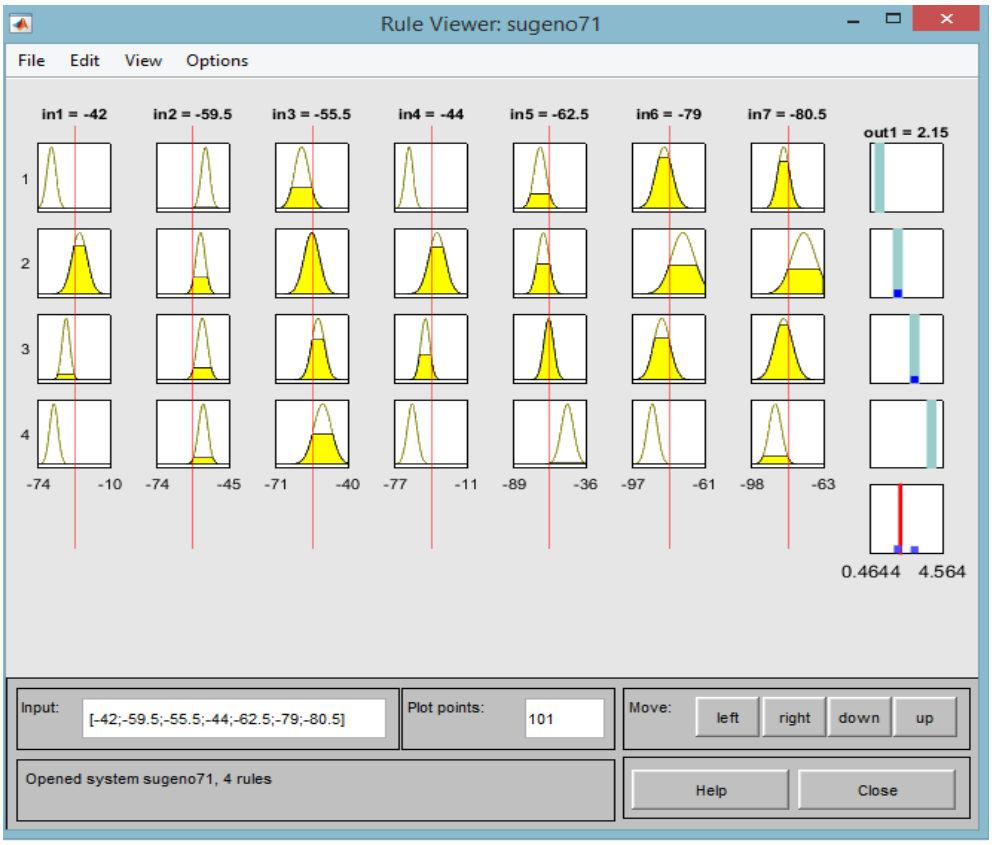
Το μοντέλο όπως βλέπουμε δούλεψε πολύ καλά για όλο το σύνολο των προβλέψεων με μόλις 10 λάθος προβλέψεις σε σύνολο 400 προβλέψεων, άρα με ποσοστό επιτυχίας 390/400 = 97,5 % . Αυτό θα φανεί καλύτερα παρακάνω και στον πίνακα απόδοσης.

Η εκπαίδευση έχει γίνει για 300 εποχές. Από το παρακάτω διάγραμμα μάθησης βλέπουμε ότι η εκπαίδευση δεν είχε κάποιο πρόβλημα,σε κάποιο σημείο στο οποίο το σφάλμα για τα δεδομένα evaluation ελαττώθηκε και έγινε μικρότερο από το σφάλμα των training δεδομένων κάτι που μάλλον είναι τυχαίο αφού τα δεδομένα με τα οποία εκπαιδεύεται το μοντέλο πρέπει να έχουν μικρότερο σφάλμα από τα δεδομένα evaluation αφού με βάση αυτά τα δεδομένα γίνεται και η εκπαίδευση του μοντέλου,δηλαδή με βάση την μείωση του σφάλματος αυτών των δεδομένων εκπαιδεύεται μοντέλο. Συνολικά βλέπουμε το σφάλμα τελικά κατά την διάρκεια της εκπαίδευση ελαττώθηκε και πήρε μια ελάχιστη τιμή για την εκπαίδευση των 300 εποχών.

Διάγραμμα μάθησης

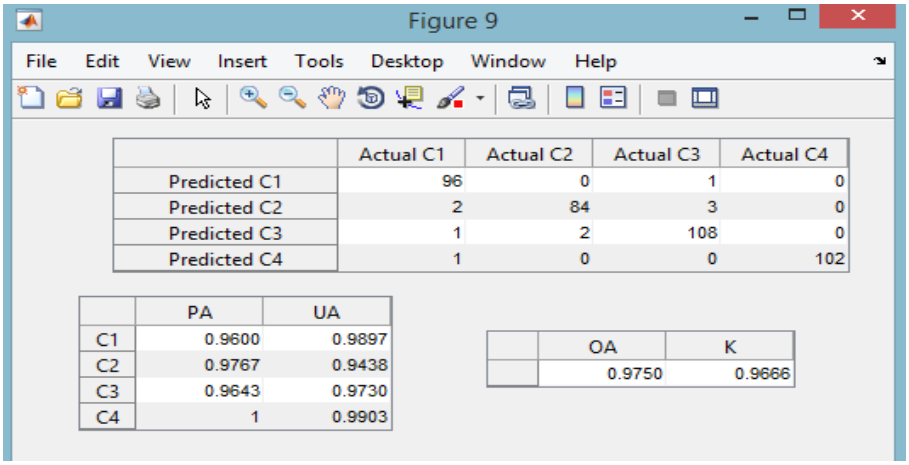


Βάση κανόνων εκπαιδευμένου μοντέλου



Τέλος δίνεται ο πίνακας σφαλμάτων ταξινόμησης(πάνω πίνακας) και οι τιμές των τιμών απόδοσης (κάτω πίνακες). Για τον υπολογισμό των πινάκων έχουν χρησιμοποιηθεί τα check δεδομένα.Για την απεικόνιση τους έχει χρησιμοποιηθεί η συνάρτηση unitable του Matlab.

Πίνακας απόδοσης



	Actual C1	Actual C2	Actual C3	Actual C4
Predicted C1	96	0	1	0
Predicted C2	2	84	3	0
Predicted C3	1	2	108	0
Predicted C4	1	0	0	102

	PA	UA
C1	0.9600	0.9897
C2	0.9767	0.9438
C3	0.9643	0.9730
C4	1	0.9903

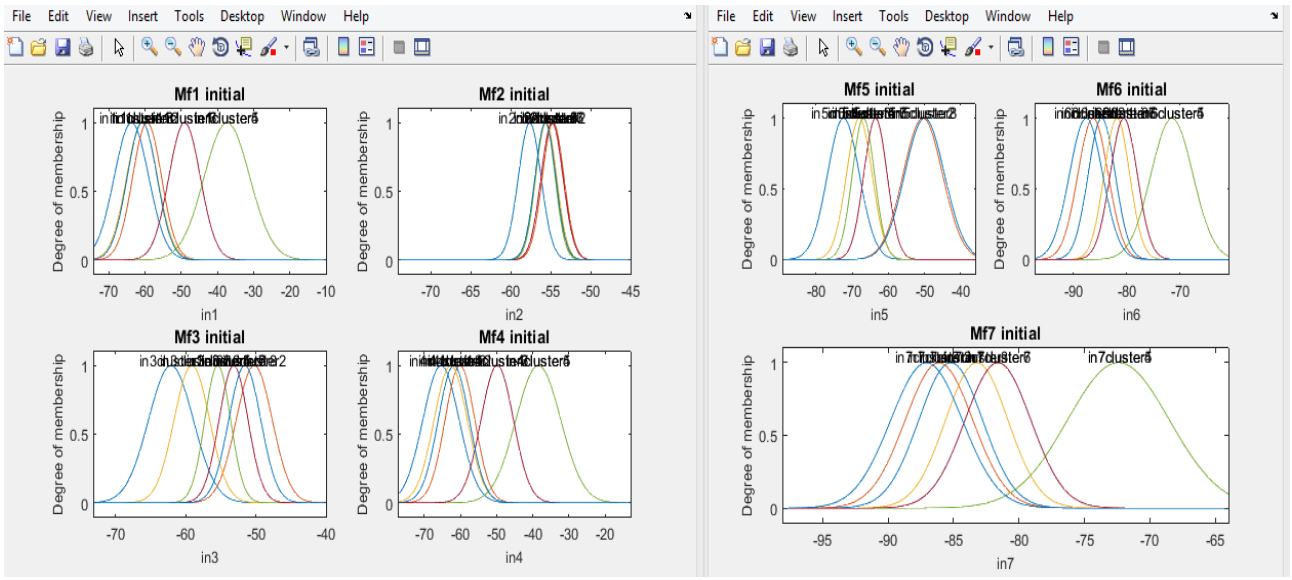
	OA	K
	0.9750	0.9666

Όπως και από τα προηγούμενα διαγράμματα έτσι και από τις τιμές των δεικτών απόδοσης βλέπουμε ότι η απόδοση του μοντέλου είναι πολύ καλή και το μοντέλο κάνει πολύ καλές προβλέψεις.

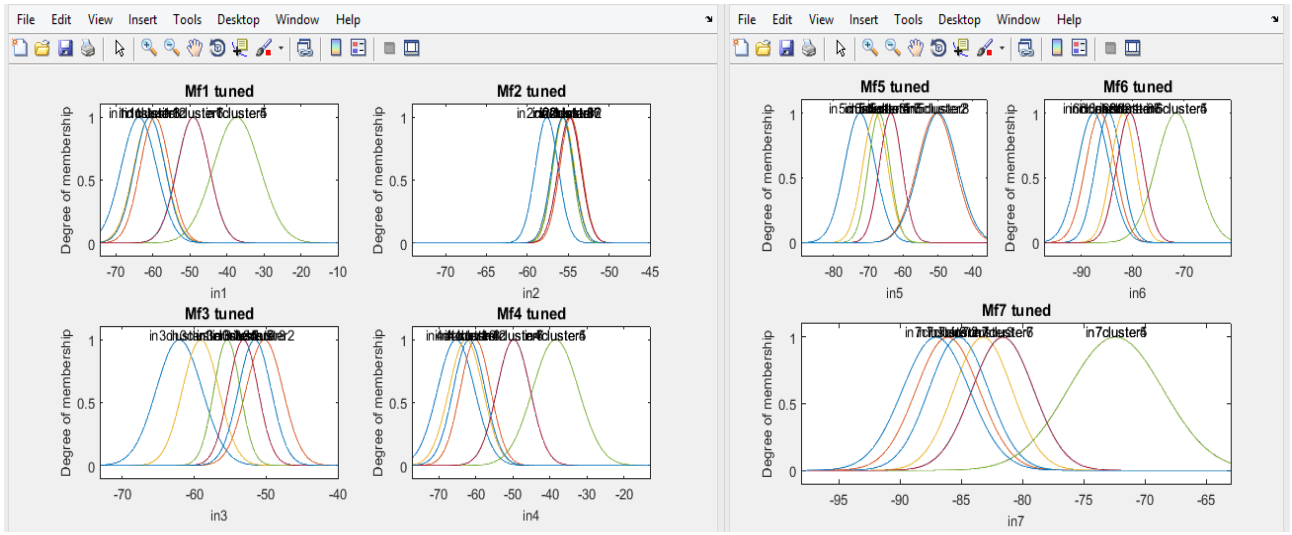
TSK model 2 με 8 κανόνες

Στο μοντέλο 2 έχουμε 8 κανόνες, Όπως και στο μοντέλο 1 η εκπαίδευση του μοντέλου έχει γίνει για 300 εποχές. Παρακάτω παρουσιάζονται οι μορφές των ασαφών συνόλων πριν και μετά την διαδικασία εκπαίδευσης του μοντέλου για τις 7 εισόδους/χαρακτηριστικά.

Ασαφής σύνολα πριν την εκπαίδευση



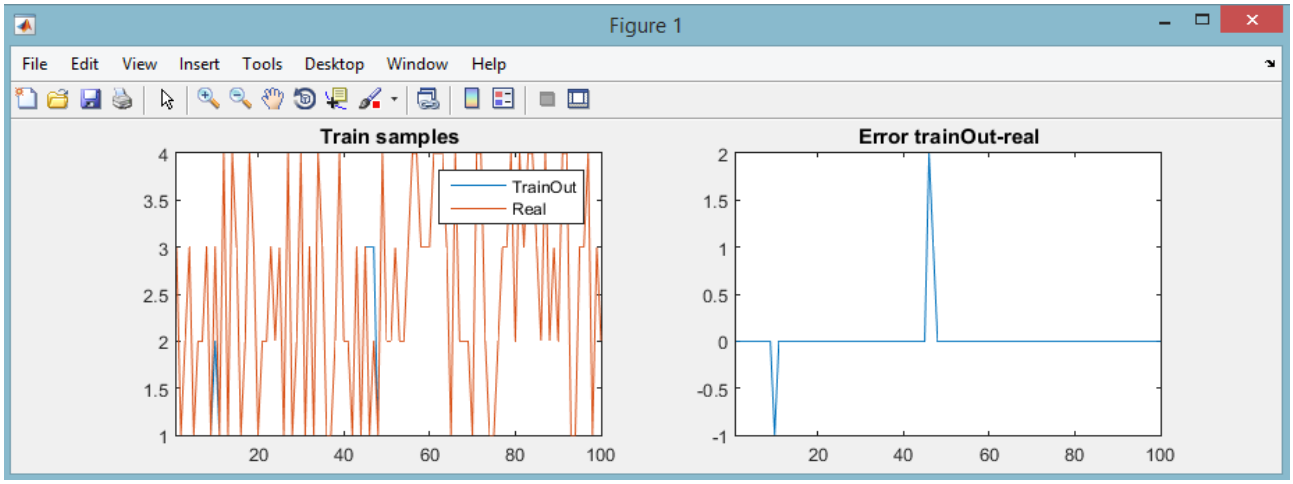
Ασαφή σύνολα μετά την εκπαίδευση του μοντέλου



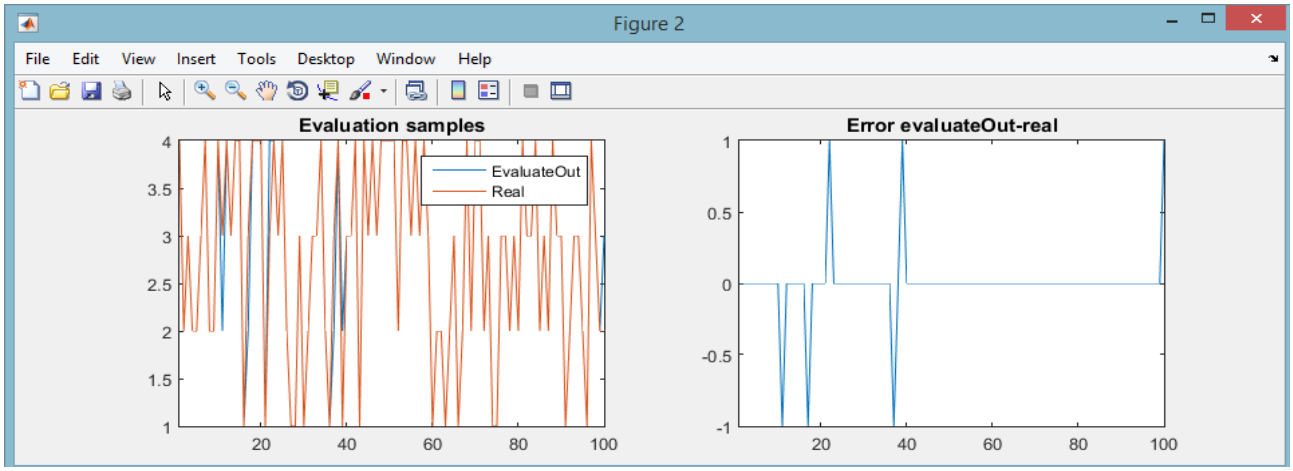
Όπως και στην προηγούμενη βλέπουμε ότι η μεταβολές στα ασαφή σύνολα μετά την διαδικασία της εκπαίδευσης δεν είναι αισθητή,αλλά η μικρές αλλαγές που έγινα από τον αλγόριθμο οπισθοδιάδοσης (backpropagation algorithm) ήταν αρκετές για να παράξουν ένα μοντέλο με καλή απόδοση.

