ΕΡΓΑΣΙΑ ΣΤΟ ΜΑΘΗΜΑ
ΑΣΑΦΗ ΣΥΣΤΗΜΑΤΑ

Επίλυση προβλήματος ταξινόμησης με χρήση μοντέλων TSK

Στόχος της εργασίας αυτής είναι να διερευνηθεί η ικανότητα των TSK ασαφών μοντέλων στην επίλυση προβλημάτων ταξινόμησης (classification). Συγκεκριμένα, επιλέγονται δύο σύνολα δεδομένων με σκοπό την ταξινόμηση, από τα διαθέσιμα δεδομένα, δειγμάτων στις εκάστοτε κλάσεις τους, με χρήση ασαφών νευρωνικών μοντέλων. Η εργασία αποτελείται από δύο μέρη, το πρώτο από τα οποία προορίζεται για μια απλή διερεύνηση της διαδικασίας εκπαίδευσης και αξιολόγησης των TSK μοντέλων, ενώ το δεύτερο περιλαμβάνει μια πιο συστηματική προσέγγιση στο πρόβλημα της εκμάθησης από δεδομένα, σε συνδυασμό με προεπεξεργαστικά βήματα όπως επιλογή χαρακτηρηστικών (feature selection) και μεθόδους βελτιστοποίησης των μοντέλων μέσω της διασταυρωμένης επικύρωσης (cross validation).

Λίγα λόγια για την ταξινόνηση δεδομένων: Το πρόβλημα της ταξινόμησης δεδομένων αποτελεί ένα υποσύνολο του γενικότερου ζητήματος της αναγνώρισης προτύπων, και ασχολείται με την κατηγοριοποίηση δεδομένων σε δύο ή περισσότερες κλάσεις. Συγκεκριμένα, ας υποθέσουμε ότι έχουμε ένα σύνολο δειγμάτων $X=\{\vec{x}_1,\vec{x}_2,...,\vec{x}_N\}$, όπου N είναι το πλήθος των δειγμάτων και $\vec{x}_i\in R^M$, με M το πλήθος των χαρακτηριστικών (attributes), ένα σύνολο κλάσεων (labels) $Y=\{y_1,y_2,...,y_k\}$, και ότι κάθε στοιχείο του συνόλου X ανήκει σε κάποια από τις $k\in R$ κλάσεις. Στόχος μας είναι, με χρήση των διαθέσιμων δεδομένων, να κατασκευάσουμε μια συνάρτηση $f(\vec{x})$ η οποία να αναθέτει σε κάθε στοιχείο $\vec{x}_i\in R^M$ μια ετικέτα (label) η οποία να ισοδυναμεί με την αντίστοιχη κλάση στην οποία και ανήκει. Τελικός μας σκοπός, είναι η χρήση της συνάρτησης αυτής σε ένα νέο σύνολο δειγμάτων, $\hat{X}$ για τη σωστή ταξινόμηση των στοιχείων του στις αντίστοιχες κλάσεις τους.

1) Μια πρώτη εφαρμογή σε UCI dataset: Στην πρώτη φάση της εργασίας, επιλέγεται από το UCI repository το Wifi-localization dataset, το οποίο περιλαμβάνει 2000 δείγματα (instances), από 7 χαρακτηριστικά (attributes) το καθένα. Τα χαρακτηριστικά του συνόλου αποτελούνται από μετρήσεις της έντασης του σήματος wifi , όπως αυτές έγιναν από διαφορετικά smartphones. Στόχος είναι να χρησιμοποιηθούν τα δεδομένα αυτά για την κατασκευή ενός μοντέλου το οποίο να ταυτοποιεί τη θέση (ένα από τέσσερα δωμάτια) στα οποία γίνονται αυτές οι μετρήσεις. Για την επίλυση του προβλήματος θα ακολουθήσουμε τα παρακάτω βήματα:

- *Διαχωρισμός σε σύνολα εκπαίδευσης-επικύρωσης-ελέγχου:* Σε πρώτη φάση είναι απαραίτητος ο διαχωρισμός του συνόλου δεδομένων σε τρία μη επικαλυπτόμενα υποσύνολα, $\{D_{trn},D_{val},D_{chk}\}$ από τα οποία το πρώτο θα χρησιμοποιηθεί για εκπαίδευση, το δεύτερο γα επικύρωση και αποφυγή του φαινομένου υπερκπαίδευσης, και το τελευταίο για τον έλεγχο της απόδοσης του τελικού μοντέλου. Προτείνεται να χρησιμοποιηθεί το 60% του συνόλου των δειγμάτων για το υποσύνολο εκπαίδευσης και από 20% του συνόλου των δειγμάτων για κάθε ένα από τα δύο εναπομείναντα υποσύνολα. Ένα σημείο στο οποίο θα πρέπει να δοθεί προσοχή είναι το ότι για να επιτύχουμε καλή απόδοση, θα πρέπει 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:* Σε αυτό το στάδιο θα εξεταστούν διάφορα μοντέλα TSK όσον αφορά την απόδοσή τους στο σύνολο ελέγχου. Συγκεκριμένα, θα εκπαιδευτούν τέσσερα TSK μοντέλα, στα οποία θα μεταβάλλεται το πλήθος των ασαφών IF-THEN κανόνων. Σκοπός είναι να μελετηθεί η επίδραση της διαμέρισης του χώρου εισόδου – σε συνάρτηση με την πολυπλοκότητα που αυτή επιφέρει, στην απόδοση του ταξινομητή. 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. Εφόσον η έξοδός μας αποτελείται από έναν αριθμό, ενδεικτικό της κλάσης στην οποία ανήκει το εκάστοτε δείγμα, προτείνεται η χειροκίνητη αλλαγή του τύπου συνάρτησης εξόδου από linear σε constant. Ο αριθμός των κανόνων να λαμβάνει τιμές από το σύνολο $\{4,8,12,16\}$. Και τα τέσσερα μοντέλα να εκπαιδευτούν με την υβριδική μέθοδο, σύμφωνα με την οποία οι παράμετροι των συναρτήσεων συμμετοχής βελτιστοποιούνται μέσω της μεθόδου της οπισθοδιάδοσης (backpropagation algorithm), ενώ οι παράμετροι της συνάρτησης εξόδου βελτιστοποιούνται μέσω της μεθόδου των ελαχίστων τετραγώνων (Least Squares).
- *Αξιολόγηση μοντέλων:* Για την αξιολόγηση της ταξινόμησης των δειγμάτων από τα διάφορα μοντέλα, θα χρησιμοποιηθούν οι εξής δείκτες απόδοσης:
  1. *Error matrix:* Ο πίνακας σφαλμάτων ταξινόμησης είναι ένας $k \times k$ πίνακας, με k τον αριθμό των κλάσεων ο οποίος βοηθά στην οπτικοποίηση της απόδοσης ενός ταξινομητή και μέσω του οποίου αποκτούμε πρόσβαση σε μια σειρά δεικτών απόδοσης. Η γενική του δομή έχει ως εξής:

Error Matrix	***Actual: C_1***	***Actual: C_2***	***...***	***Actual: C_k***
Predicted: C_1	x_{11}	x_{12}	…	x_{1k}
Predicted: C_2	x_{21}	x_{22}	…	x_{2k}
...	…	…	…	…
Predicted: C_k	x_{k1}	x_{k2}	…	x_{kk}