Τα διαγράμματα εξόδου σε σύγκριση με την πραγματική τιμή και τα διαγράμματα σφάλματος για διάστημα 100 τιμών είναι:

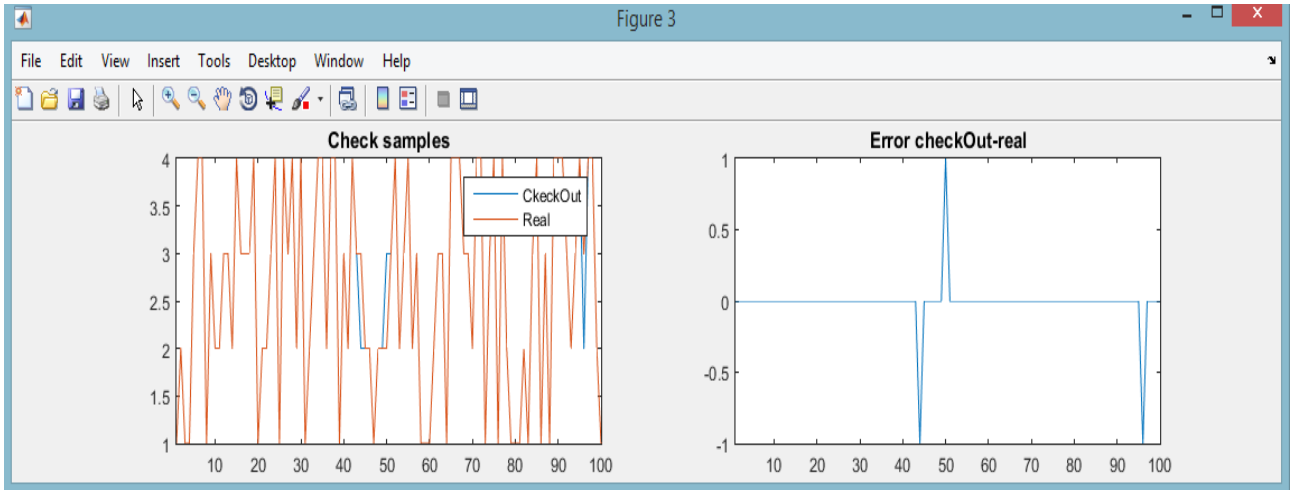
Δεδομένα εκπαίδευσης



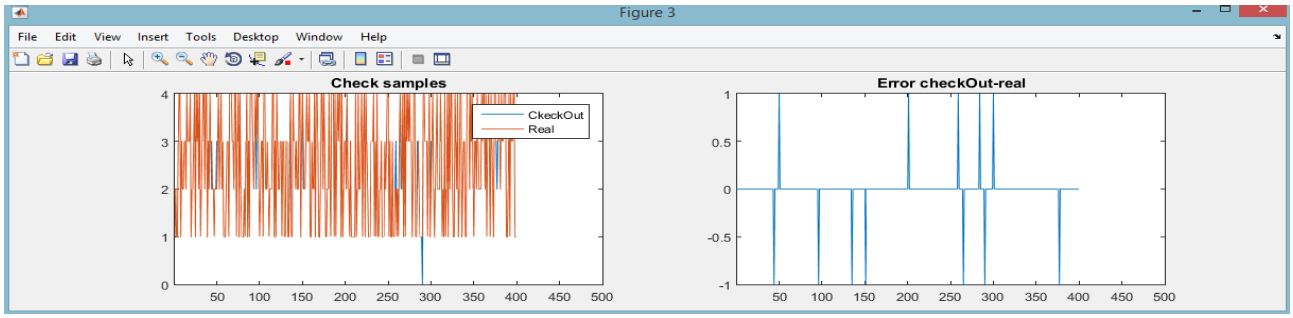
Δεδομένα επικύρωσης



Δεδομένα ελέγχου

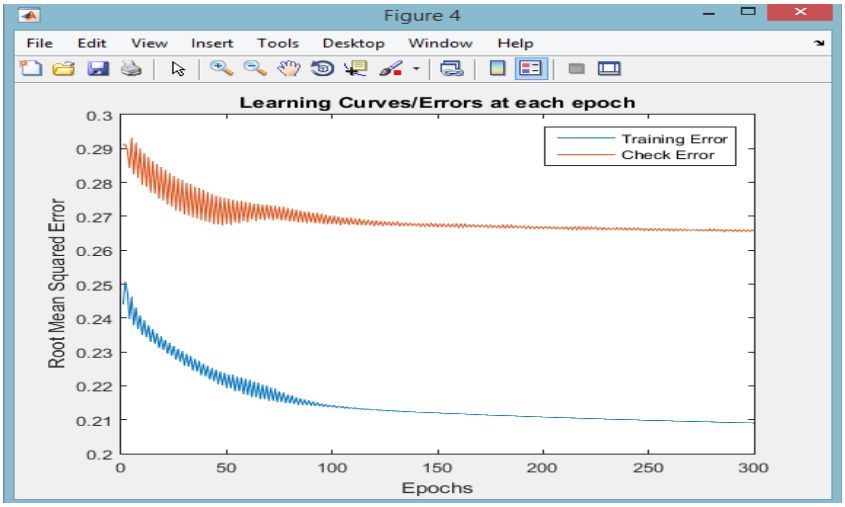


Απεικόνιση ολόκληρου του συνόλου πρόβλεψης των check δεδομένων



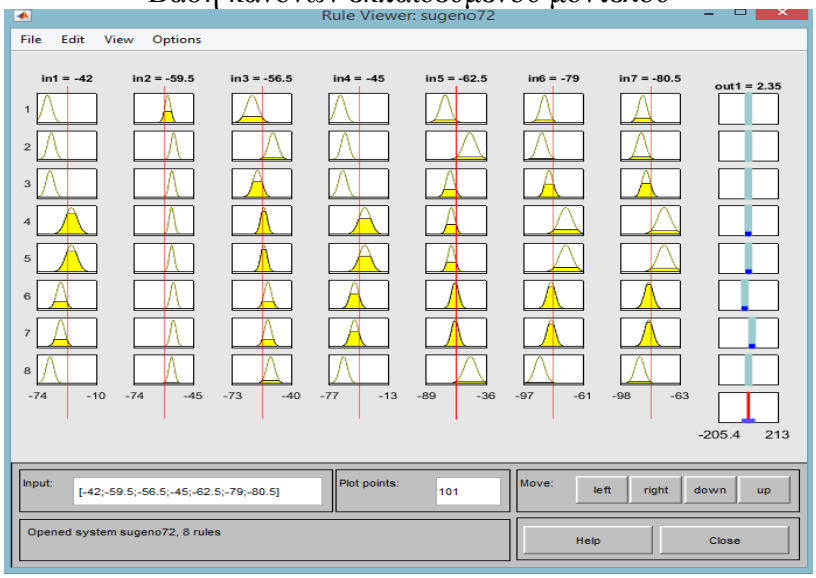
Με βάση τα διαγράμματα σφάλματος παρατηρούμε πολύ καλή απόδοση για το μοντέλο,μπορούμε να μετρήσουμε μόνο 12 σφάλματα στις 400 προβλέψεις,πολύ κοντά με το προηγούμενο μοντέλο που είχε μόλις 10 σφάλματα.

Διάγραμμα μάθησης



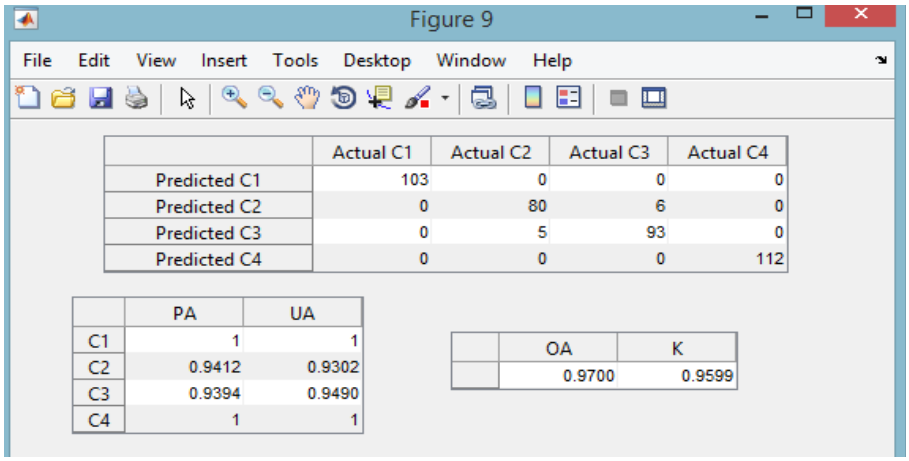
Η εκπαίδευση έχει γίνει για 300 εποχές. Στο διάγραμμα βλέπουμε κάποια ανεβοκατεβάσματα στο σφάλμα και το οποίο μπορεί να συμβαίνει λόγω υπερεκπαίδευσης, συνολικά βλέπουμε το σφάλμα τελικά κατά την διάρκεια της εκπαίδευση ελαττώθηκε και πήρε μια ελάχιστη τιμή για την εκπαίδευση των 300 εποχών.

Βάση κανόνων εκπαιδευμένου μοντέλου



Τέλος δίνεται ο πίνακας σφαλμάτων ταξινόμησης(πάνω πίνακας) και οι τιμές των τιμών απόδοσης (κάτω πίνακες). Για την απεικόνιση τους έχει χρησιμοποιηθεί η συνάρτηση unitable του Matlab.

Πίνακας απόδοσης



	Actual C1	Actual C2	Actual C3	Actual C4
Predicted C1	103	0	0	0
Predicted C2	0	80	6	0
Predicted C3	0	5	93	0
Predicted C4	0	0	0	112

	PA	UA
C1	1	1
C2	0.9412	0.9302
C3	0.9394	0.9490
C4	1	1

OA	K
0.9700	0.9599

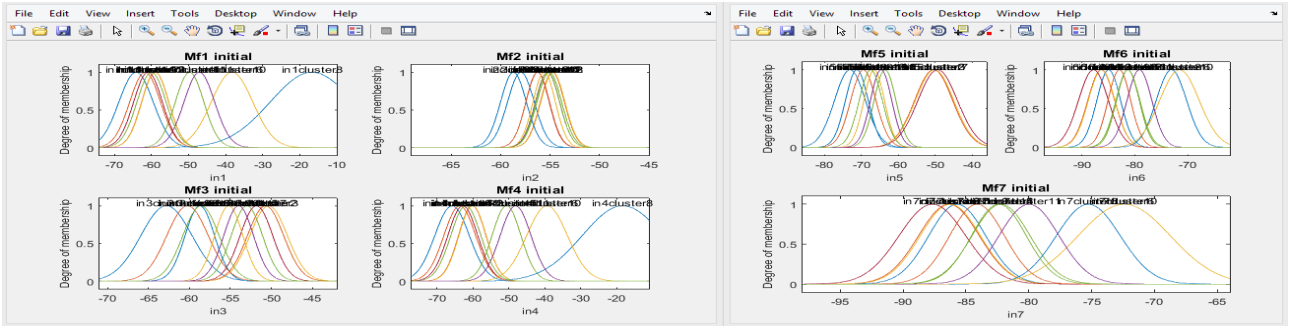
Πολύ καλή απόδοση βλέπουμε και σε αυτό το μοντέλο. Ειδικά από το πίνακα απόδοσης βλέπουμε ότι για την πρώτη και τέταρτη κατηγορία ταξινόμησης έχουμε 100% επιτυχία. Για τις δυο ενδιάμεσες έχουμε κάποια σφάλματα αλλά είναι πολύ μικρά.

Συνεχίζουμε με την παράθεση των διαγραμμάτων για τα 2 εναπομείναντα μοντέλα.

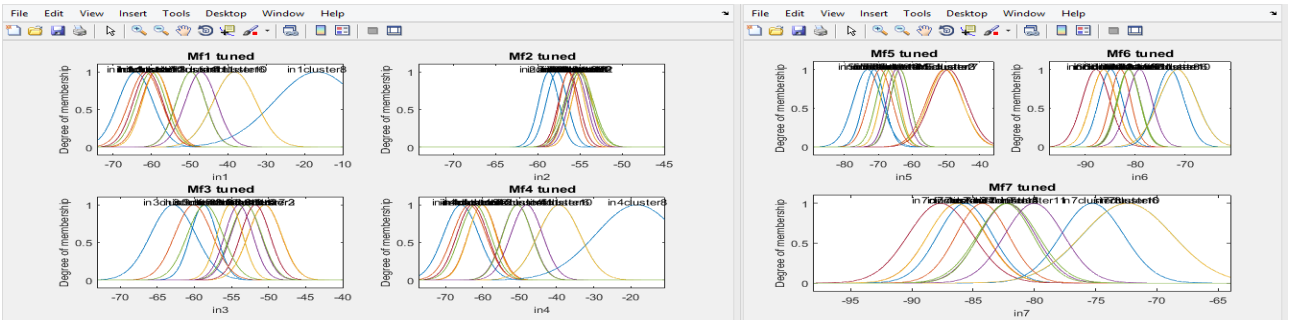
TSK model 3 με 12 κανόνες

Στο μοντέλο 3 έχουμε 12 κανόνες. Όπως και στα προηγούμενα μοντέλα η εκπαίδευση του μοντέλου έχει γίνει για 300 εποχές. Παρακάτω παρουσιάζονται οι μορφές των ασαφών συνόλων πριν και μετά την διαδικασία εκπαίδευσης του μοντέλου για τις 7 εισόδους/χαρακτηριστικά.

Ασαφής σύνολα πριν την εκπαίδευση



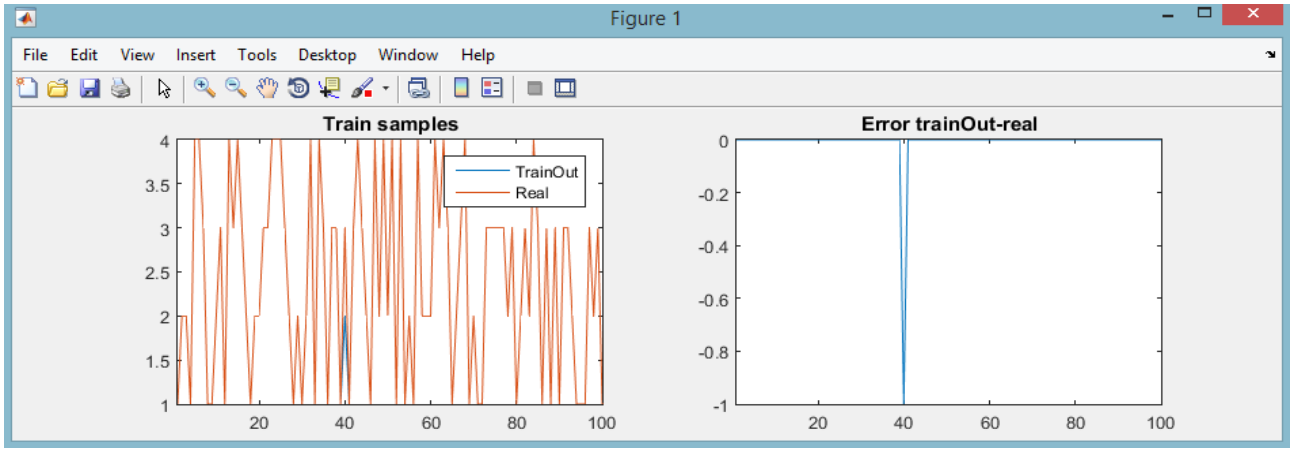
Ασαφή σύνολα μετά την εκπαίδευση του μοντέλου



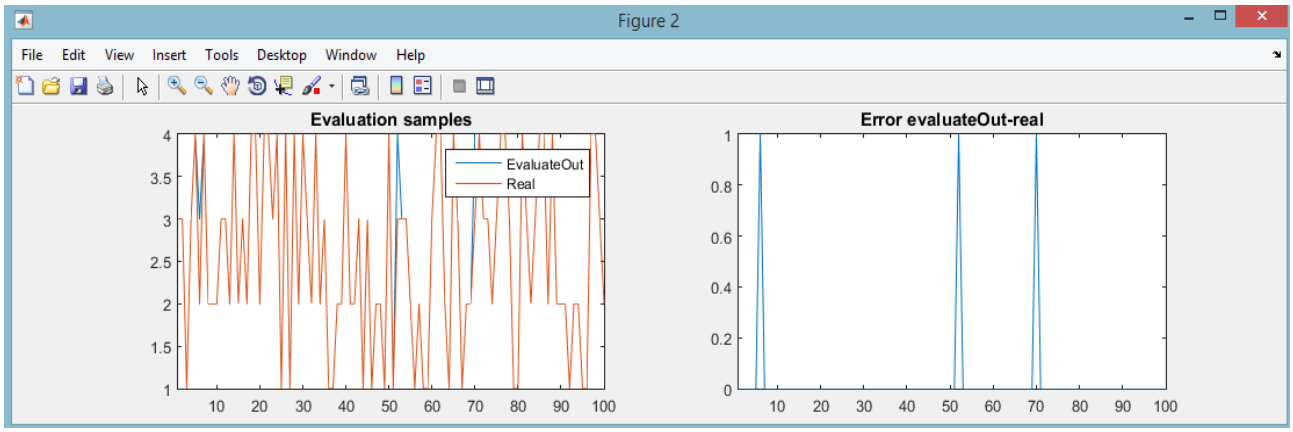
Τα ασαφή σύνολα μετά την εκπαίδευση τους έχουν προσαρμοστεί κατάλληλα από την συνάρτηση anfis και οι παράμετροι τους έχουν βελτιστοποιηθεί μέσω της μεθόδου της οπισθοδιάδοσης (backpropagation algorithm) που χρησιμοποιείται από την υβριδική μέθοδο.

Τα αποτελέσματα από τις εξόδους του εκπαιδευμένου μοντέλου σε σχέση με τις πραγματικές τιμές για κάθε κατηγορία δεδομένων φαίνεται παρακάτω. Επιπλέον στα δεξιά έχουμε το σφάλμα της πραγματικής τιμή από την τιμή της πρόβλεψης. Τα διαγράμματα έχουν γίνει για διάστημα 100 τιμών για την καλύτερη απεικόνιση των δεδομένων.

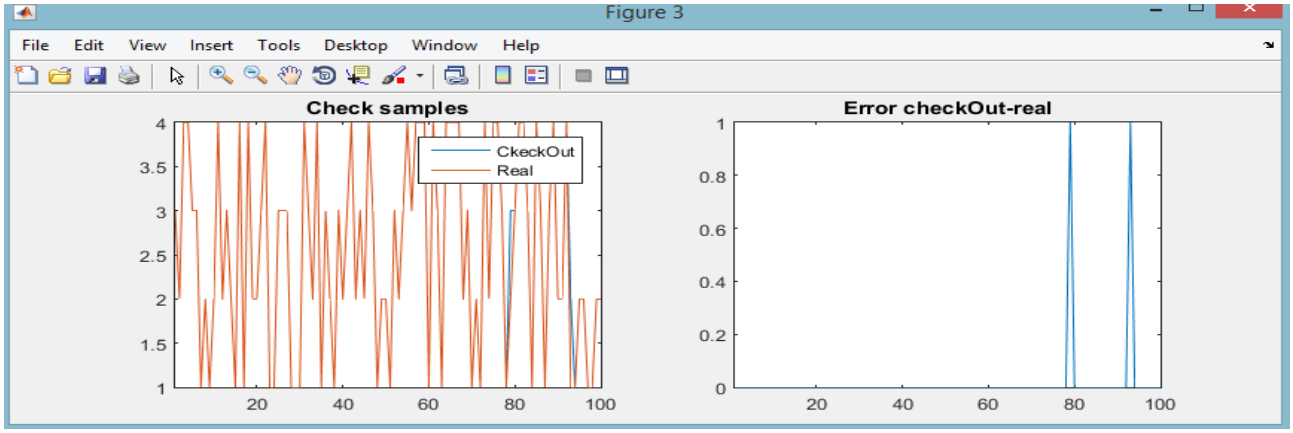
Δεδομένα εκπαίδευσης



Δεδομένα επικύρωσης

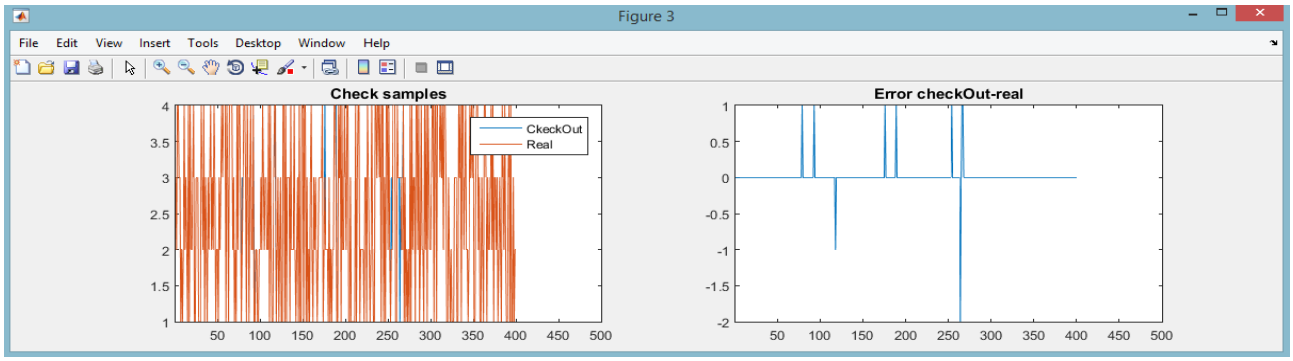


Δεδομένα ελέγχου



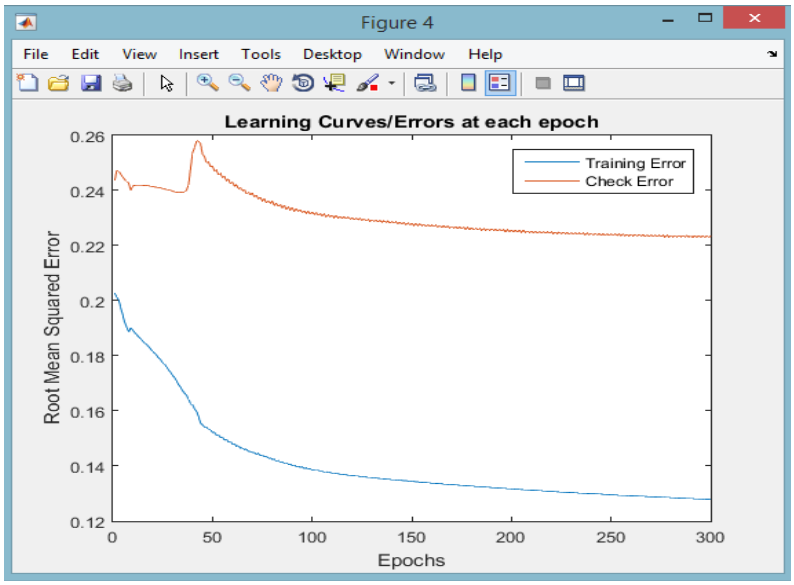
Έχουμε ποσοστά επιτυχίας 99%,97% και 98% τα οποία είναι συγκρίσιμα με τα προηγούμενα μοντέλα.

Απεικόνιση ολόκληρου του συνόλου πρόβλεψης των check δεδομένων



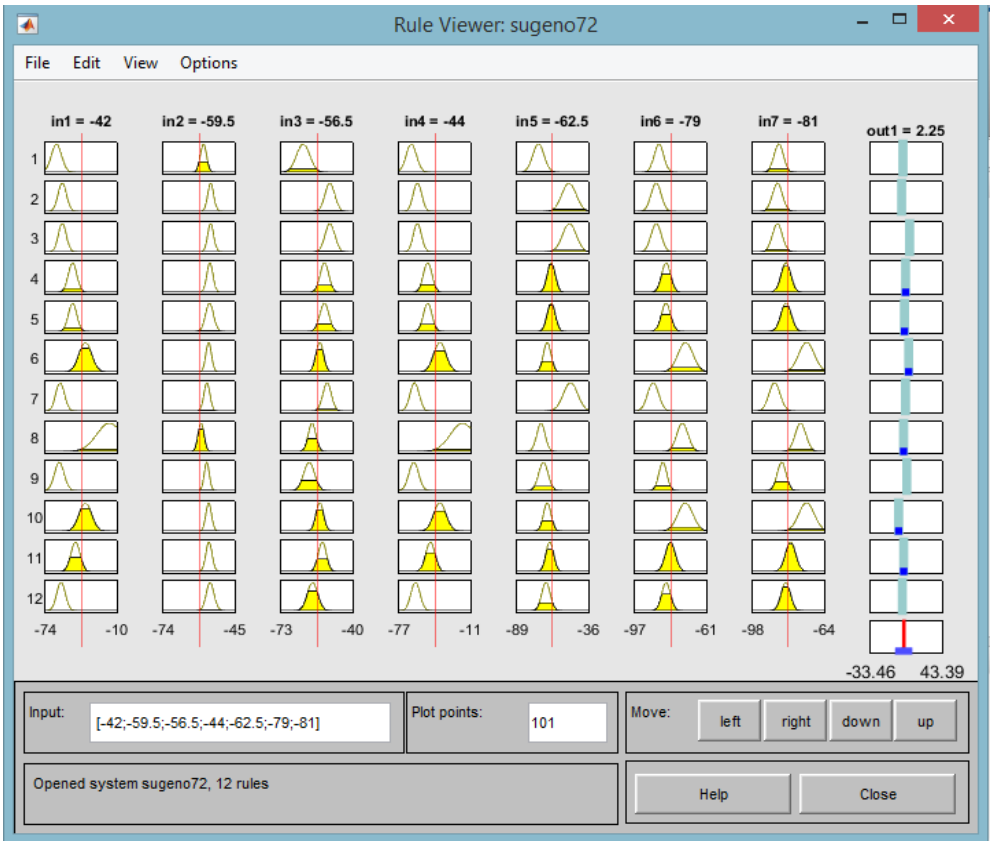
Για όλο το σύνολο των προβλέψεων έχουμε μόλις 8 λάθος προβλέψεις σε σύνολο 400 προβλέψεων, άρα με ποσοστό επιτυχίας 392/400 = 98 % .

Διάγραμμα μάθησης



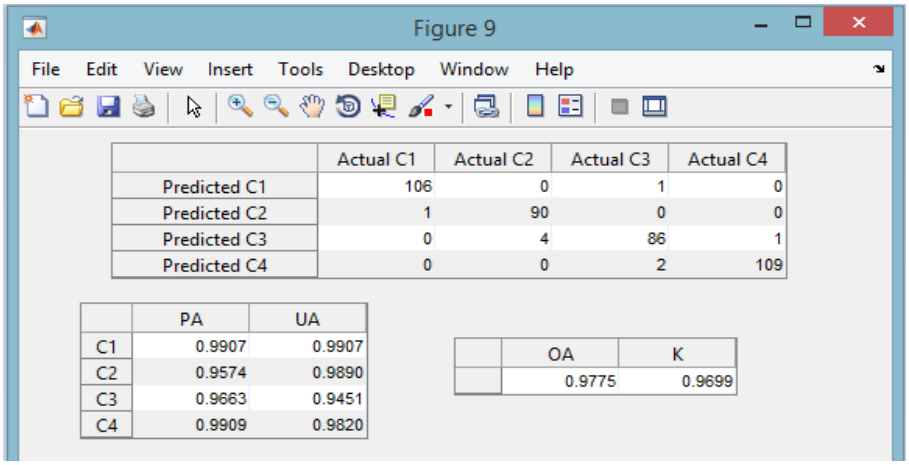
Η εκπαίδευση έχει γίνει για 300 εποχές. Συνολικά βλέπουμε το σφάλμα κατά την διάρκεια της εκπαίδευση ελαττώθηκε και πήρε μια ελάχιστη τιμή για την εκπαίδευση των 300 εποχών. Παρατηρούμε όμως μία απότομη αύξηση του σφάλματος για τα evaluation δεδομένα που μπαίνουν ως παράμετρος στην συνάρτηση anfis και αντιστοιχούν στην καμπύλη check του παραπάνω σχήματος(την καμπύλη την ονομάζουμε check απλά επειδή έτσι ονομάζεται και από το manual του anfis του matlab, δεν έχει όμως σχέση με τα check δεδομένα). Αυτή η αύξηση υποδηλώνει το πρόβλημα υπερεκπαίδευσης που πιθανόν αρχίζει να εμφανίζεται λόγο αύξησης του αριθμού κανόνων. Παρόλα αυτά το μοντέλο τελικά δεν υπερεκπαιδεύεται και παίρνει την βέλτιστη τιμή του,δηλαδή την ελάχιστη τιμή για το σφάλμα των evaluation δεδομένων.

Βάση κανόνων εκπαιδευμένου μοντέλου



Τέλος δίνεται ο πίνακας σφαλμάτων ταξινόμησης(πάνω πίνακας) και οι τιμές των τιμών απόδοσης (κάτω πίνακες). Για τον υπολογισμό των πινάκων έχουν χρησιμοποιηθεί τα check δεδομένα.Για την απεικόνιση τους έχει χρησιμοποιηθεί η συνάρτηση unitable του Matlab.

Πίνακας απόδοσης



	Actual C1	Actual C2	Actual C3	Actual C4
Predicted C1	106	0	1	0
Predicted C2	1	90	0	0
Predicted C3	0	4	86	1
Predicted C4	0	0	2	109

	PA	UA
C1	0.9907	0.9907
C2	0.9574	0.9890
C3	0.9663	0.9451
C4	0.9909	0.9820

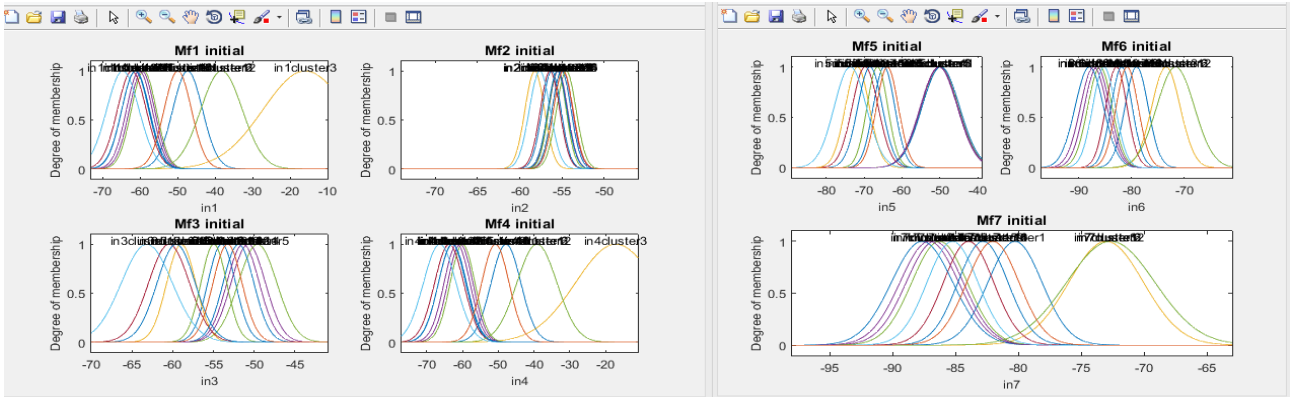
OA	K
0.9775	0.9699

Η απόδοση του μοντέλου είναι πολύ καλή και το μοντέλο κάνει καλές προβλέψεις, παρατηρούμε μεγάλες τιμές για τους δείκτες OA και K όπως και για τις τιμές PA και UA.

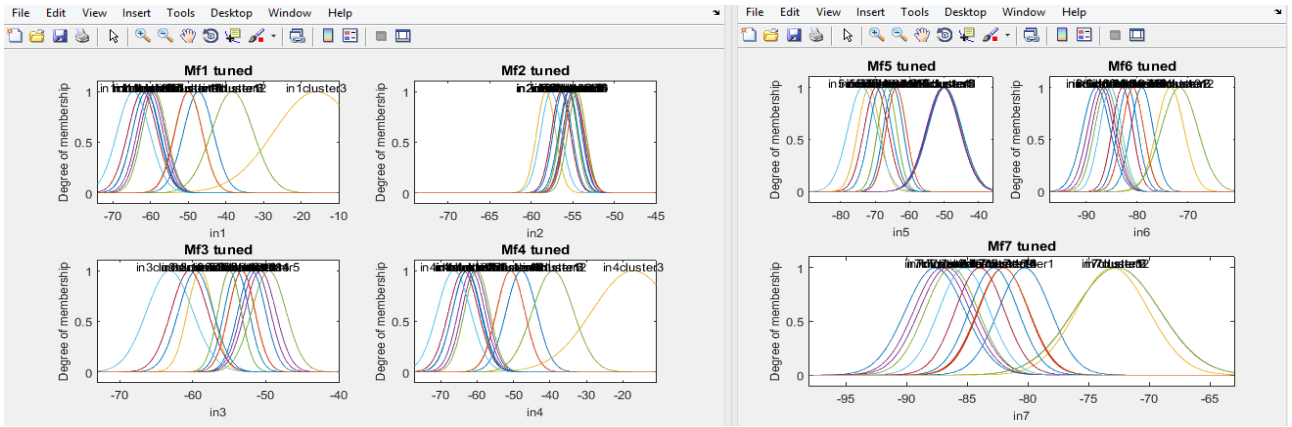
TSK model 4 με 16 κανόνες

Για το μοντέλο 4 έχουμε 16 κανόνες. Η εκπαίδευση του μοντέλου έχει γίνει για 300 εποχές. Παρακάτω παρουσιάζονται οι μορφές των ασαφών συνόλων πριν και μετά την διαδικασία εκπαίδευσης του μοντέλου για τις 7 εισόδους/χαρακτηριστικά.

Ασαφής σύνολα πριν την εκπαίδευση



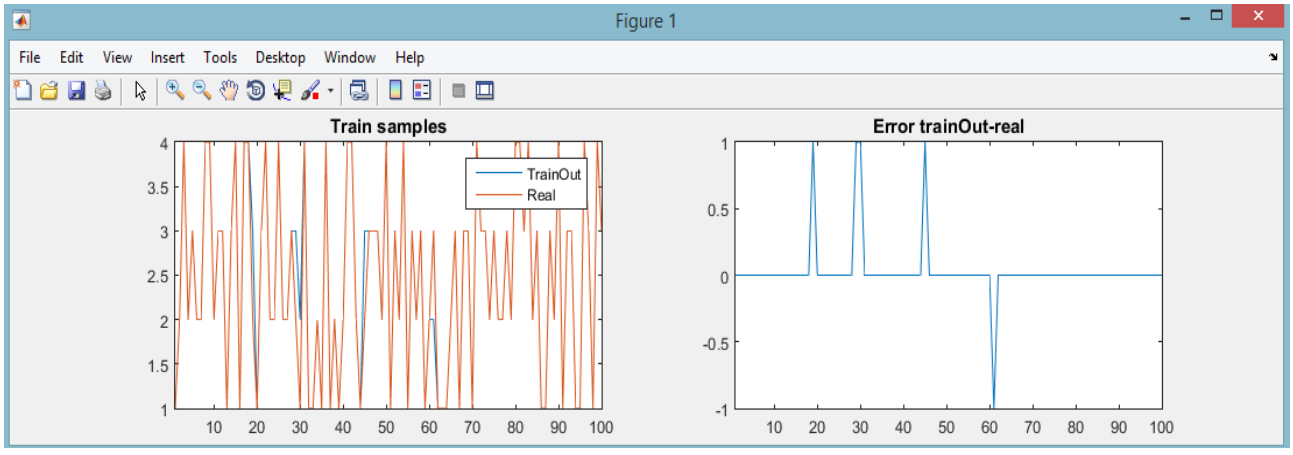
Ασαφή σύνολα μετά την εκπαίδευση του μοντέλου



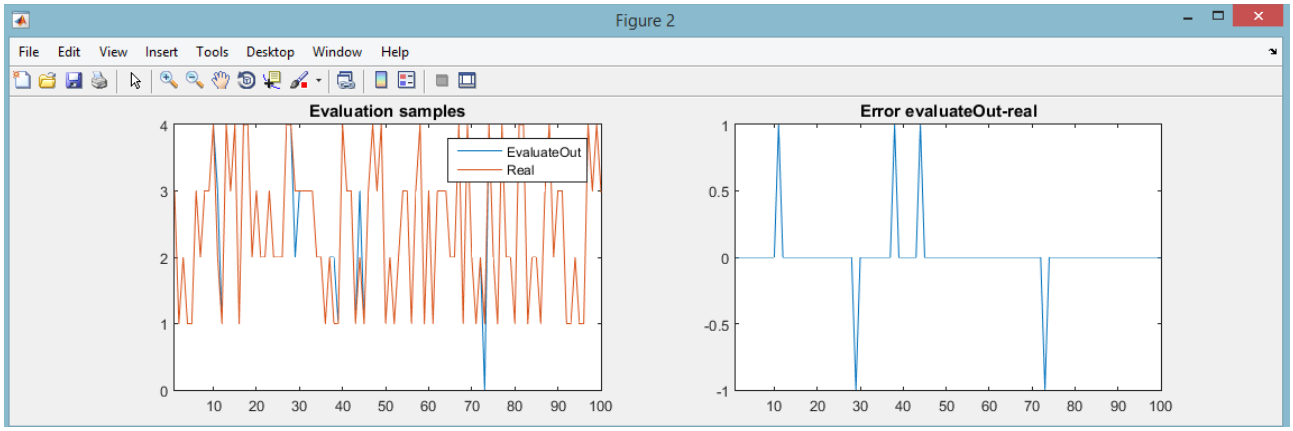
Η μεταβολές των ασαφών συνόλων είναι πολύ μικρές και δύσκολα παρατηρήσιμες.

Τα διαγράμματα έχουν γίνει για διάστημα 100 τιμών για την καλύτερη απεικόνιση των δεδομένων.

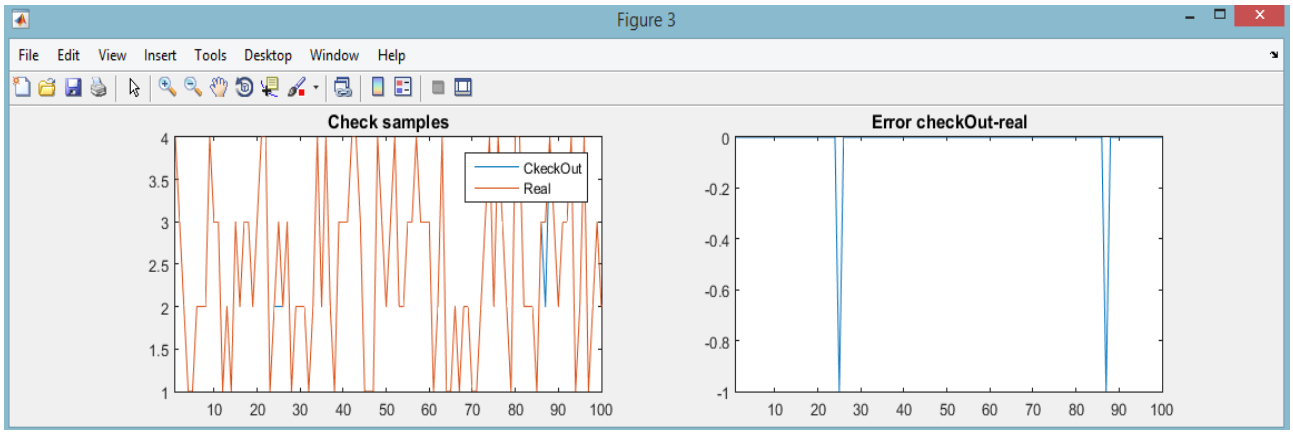
Δεδομένα εκπαίδευσης



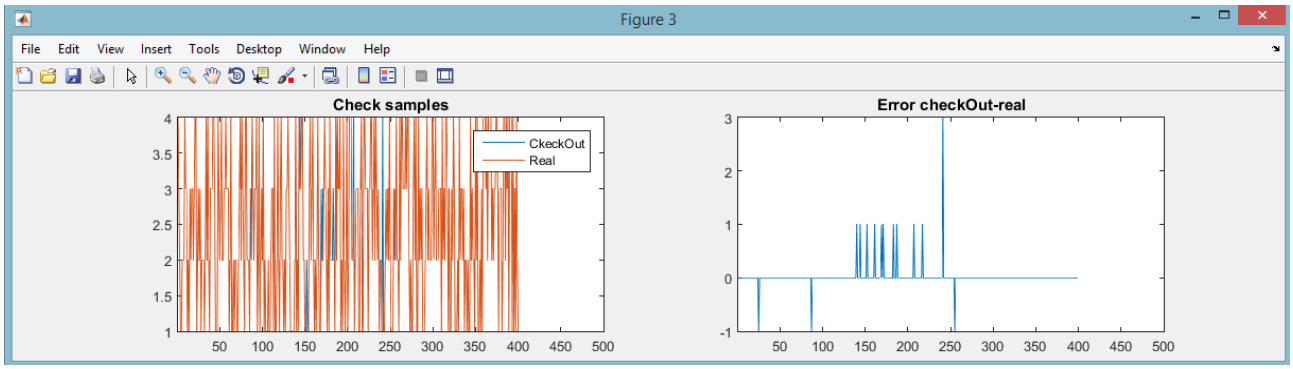
Δεδομένα επικύρωσης



Δεδομένα ελέγχου



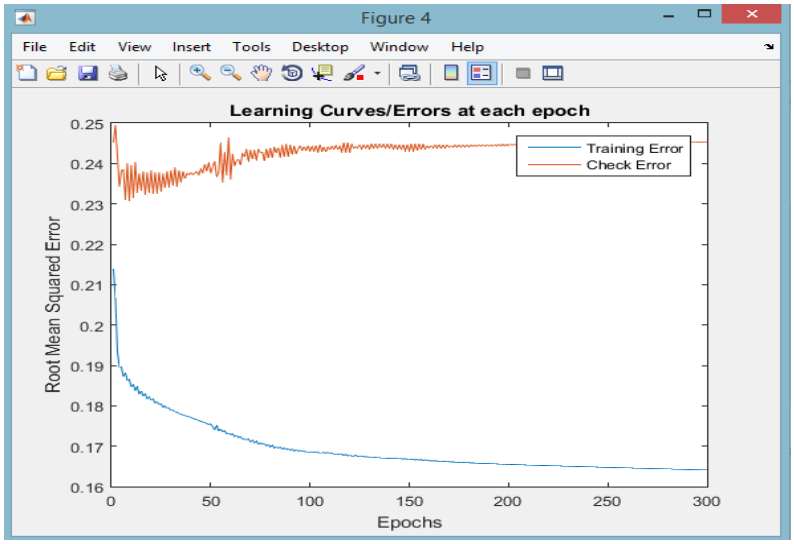
Απεικόνιση ολόκληρου του συνόλου πρόβλεψης των check δεδομένων



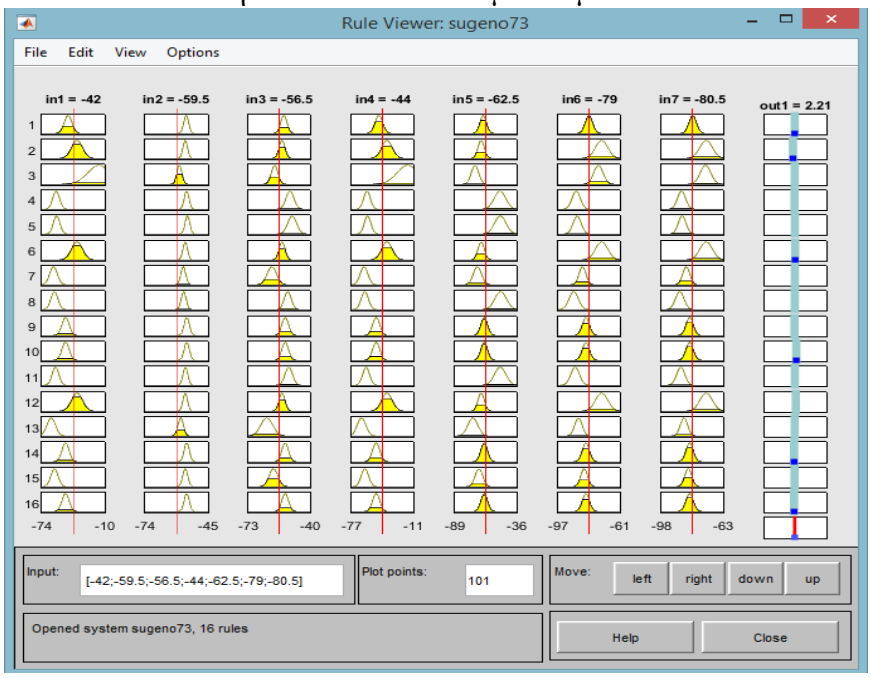
Το μοντέλο όπως βλέπουμε δεν έχει την ίδια απόδοση με τα προηγούμενα μοντέλα.

Η εκπαίδευση έχει γίνει για 300 εποχές. Από το παρακάτω διάγραμμα μάθησης βλέπουμε ότι η εκπαίδευση του μοντέλου λόγο αύξησης του αριθμού των κανόνων εμφανίζει κάποια ανεπιθύμητα προβλήματα που σχετίζονται με την υπερεκπαίδευση. Όπως φαίνεται στο σχήμα ενώ το σφάλμα των training δεδομένων μειώνεται, το σφάλμα των evaluation δεδομένων της συνάρτησης anfis αρχικά μειώνεται αλλά πολύ γρήγορα σταματάει την μείωση και ξεκινάει να αυξάνεται. Αυτό είναι προφανώς πρόβλημα υπερεκπαίδευσης του μοντέλου, το μοντέλο προσαρμόζεται στα δεδομένα εκπαίδευσης και χάνει την ικανότητα του να κάνει προβλέψεις και για άλλα δεδομένα,αυτό προφανώς δεν είναι επιθυμητό.

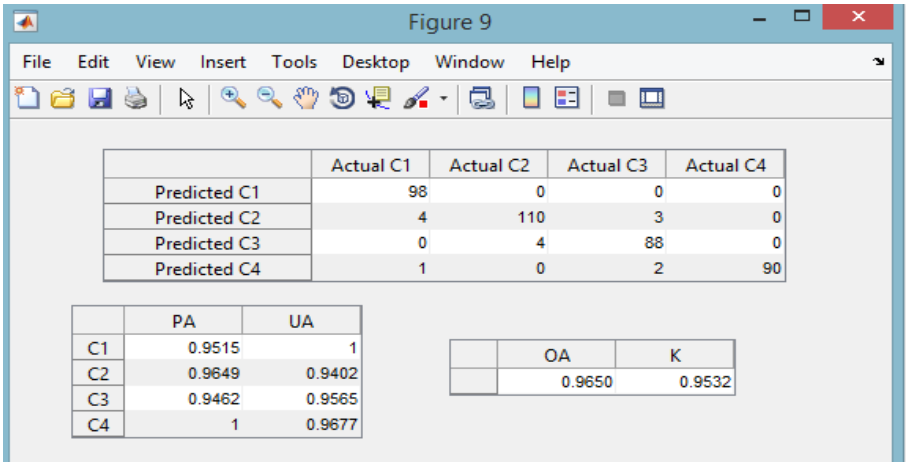
Διάγραμμα μάθησης



Βάση κανόνων εκπαιδευμένου μοντέλου



Πίνακας απόδοσης



	Actual C1	Actual C2	Actual C3	Actual C4
Predicted C1	98	0	0	0
Predicted C2	4	110	3	0
Predicted C3	0	4	88	0
Predicted C4	1	0	2	90

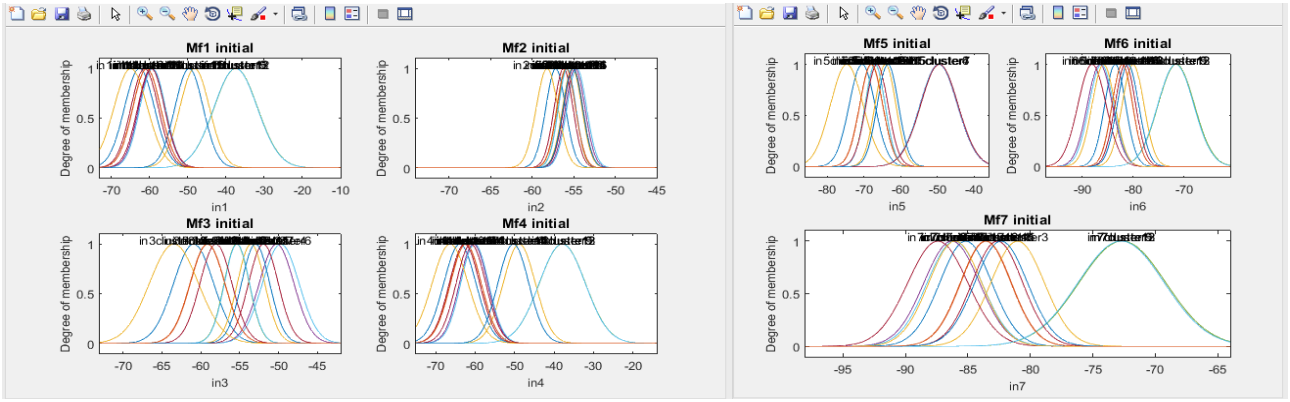
	PA	UA
C1	0.9515	1
C2	0.9649	0.9402
C3	0.9462	0.9565
C4	1	0.9677

OA	K
0.9650	0.9532

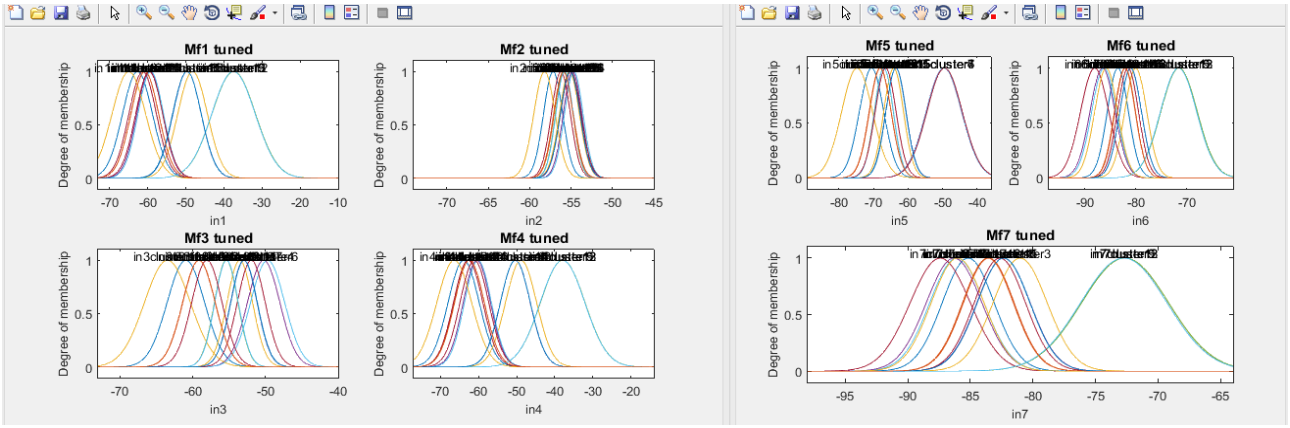
Επανάληψη εκπαίδευσης 4ου μοντέλου

Όπως είδαμε από την προηγούμενη ανάλυση λόγο του αυξημένου αριθμού των κανόνων το μοντέλο παρουσιάσει πρόβλημα υπερεκπαίδευσης. Επομένως είναι ορθό να σταματήσουμε την εκπαίδευση του μοντέλου πριν ξεκινήσει να προσαρμόζεται στα δεδομένα εκπαίδευση κρατώντας την γενικότητα του. Για αυτό τον σκοπό θα ξανά εκπαιδεύσουμε το 4ο μοντέλο σταματώντας την εκπαίδευση νωρίτερα στις 45 εποχές.

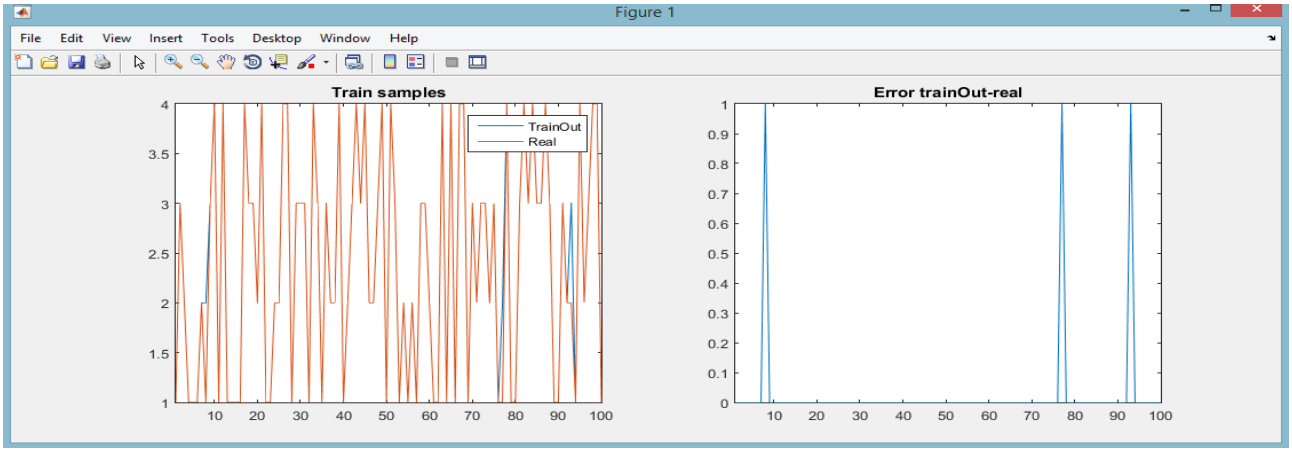
Ασαφής σύνολα πριν την εκπαίδευση



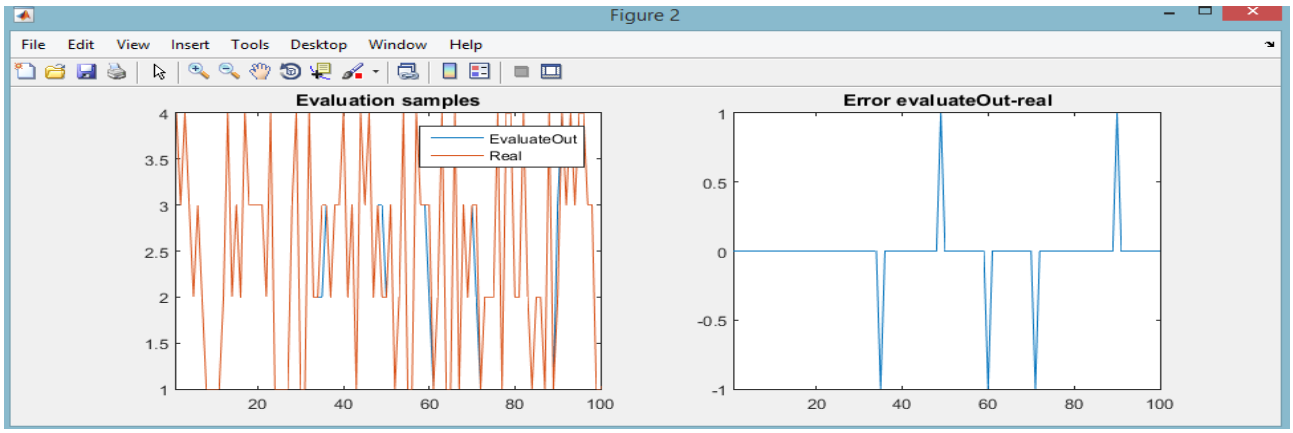
Ασαφή σύνολα μετά την εκπαίδευση του μοντέλου



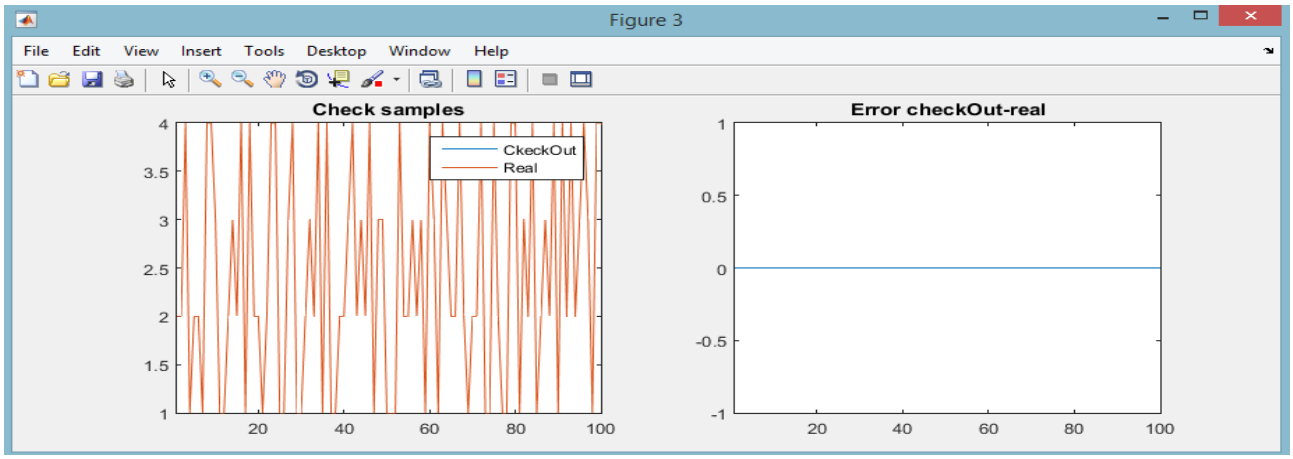
Δεδομένα εκπαίδευσης



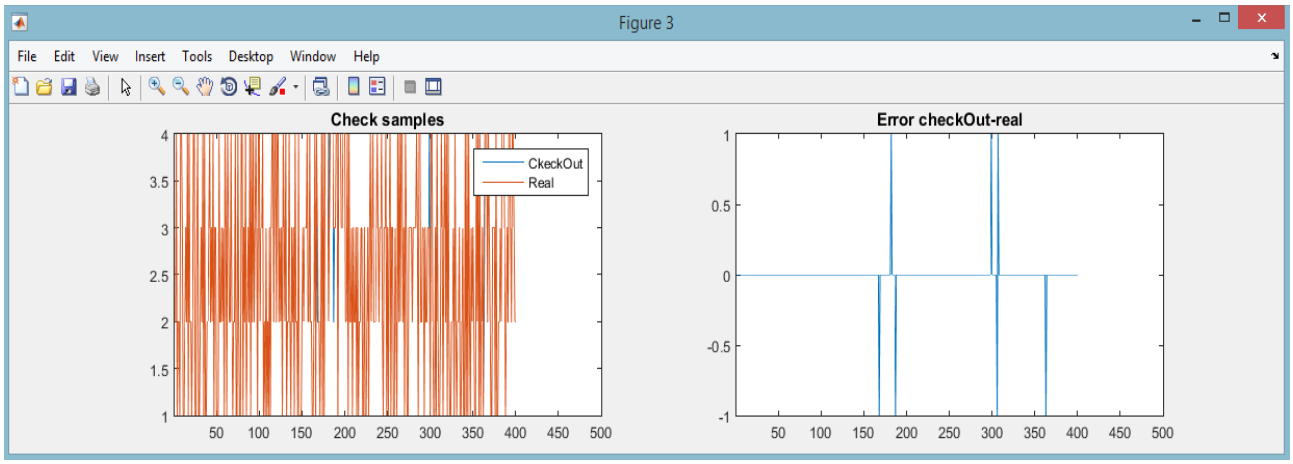
Δεδομένα επικύρωσης



Δεδομένα ελέγχου

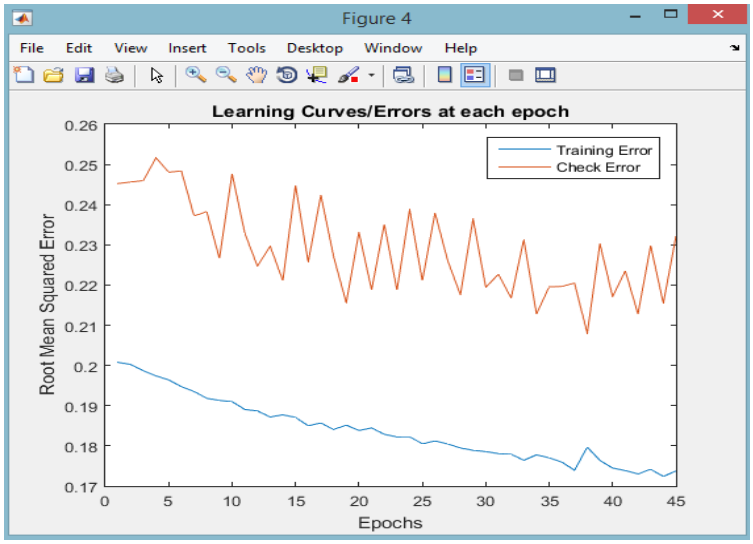


Απεικόνιση ολόκληρου του συνόλου πρόβλεψης των check δεδομένων

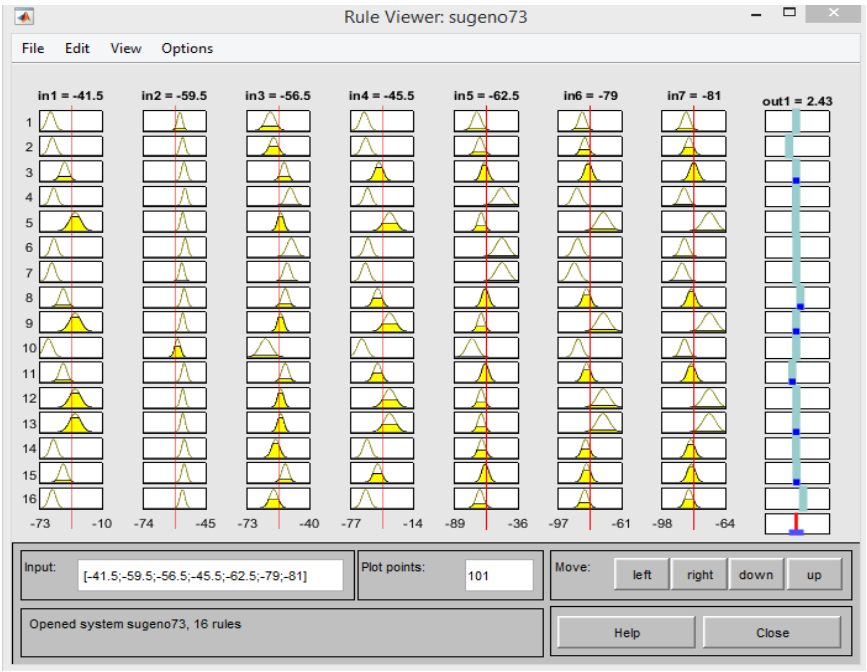


Τα αποτελέσματα είναι πολύ καλύτερα από πριν με μόλις 7 σφάλματα για τα δεδομένα ελέγχου για τις συνολικά 400 προβλέψεις,συνεχίζουμε με τα επόμενα διαγράμματα.

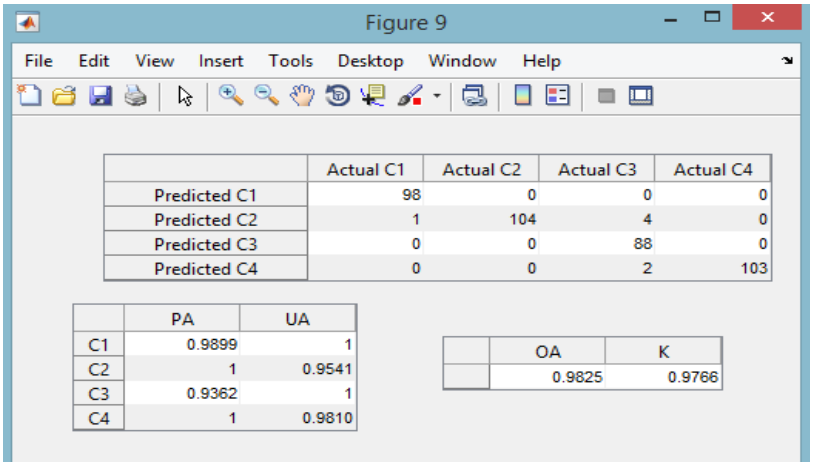
Διάγραμμα μάθησης



Βάση κανόνων εκπαιδευμένου μοντέλου



Πίνακας απόδοσης



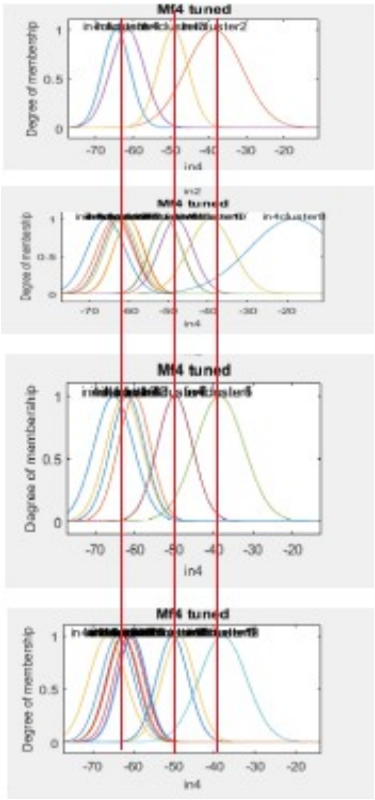
	Actual C1	Actual C2	Actual C3	Actual C4
Predicted C1	98	0	0	0
Predicted C2	1	104	4	0
Predicted C3	0	0	88	0
Predicted C4	0	0	2	103

	PA	UA
C1	0.9899	1
C2	1	0.9541
C3	0.9362	1
C4	1	0.9810

	OA	K
	0.9825	0.9766

Συμπεράσματα

Παρατηρώντας και τα 4 μοντέλα βλέπουμε ότι οι απόδοση τους είναι πολύ κοντά και δεν υπάρχουν πολύ μεγάλες διαφορές. Τα ασαφή σύνολα δεν μεταβάλλονται σε μεγάλο βαθμό κατά την εκπαίδευση τους παρόλα αυτά παρατηρούμε ότι όσο αυξάνονται οι κανόνες τόσο περισσότερα σύνολα δημιουργούνται τα οποία εξαπλώνονται για να καλύψουν την περιοχή της κάθε εισόδου. Κάθε σύνολο συγκεντρώνεται γύρο από συγκεκριμένες περιοχές για κάθε είσοδο, με αυτό τον τρόπο προσαρμόζεται η βαρύτητα της κάθε εισόδου στην εξαγωγή του συμπεράσματος, δηλαδή ο βαθμός ενεργοποίησης της εξόδου για κάποια τιμή εισόδου. Για παράδειγμα για την δεύτερη είσοδο με βάση τα διαγράμματα βλέπουμε ότι τα ασαφή σύνολα συγκεντρώνονται περίπου γύρο από τις τιμές -50 ως -60. Άρα για μία τιμή εισόδου -55 η έξοδος του κανόνα που συμπεριλαμβάνει το συγκεκριμένο ασαφή σύνολο έχει μεγάλο βαθμό ενεργοποίησης. Άρα αυτό μπορεί να δηλώνει για παράδειγμα ότι μία κατηγορία δεδομένων έχει χαρακτηριστική τιμή -55 για αυτή την είσοδο άρα το μοντέλο στην ουσία με αυτό τον τρόπο φανερώνει την κατηγορία στην οποία αντιστοιχούν τα δεδομένα εισόδου. Άρα με αυτό τον τρόπο οι είσοδοι με ασαφή σύνολα τα οποία είναι περισσότερο διασκορπισμένα χωρίς επικαλύψεις στην ουσία δείχνουν ότι αυτή η είσοδος δεν δείχνει άμεσα μια κατηγορία αλλά οι τιμές της μπορούν να φανερώσουν διάφορες κατηγορίες. Για παράδειγμα η τέταρτη είσοδος για τιμές γύρο του -65 μπορεί να δηλώνει μια κατηγορία δεδομένων,για -50 μία άλλη κατηγορία και για τιμή -35 κάποια άλλη κατηγορία δεδομένων. Συμπερασματικά, οι είσοδοι με ασαφή σύνολα για τα οποία οι προβολές τους έχουν μεγάλη επικάλυψη φανερώνουν ότι η είσοδος δεν συμμετέχει σε πολλές διαφορετικές κατηγορίες, δηλαδή μπορεί να φανερώσει ότι τα δεδομένα ανήκουν σε κάποιες συγκεκριμένες κατηγορίες, ενώ σύνολα με περισσότερο διασκορπισμένα ασαφή σύνολα δείχνουν συμμετοχή της εισόδου σε περισσότερες κατηγορίες. Για παράδειγμα για την 4η είσοδο των τεσσάρων μοντέλων όπως βλέπουμε στο παρακάτω σχήμα αριστερά, τα ασαφή σύνολα και των 4 μοντέλων συγκεντρώνονται γύρο από συγκεκριμένες τιμές,όπως φαίνεται από τις κατακόρυφες κόκκινες γραμμές, και αυτό πιθανόν δηλώνει ότι αυτές οι τιμές χαρακτηρίζουν κάποιες από τις κατηγορίες εξόδου. Για τα μοντέλα με περισσότερους κανόνες άρα και ασαφή σύνολα, τα παραπάνω σύνολα συγκεντρώνονται περίπου γύρο από την ίδια περιοχή δημιουργώντας κοινές προβολές.



chkData

400x8 double

	1	2	3	4	5	6	7	8	9
128	-65	-57	-55	-64	-50	-92	-85	4	
129	-45	-54	-48	-49	-65	-78	-81	3	
130	-51	-57	-59	-45	-77	-75	-81	2	
131	-62	-58	-49	-64	-42	-86	-89	4	
132	-61	-57	-59	-62	-69	-78	-83	1	
133	-39	-54	-56	-40	-70	-77	-78	2	
134	-34	-57	-55	-35	-59	-71	-69	2	
135	-64	-60	-53	-60	-48	-89	-89	4	
136	-40	-54	-49	-43	-62	-70	-65	2	
137	-62	-56	-59	-68	-74	-95	-84	1	
138	-50	-49	-52	-56	-69	-80	-81	3	
139	-19	-60	-56	-36	-63	-70	-75	2	
140	-46	-54	-53	-48	-64	-78	-79	3	
141	-64	-57	-60	-70	-69	-86	-83	1	
142	-64	-57	-47	-58	-46	-89	-86	4	
143	-16	-58	-58	-13	-68	-76	-90	2	
144	-41	-52	-57	-41	-63	-73	-68	2	
145	-66	-58	-64	-63	-73	-83	-86	1	
146	-59	-52	-56	-58	-55	-88	-87	4	
147	-68	-56	-66	-69	-75	-82	-83	1	

Command Window

Για περαιτέρω επεξήγηση θα χρησιμοποιήσου ένα απόσπασμα από τα δεδομένα τα οποία απεικονίζονται στο παραπάνω από δεξιά σχήμα. Με κόκκινο έχουμε σημειώσει την 4η είσοδο και τις κατηγορίες(έξοδο), με μπλε έχουμε σημειώσει την κατηγορία που αντιστοιχεί σε αριθμό εξόδου 4 και με πράσινο την κατηγορία που αντιστοιχεί σε αριθμό εξόδου 2. Μπορούμε να παρατηρήσουμε ότι για την κατηγορία 4 η είσοδος 4 παίρνει τιμές περίπου στο 60 ενώ για την κατηγορία 2 η είσοδος 4 παίρνει τιμές γύρο στο 40. Αυτή η κατανομή φαίνεται και στο σχήμα αριστερά με τα ασαφή σύνολα, δηλαδή η είσοδος 4 μπορεί να μας δώσει την πληροφορία για να συμπεράνουμε αν τα δεδομένα μας ανήκουν στην κατηγορία 4 ή 2.

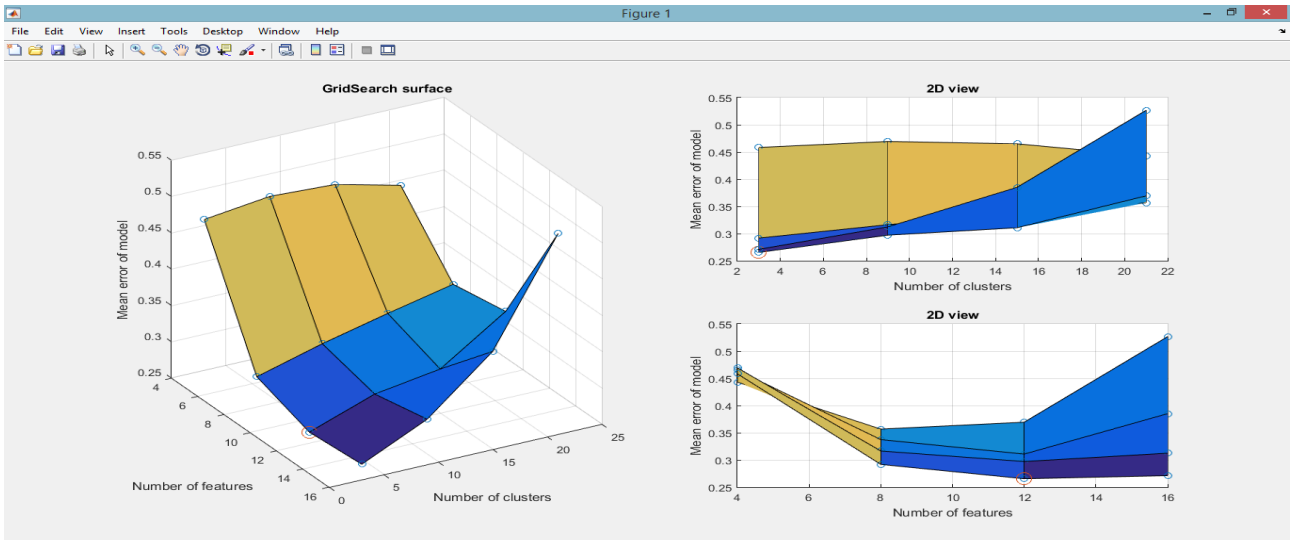
Από ότι είδαμε από τα προηγούμενα διαγράμματα των μοντέλων,καθώς ο αριθμός των κανόνων αυξάνεται,δημιουργούνται προβλήματα υπερεκπαίδευσης. Από το μοντέλο 4 το πρόβλημα της υπερεκπάιδευσης μπορε να κατανοηθεί εύκολα αφού όπως είδαμε το μοντέλο που παράχθηκε μετά από 300 εποχές δεν είχε το ελάχιστο σφάλμα για τα evaluation δεδομένα. Σταματώντας επομένως την εκπαίδευση στις 45 εποχές καταφέραμε τελικά να δημιουργήσουμε ένα πολύ καλύτερο ταξινομητή, που ουσιαστικά από ότι μπορούμε να παρατηρήσουμε είχε την καλύτερη απόδοση και από τα 4 μοντέλα.
Τέλος, μια πρόταση για την βελτίωση του ταξινομητη είναι η χρήση διαφορικής μορφής ασαφών συνόλων. Επίσης θα μπορούσαμε να χρησιμοποιήσου διαφορετικό αριθμό ασαφών συνόλων ανά είσοδο, προσαρμόζοντας κατάλληλα τον αριθμό για κάθε είσοδο,αφαιρώντας με αυτό τον τρόπο πολυπλοκότητα που μπορεί να οδηγεί σε εσφαλμένα αποτελέσματα. Μία τελευταία ιδέα είναι να μην χρησιμοποιήσουμε όλες τις εισόδους καθώς κάποιες από αυτές μπορεί να αποτελούνται από θόρυβο και να εισάγουν σφάλματα,θα μπορούσαμε μέσω κάποιου αλγόριθμου όπως του Relief να επιλέξουμε τις καλύτερες εισόδους που θα μπορούσαν να χρησιμοποιηθούν ως προβλεπτές και να χρησιμοποιούσαμε μόνο αυτές τις εισόδους.

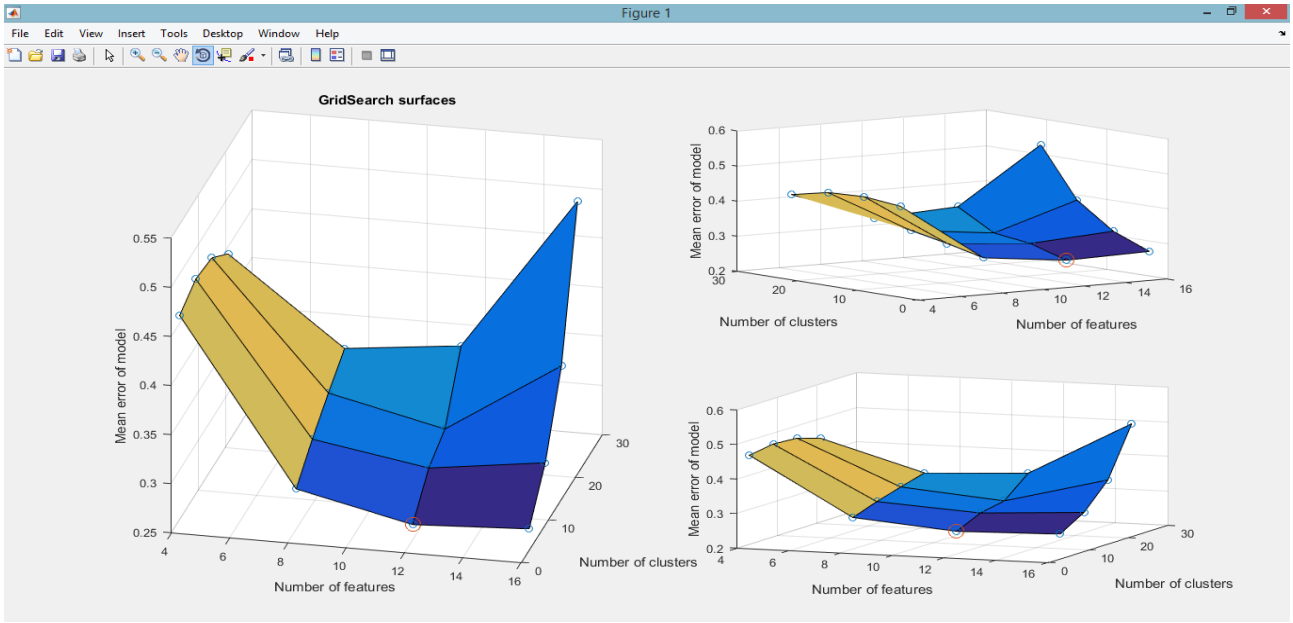
2) Εφαρμογή σε dataset με υψηλή διαστασιμότητα

Στο δεύτερο μέρος της εργασίας θα χρησιμοποιήσου τα δεδομένα Waveform Generation Dataset από το UCI repository. Θα χρησιμοποιήσου αναζήτηση πλέγματος (grid search) και αξιολόγηση μέσω 5-πτυχης διασταυρωμένης επικύρωσης (5-fold cross validation) για την επιλογή των βέλτιστων τιμών των παραμέτρων. Ως μέθοδος επιλογής χαρακτηριστικών επιλέγεται ο αλγόριθμος Relief και ως μέθοδος διαμέρισης διασκορπισμού ο αλγόριθμος Fuzzy C-Means (FCM – genfis3). Πριν να διαχωρίσουμε τα δεδομένα στις 3 γνωστές ομάδες εκπαίδευσης, επικύρωσης και ελέγχου, ανακατεύουμε τα δεδομένα για πετύχουμε ομοιόμορφο διασκορπισμό και συχνότητα εμφάνισης των δεδομένων και στις 3 κατηγορίες. Αφού βρούμε τις βέλτιστες τιμές των ελεύθερων παραμέτρων στην συνέχεια θα περάσουμε στην δοκιμή του μοντέλου και την δημιουργία των απαραίτητων διαγραμμάτων για την μελέτη του.

Αναζήτηση πλέγματος (Grid Search)

Η διάσταση του πλέγματος είναι 4x4 δηλαδή επιλέγουμε 4 διαφορετικούς αριθμούς χαρακτηριστικών και 4 διαφορετικούς αριθμούς κανόνων για κάθε αριθμό χαρακτηριστικών και η εκπαίδευση έχει γίνει για 110 εποχές. Σύμφωνα με την εκφώνηση της εργασίας οι τιμές των χαρακτηριστικών που θα χρησιμοποιηθούν δίνονται από το σύνολο $NF=\{4,8,12,16\}$ ενώ για τον αριθμό κανόνων (αριθμός των clusters), θα λαμβάνει τιμές από το σύνολο $NR=\{3,9,15,21\}$.
Κατά την διαδικασία της 5-fold επικύρωσης διαχωρίζουμε τα training data σε 5 κατηγορίες και σε κάθε επανάληψη χρησιμοποιούμε την μία από αυτές ως evaluation δεδομένα και τις υπόλοιπες 4 ως training δεδομένα. Στο τέλος υπολογίζουμε το μέσο σφάλμα που προέκυψε από όλους τους συνδυασμούς των 5 κατηγοριών. Εκτελώντας επομένως την 5-fold διασταυρωμένη επικύρωση για κάθε συνδυασμό των ελεύθερων μεταβλητών προκύπτουν τα παρακάτω διαγράμματα στα οποίο φαίνεται το μέσο σφάλμα της 5-fold διασταυρωμένης επικύρωσης συναρτήσει του αριθμού των χαρακτηριστικών και του αριθμού των κανόνων σε μορφή 3D απεικόνισης.





Σχόλια

Από το διάγραμμα φαίνεται ότι όσο αυξάνεται ο αριθμός των χαρακτηριστικών το σφάλμα μειώνεται μέχρι ο αριθμός χαρακτηριστικών να φτάσει τα 12 χαρακτηριστικά, για τα 16 χαρακτηριστικά βλέπουμε ότι το σφάλμα αυξάνεται,ελάχιστα όμως για την περίπτωση των 3 κανόνων οπότε θα μπορούσαμε να πούμε ότι η αύξηση είναι αμελητέα και τα σημεία είναι ισοδύναμα,για μικρότερη όμως υπολογιστική πολυπλοκότητα θα επιλέξουμε τα 12 χαρακτηριστικά. Φυσικά τόσο μικρές διαφορές μπορεί να είναι απλά τυχαίες. Το σφάλμα επίσης αυξάνεται και με την αύξηση των cluster εκτός της περίπτωσης που έχουμε 4 χαρακτηριστικά που όπως φαίνεται σε αυτή την περίπτωση έχουμε μια σταθερή περίπου τιμή σφάλματος και για την περίπτωση των 21 cluster μικρή μείωση στο σφάλμα. Παρόλα αυτά το σφάλμα είναι μεγάλο για 4 χαρακτηριστικά οπότε δεν έχει κάποια πρακτική αξία ο συγκεκριμένος συνδυασμός μεταβλητών. Επίσης το μεγαλύτερο σφάλμα εμφανίζεται για 16 χαρακτηριστικά και 21 κανόνες, δηλαδή τον μεγαλύτερο συνδυασμό τιμών. Το αυξημένο σφάλμα πιθανόν σχετίζεται με το πρόβλημα της υπερεκπαίδευσης που εμφανίζεται για μεγάλο αριθμό κανόνων. Η περιοχή που μας ενδιαφέρει είναι η περιοχή με το έντονο μπλε σκούρο χρώμα στο οποίο έχουμε τις ελάχιστες τιμές σφάλματος,μέσα στον κόκκινο κύκλο βρίσκεται ο βέλτιστος συνδυασμός των ελεύθερων μεταβλητών που δίνουν ελάχιστο σφάλμα.

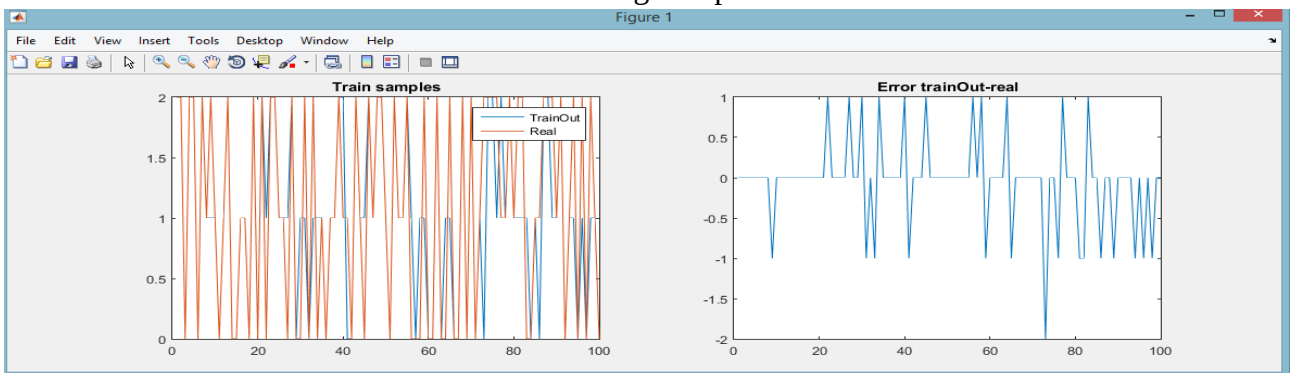
Από τα παραπάνω καταλήγουμε στο ότι οι βέλτιστες τιμές του μοντέλου είναι η χρήση NF =12 χαρακτηριστικών και NR = 3 κανόνων.

Συνεχίζουμε την ανάλυση του δεύτερου μέρους της εργασίας δημιουργώντας ένα μοντέλο με τις ίδιες παραμέτρους και το εκπαιδεύουμε με χρήση των training data και evaluation data για τον έλεγχο υπερεκπαίδευσης. Μετά την εκπαίδευση χρησιμοποιούμε το check δεδομένα για την μελέτη της απόδοσης του μοντέλου όπως ακριβώς κάναμε και στο πρώτο κομμάτι της εργασίας. Τα αποτελέσματα από την εκπαίδευση του μοντέλου είναι:

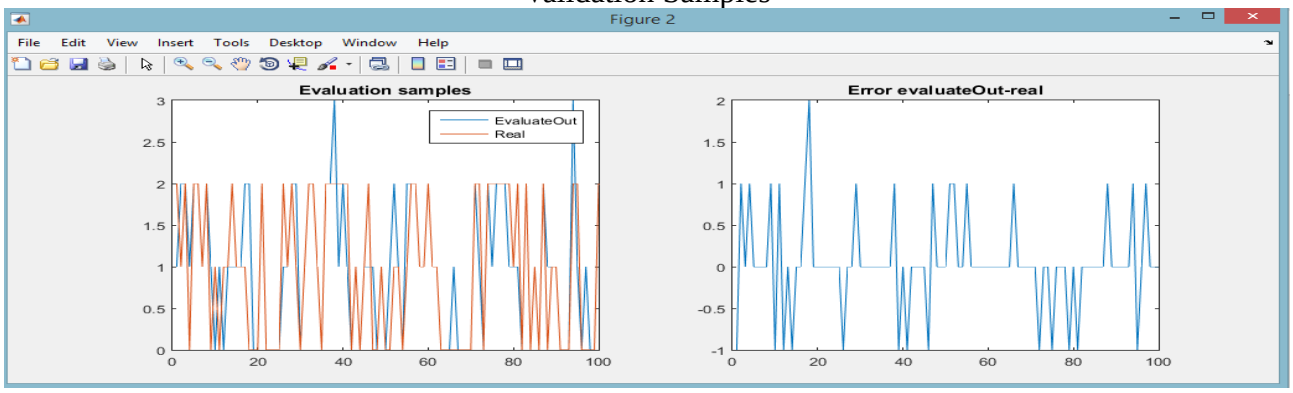
Βέλτιστο μοντέλο ταξινομητή

Η εκπαίδευση του μοντέλου έχει γίνει για 300 εποχές, τα διαγράμματα που απεικονίζουν το σφάλμα και την πραγματική τιμή της εξόδου για ένα διάστημα 100 τιμών για τις τρις κατηγορίες δεδομένων είναι.

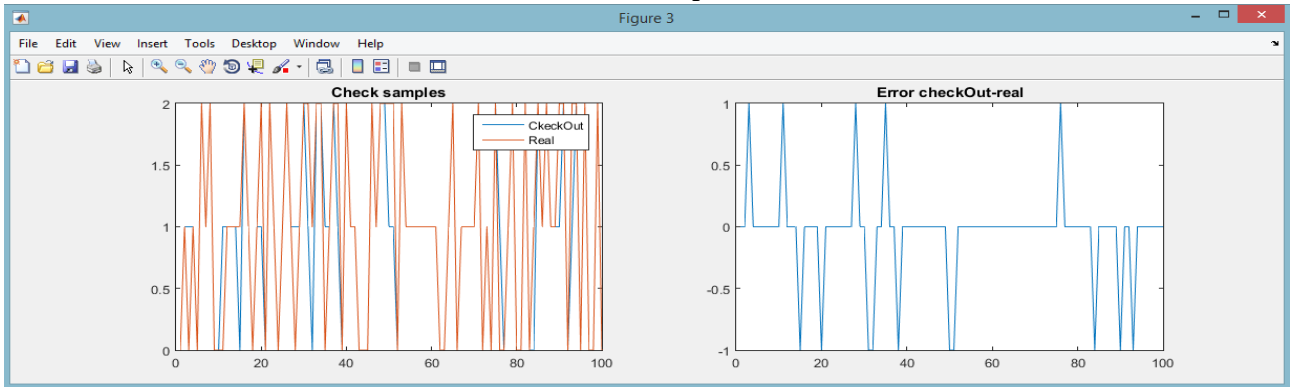
Training Sample



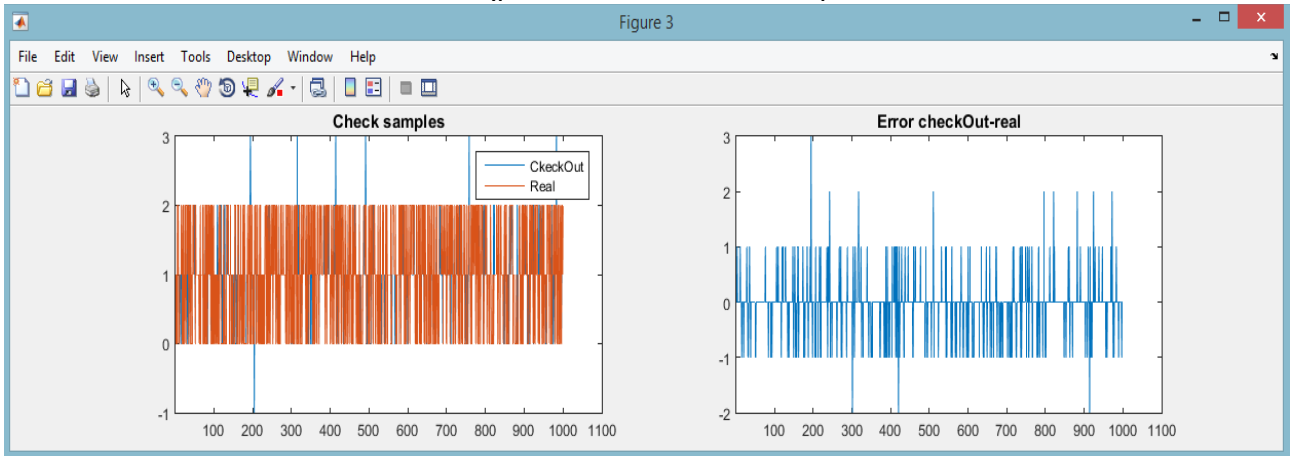
Validation Samples



Check samples



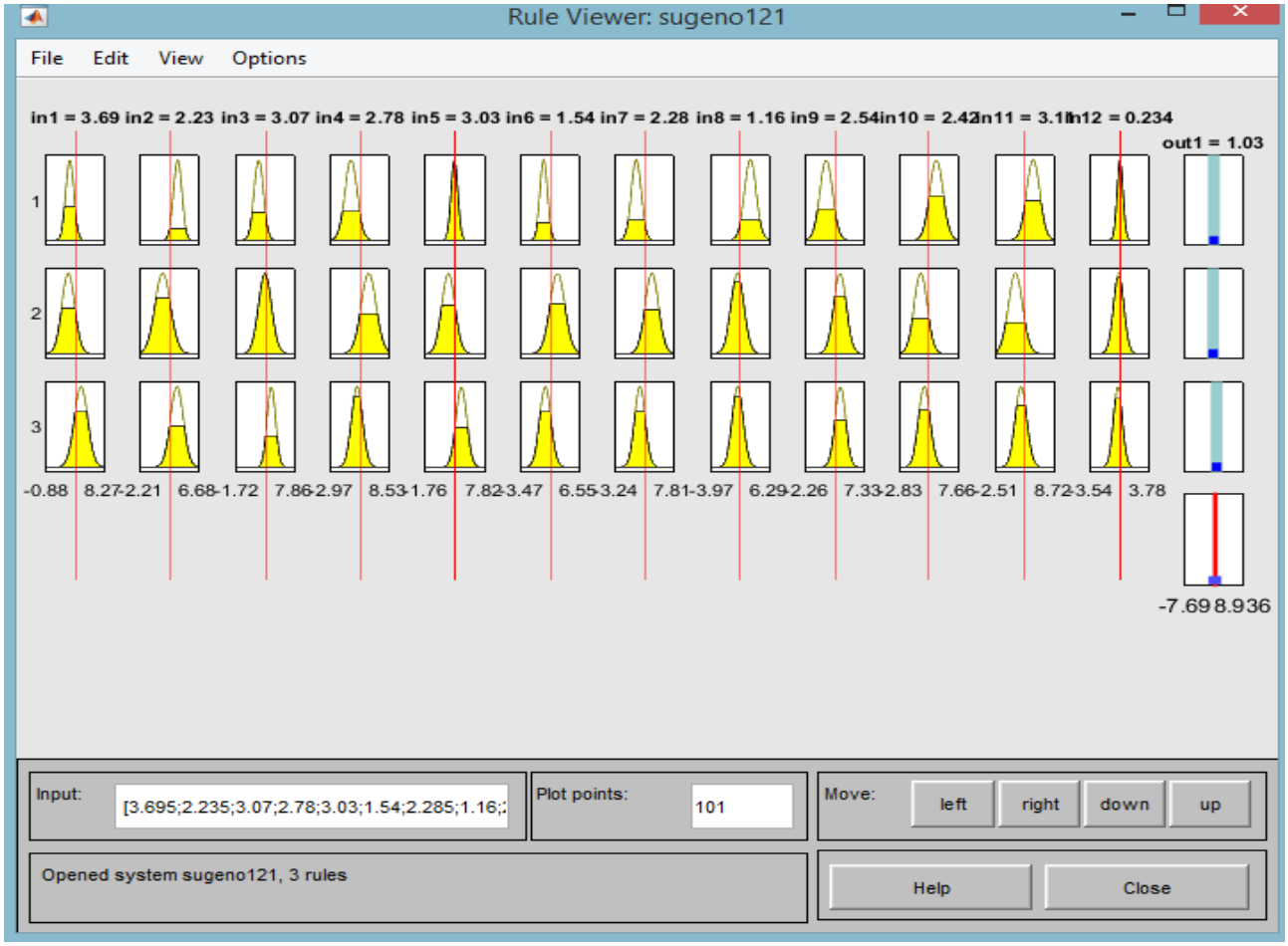
Ολόκληρο το σύνολο check δεδομένων



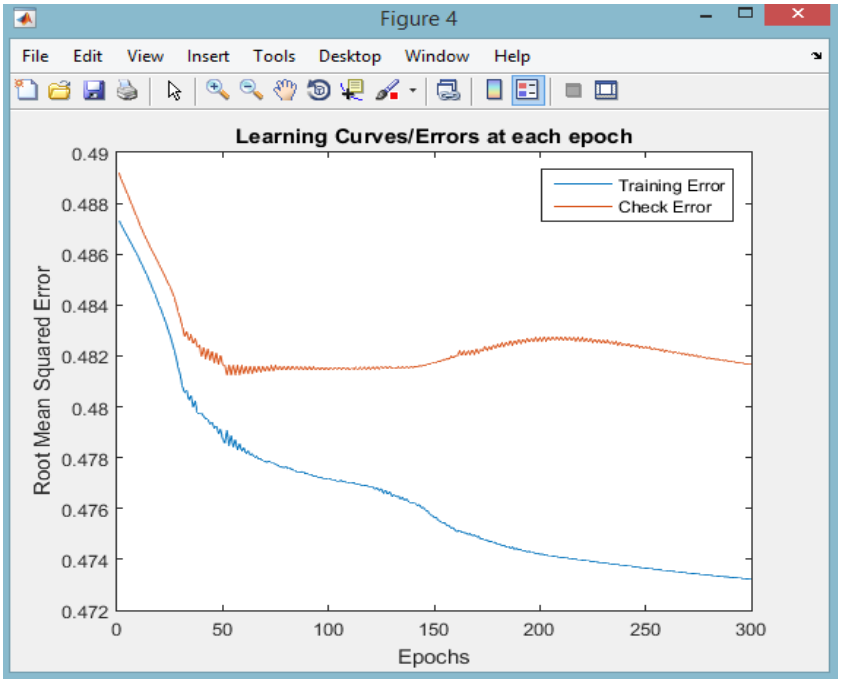
Τα αποτελέσματα δεν είναι τόσο καλά όσο στο πρώτο κομμάτι της εργασίας,πάντως ο ταξινομητής δουλεύει στα δεδομένα και κάνει προβλέψεις,συνεχίζουμε για να δούμε και τα υπόλοιπα διαγράμματα.

Παρουσιάζονται οι 3 κανόνες που διαμορφώνονται για 12 εισόδους . Για την παρουσίαση των κανόνων του μοντέλου καλούμε την fuzzy(tuned_fis) και επιλέγουμε από το GUI view rules, έτσι παίρνουμε.

Κανόνες βάσης

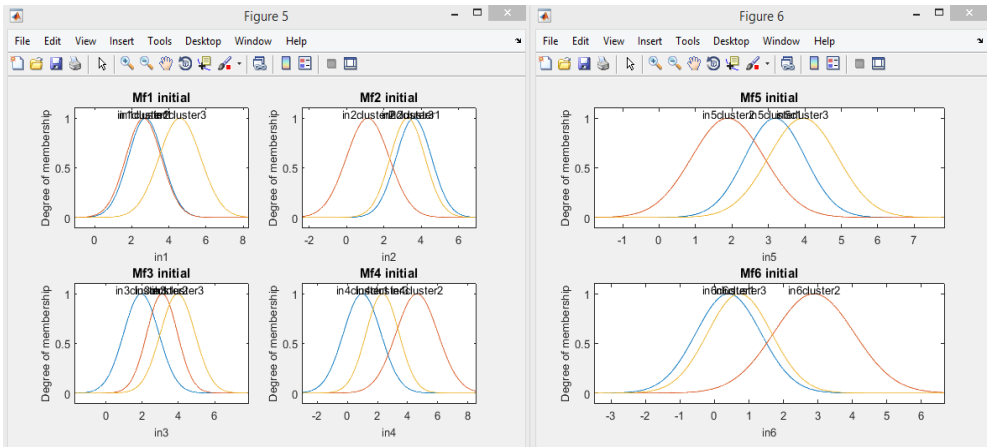


Διάγραμμα μάθησης (Learning Curves)

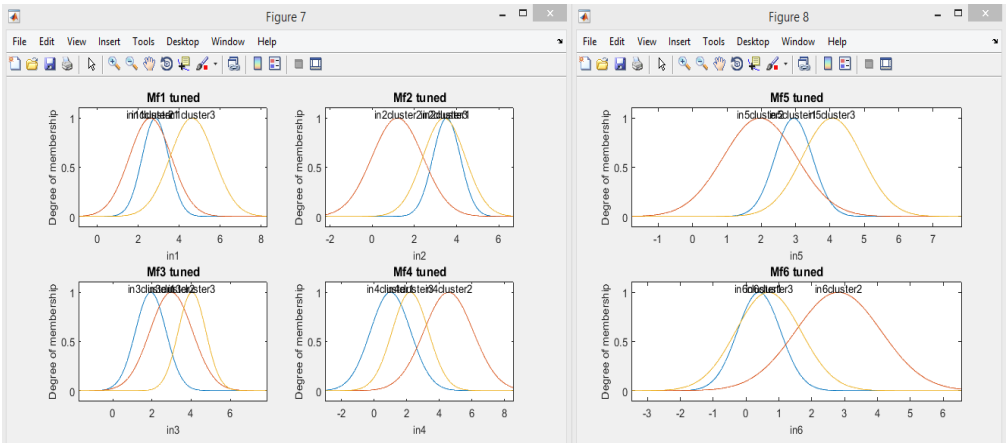


Τα ασαφή σύνολα για τις εισόδους 1,2,3,4,5,6 πριν και μετά την εκπαίδευση φαίνονται στα παρακάτω διαγράμματα.

Αρχικό μοντέλο

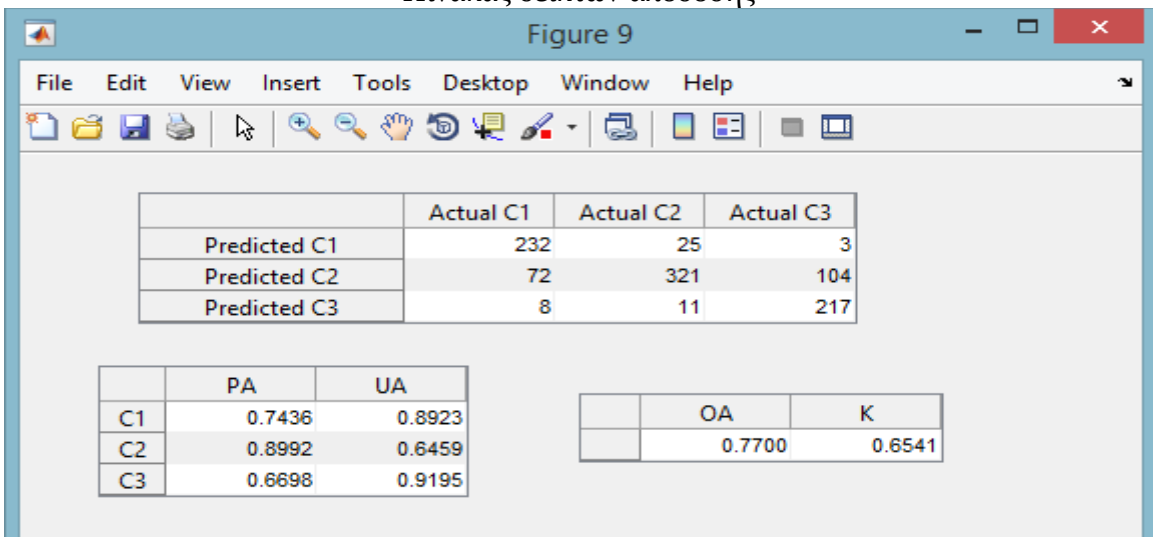


Εκπαιδευμένο μοντέλο



Τέλος παρουσιάζουμε τους δείκτες απόδοσης που αναφέρονται στην εκφώνηση. Με χρήση της συνάρτησης unitable του Matlab έχουμε τοποθετήσει τους δείκτες σε πίνακα για την καλύτερη παρουσίαση.

Πίνακας δεικτών απόδοσης



	Actual C1	Actual C2	Actual C3
Predicted C1	232	25	3
Predicted C2	72	321	104
Predicted C3	8	11	217

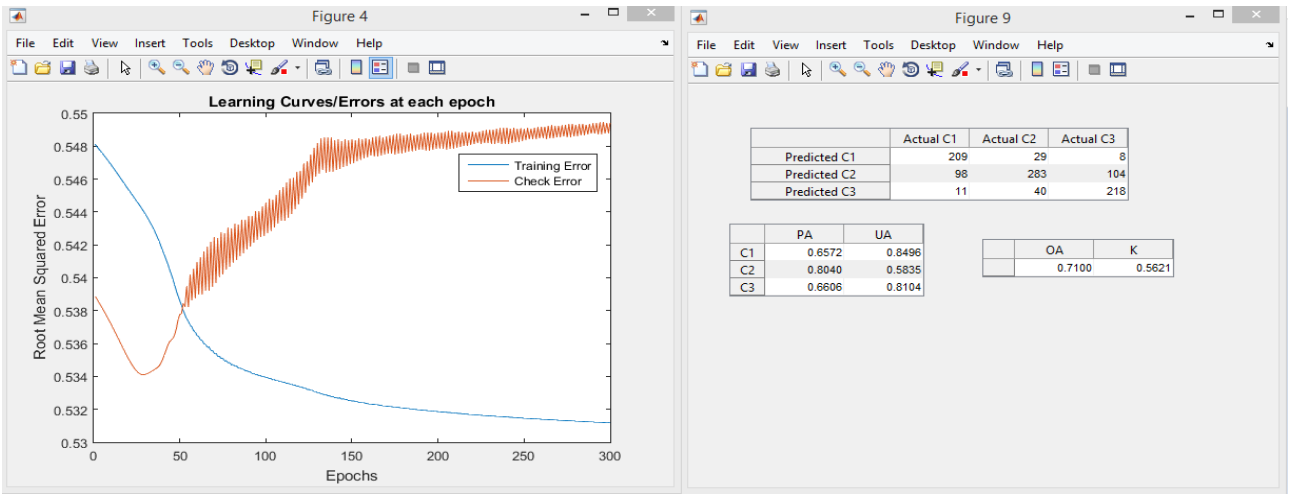
	PA	UA
C1	0.7436	0.8923
C2	0.8992	0.6459
C3	0.6698	0.9195

	OA	K
	0.7700	0.6541

Συμπεράσματα

Για τον βέλτιστο ταξινομητή έχουν επιλεγεί μόλις 3 κανόνες, κάτι που δεν είναι περίεργο αφού και οι διαφορετικές κατηγορίες πρόβλεψης είναι 3 {0,1,2}.Αυτό μπορεί να σημαίνει ότι ο κάθε κανόνας μπορεί να αντιστοιχεί σε κάθε μία από τις διαφορετικές κατηγορίες ή τουλάχιστον να συμμετέχει σε μεγάλο βαθμό για τον συμπερασμό της κατηγορίας. Για την κατανομή των συναρτήσεων συμμετοχής ισχύει ότι και στο πρώτο μέρος της εργασίας, βλέπουμε δηλαδή ότι περιορίζονται γύρο από συγκεκριμένες περιοχές για κάθε είσοδο και αυτό συνεπάγεται την συσχέτιση της εκάστοτε περιοχής με κάποια από τις κατηγορίες. Επίσης από τον πίνακα απόδοσης βλέπουμε ότι το μοντέλο έχει κάνει αρκετές λάθος προβλέψεις και κυρίως για την περίπτωση της 2ης κατηγορίας, φαίνεται να έχει πρόβλημα στο να προβλέψει την δεύτερη κατηγορία που αντιστοιχεί σε έξοδο 1. Για τις άλλες δύο κατηγορίες ο αριθμός σφαλμάτων είναι σχετικά μικρός. Επίσης από το διάγραμμα μάθησης παρατηρούμε ότι εμφανίζεται και πρόβλημα υπερεκπαίδευσης το οποίο όπως φαίνεται τελικά είναι αμελητέο. Γενικά το μοντέλο μπορεί να μην κάνει τις καλύτερες προβλέψεις, παρόλα αυτά είναι το βέλτιστο μοντέλο, για επιβεβαίωση θα εκτελέσουμε εκπαίδευση και για άλλους συνδυασμούς των ελεύθερων μεταβλητών και θα προβάλουμε το διάγραμμα μάθησης και το πίνακα απόδοσης, η εκπαίδευση θα γίνει για 300 εποχές.

NF = 6, NR = 3

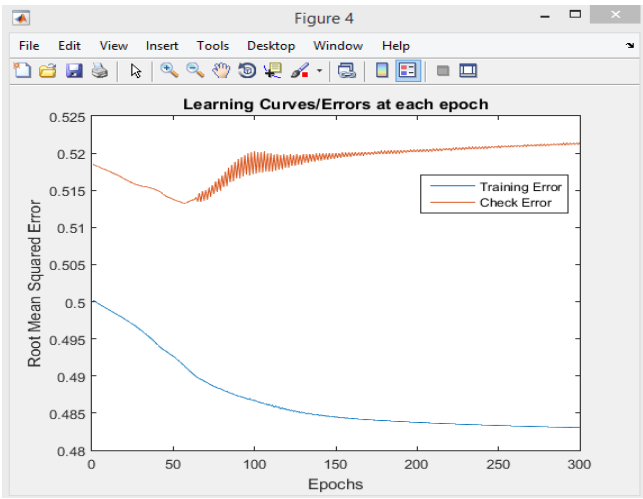


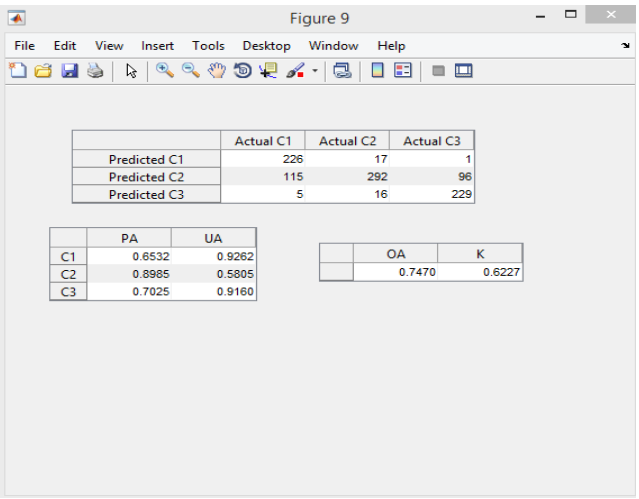
	Actual C1	Actual C2	Actual C3
Predicted C1	209	29	8
Predicted C2	98	283	104
Predicted C3	11	40	218

	PA	UA
C1	0.6572	0.8496
C2	0.8040	0.5835
C3	0.6606	0.8104

	OA	K
	0.7100	0.5621

NF = 10, NR = 3



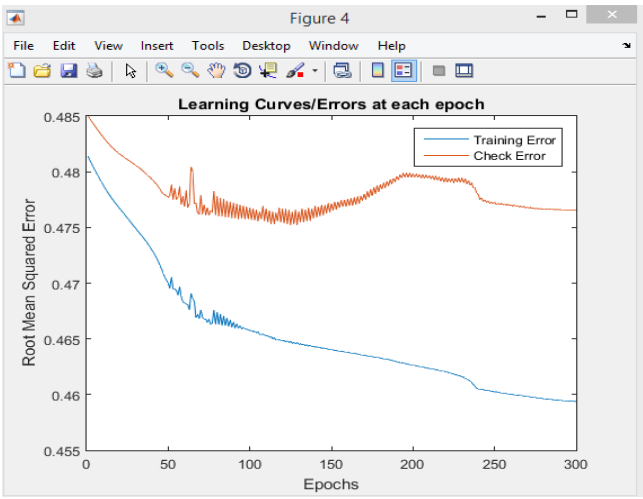


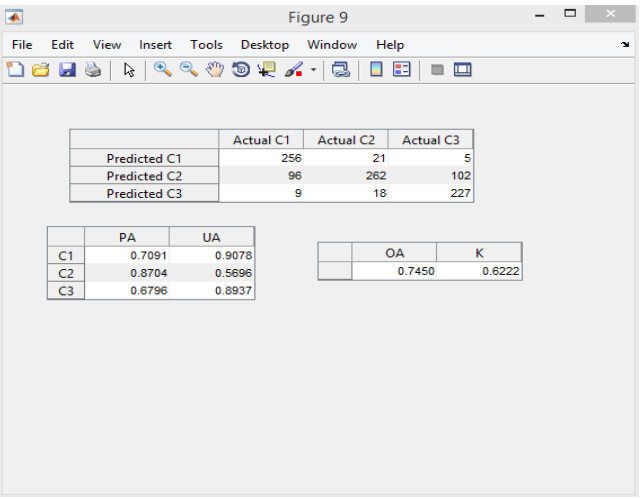
	Actual C1	Actual C2	Actual C3
Predicted C1	226	17	1
Predicted C2	115	292	96
Predicted C3	5	16	229

	PA	UA
C1	0.6532	0.9262
C2	0.8985	0.5805
C3	0.7025	0.9160

OA	K
0.7470	0.6227

NF = 16, NR = 3



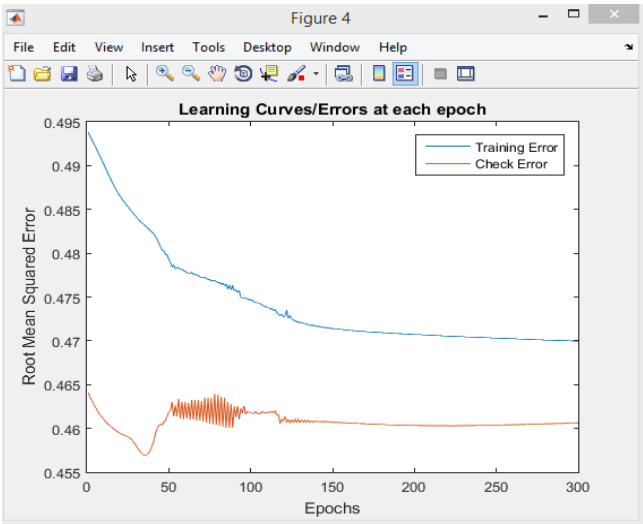


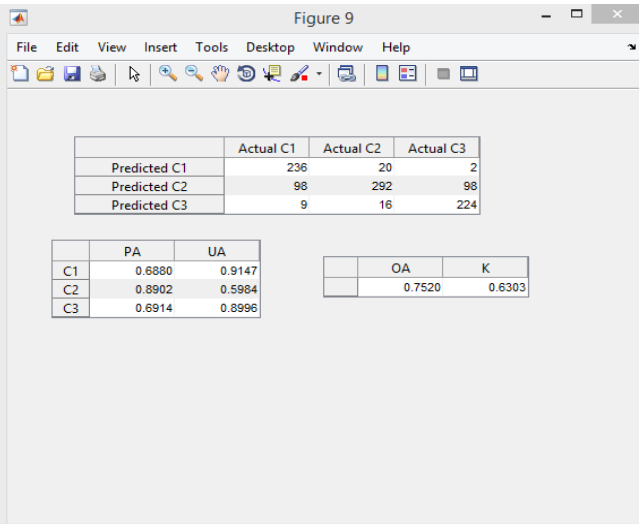
	Actual C1	Actual C2	Actual C3
Predicted C1	256	21	5
Predicted C2	96	262	102
Predicted C3	9	18	227

	PA	UA
C1	0.7091	0.9078
C2	0.8704	0.5696
C3	0.6796	0.8937

OA	K
0.7450	0.6222

NF = 21, NR = 3



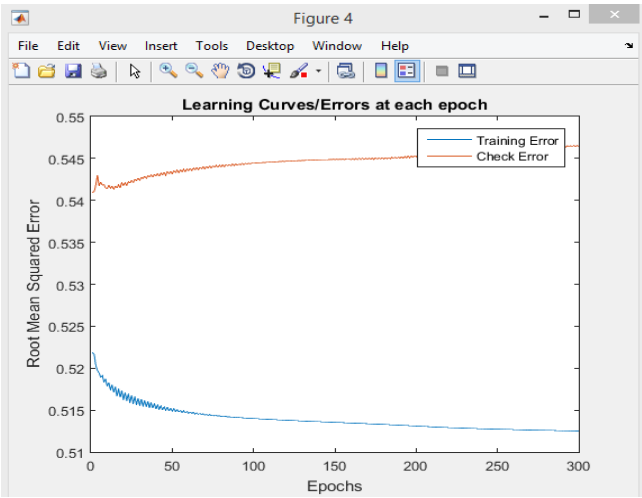


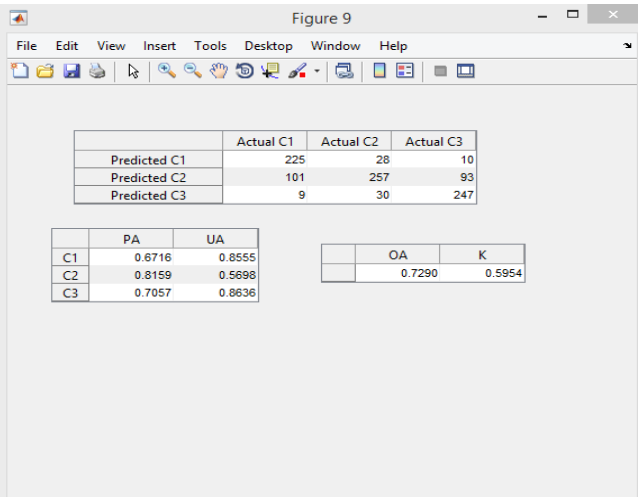
	Actual C1	Actual C2	Actual C3
Predicted C1	236	20	2
Predicted C2	98	292	98
Predicted C3	9	16	224

	PA	UA
C1	0.6880	0.9147
C2	0.8902	0.5984
C3	0.6914	0.8996

OA	K
0.7520	0.6303

NF = 6, NR = 10



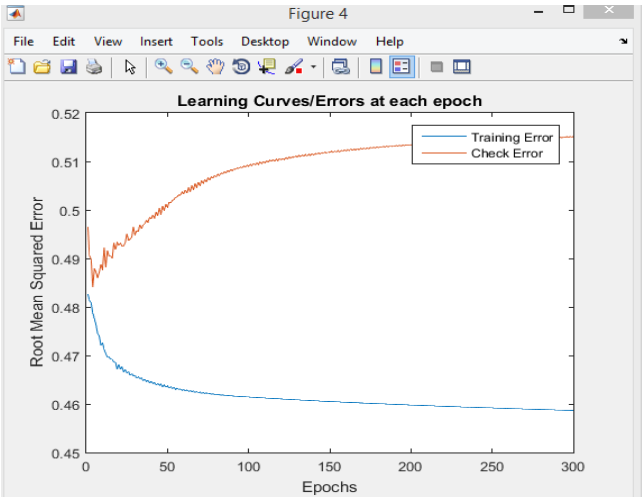


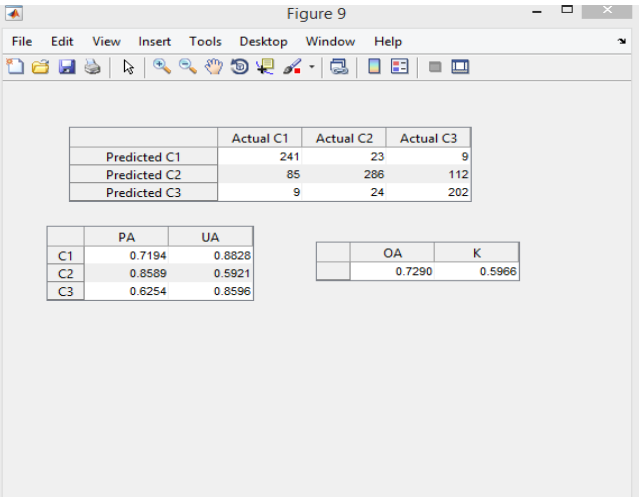
	Actual C1	Actual C2	Actual C3
Predicted C1	225	28	10
Predicted C2	101	257	93
Predicted C3	9	30	247

	PA	UA
C1	0.6716	0.8555
C2	0.8159	0.5698
C3	0.7057	0.8636

OA	K
0.7290	0.5954

NF = 10, NR = 10



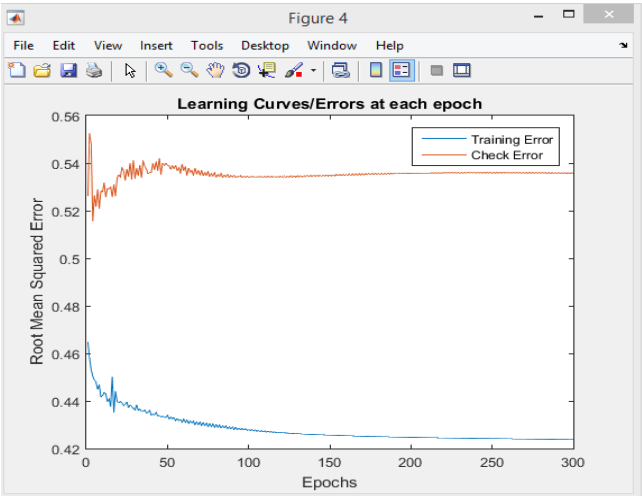


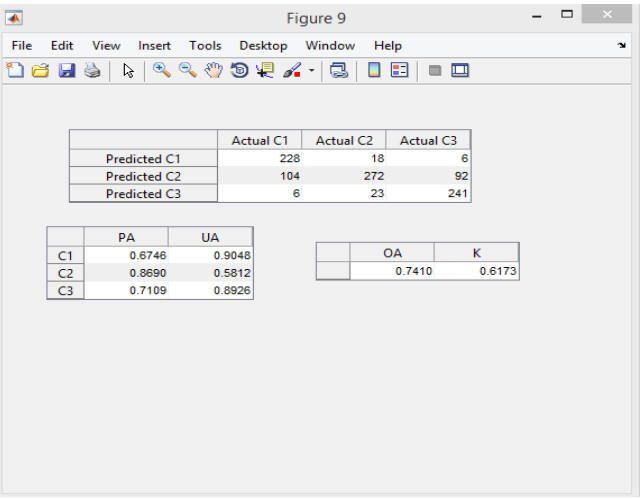
	Actual C1	Actual C2	Actual C3
Predicted C1	241	23	9
Predicted C2	85	286	112
Predicted C3	9	24	202

	PA	UA
C1	0.7194	0.8828
C2	0.8589	0.5921
C3	0.6254	0.8596

OA	K
0.7290	0.5966

NF = 15, NR = 10



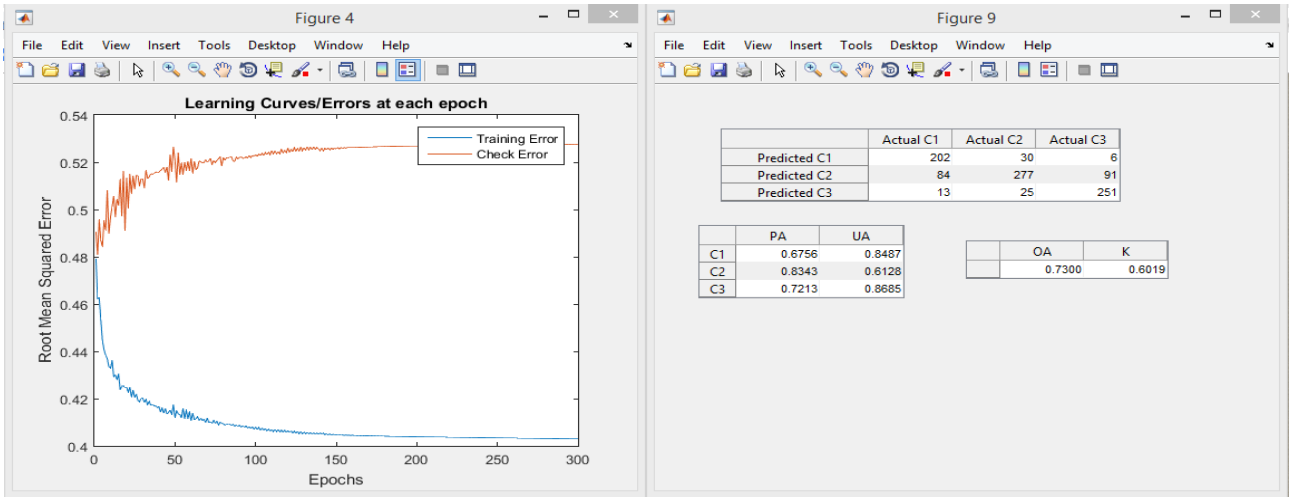


	Actual C1	Actual C2	Actual C3
Predicted C1	228	18	6
Predicted C2	104	272	92
Predicted C3	6	23	241

	PA	UA
C1	0.6746	0.9048
C2	0.8690	0.5812
C3	0.7109	0.8926

OA	K
0.7410	0.6173

NF = 15, NR = 15



Από τα διαγράμματα που παρατίθενται φαίνεται ότι το σφάλμα για διάφορους συνδυασμούς τιμών των ελεύθερων μεταβλητών είναι μεγαλύτερα από το σφάλμα που προκύπτει για τον συνδυασμό των βέλτιστων τιμών που προέκυψαν από την διαδικασία του Grid Search, κάτι που επιβεβαιώνει ότι η διαδικασία του Grid Search ήταν επιτυχής. Μια ακόμα παρατήρηση από τα παραπάνω διαγράμματα είναι το πρόβλημα της υπερεκπαίδευσης που φαίνεται ξεκάθαρα στα διαγράμματα όσο αυξάνεται ο αριθμός των κανόνων.
Τέλος,ο παραπάνω ταξινομητής χρησιμοποιεί μόλις τρις κανόνες για την εξαγωγή των συμπερασμάτων, αν είχαμε χρησιμοποιήσει grid partitioning με 2 ή 3 ασαφή σύνολα για κάθε είσοδο για τον ίδιο αριθμό χαρακτηριστικών που στην συγκεκριμένη περίπτωση είναι 12 θα είχαμε 2^{12} κανόνες για 2 ασαφή σύνολα και 3^{12} για 3 ασαφή σύνολα. Επίσης να δεν είχαμε κάνει επιλογή χαρακτηριστικών και χρησιμοποιούσαμε όλα τα χαρακτηριστικά πού είναι 40,τότε θα είχαμε 2^{40} και 3^{40} κανόνες αντίστοιχα. Προφανώς τέτοια νούμερα είναι απαγορευτικά μεγάλα για μία πρακτική εφαρμογή ταξινόμησης.

Περιγραφή αρχείων εργασίας

- **TSK_hybr_models.m :** Το αρχείο περιέχει το κώδικα για το πρώτο κομμάτι της εργασίας για τα τα 4 μοντέλα TSK .
- **gridSearch.m :** Εκτελεί τον κώδικα που αντιστοιχεί για την grid search 5-fold cross validation διαδικασία.
- **reliefCall.m :** Το script καλείται για να εκτελέσει την συνάρτηση relief ώστε γίνει η επιλογή των κατάλληλων χαρακτηριστικών/predictors. Κατά την εκτέλεση του script αποθηκεύετε η σειρά των καλύτερων χαρακτηριστικών αφού αυτή δεν αλλάζει,έτσι ώστε να μην εκτελείται συνέχεια ο αλγόριθμος σε επαναλαμβανόμενες δοκιμές στα πλαίσια της εργασίας.
- **plotGridSearch.m :** Καλείται για να δημιουργήσει την γραφική παράσταση της επιφάνειας του πλέγματος.
- **bestTSKmodel :** Καλείται για την εκπαίδευση του καλύτερου μοντέλου. Αφού εισάγουμε τις παραμέτρους για το καλύτερο μοντέλο στην συνέχεια τρέχουμε το script για να εκπαιδευτεί το μοντέλο.
- **Αρχεία δεδομένων :** Τα διάφορα αρχεία δεδομένων που περιέχονται μέσα στο φάκελο περιέχουν τα αποτελέσματα από τις προσομοιώσεις,όπως για παράδειγμα τα αποτελέσματα από το grid search, μπορούμε να χρησιμοποιήσου απευθείας το script plotGridSearch.m για το σχεδιασμό και την επίδειξη του πλέγματος.