     Τα στοιχεία της κύριας διαγωνίου περιλαμβάνουν το πλήθος των δειγμάτων που ανήκουν σε μια συγκεκριμένη κλάση και τα οποία ορθώς ταξινομήθηκαν σε αυτή από το μοντέλο μας, ενώ τα στοιχεία εκτός της διαγωνίου περιλαμβάνουν το πλήθος των δειγμάτων τα οποία λανθασμένα ταξινομήθηκαν σε διαφορετική κλάση από αυτή στην οποία στην πραγματικότητα ανήκουν.
  2. *Overall accuracy:* Η συνολική ακρίβεια ενός ταξινομητή ορίζεται ως το ποσοστό των ορθώς ταξινομημένων δειγμάτων ως προς το συνολικό πλήθος των δειγμάτων. Χρησιμοποιώντας τα στοιχεία του πίνακα σφαλμάτων, η ακρίβεια υπολογίζεται ως: $OA = \frac{1}{N} \sum_{i=1}^{k} x_{ii}$.
  3. *Producer’s accuracy – User’s accuracy:* Δύο δείκτες που παρουσιάζουν ενδιαφέρον και που αναφέρονται στην απόδοση του ταξινομητή όσον αφορά κάθε κλάση ξεχωριστά, είναι η *ακρίβεια παραγωγού* και η *ακρίβεια χρήστη.* Ορίζουμε αρχικά $x_{ir} = \sum_{j=1}^{k} x_{ij}$ το πλήθος των σημείων που ταξινομήθηκαν

στην C_i κλάση και $x_{jc}=\sum_{i=1}^{M} x_{ij}$ το πλήθος των σημείων τα οποία ανήκουν στην κλάση C_j . Με βάση τα παραπάνω, η ακρίβεια παραγωγού δίνεται από τον τύπο $PA(j)=\frac{x_{jj}}{x_{jc}}$ και η ακρίβεια χρήστη θα είναι $UA(i)=\frac{x_{ii}}{x_{ir}}$.

4. $\hat{K}$: Ένα άλλο στατιστικό μέγεθος που μπορεί να εξαχθεί από έναν πίνακα σφαλμάτων είναι το μέγεθος $\hat{k}$, το οποίο αποτελεί εκτίμηση της πραγματικής στατιστικής παραμέτρου κ . Υπολογίζεται σύμφωνα με τον τύπο $\hat{k}=\frac{N\sum_{i=1}^{M} x_{ii}-\sum_{i=1}^{M} x_{ir}x_{ic}}{N^2-\sum_{i=1}^{M} x_{ir}x_{ic}}$.

- *Ζητούμενα του προβλήματος:* Για κάθε ένα από τα πέντε TSK μοντέλα, να γίνουν οι κατάλληλες αρχικοποιήσεις και στη συνέχεια να εκτελεστεί η εκπαίδευσή τους με τις παραμέτρους που περιγράφηκαν παραπάνω. Ζητούνται τα εξής:
  1. Να δώσετε τα αντίστοιχα διαγράμματα στα οποία να απεικονίζονται οι τελικές μορφές των ασαφών συνόλων που προέκυψαν μεσω της διαδικασίας εκπαίδευσης.
  2. Να δοθούν τα διαγράμματα μάθησης (learning curves) όπου να απεικονίζεται το σφάλμα του μοντέλου συναρτήσει του αριθμού των επαναλήψεων (iterations).
  3. Να δοθεί ο πίνακας σφαλμάτων ταξινόμησης και να εξαχθούν από αυτόν τιμές των δεικτών απόδοσης $OA, PA, UA, \hat{K}$.
  4. Να σχολιάσετε τα αποτελέσματα. Ποιά είναι η επίδραση του αριθμού των κανόνων στην απόδοση του ταξινομητή; Ποιά συμπεράσματα μπορούμε να εξάγουμε σχετικά με την επικάλυψη των προβολών των ασαφών συνόλων κάθε cluster στις αντίστοιχες εισόδους όσον αφορά την ενεργοποίηση των κανόνων και γενικότερα την απόδοση του ταξινομητή; Να συνοδεύσετε τα σχόλια με διαγράμματα της επιλογής σας. Μπορείτε να προτείνετε κάποια μέθοδο για τη βελτίωση της σχεδίασης του τμήματος υπόθεσης;

- *Σημείωση για την έξοδο των μοντέλων:* Ένα σημείο το οποίο μπορεί να αποτελέσει πηγή σύγχυσης, είναι το γεγονός ότι η υλοποίηση των TSK ασαφών μοντέλων στο MATLAB είναι τέτοια ώστε η έξοδός τους να είναι πραγματική, κάτι το οποίο οδηγεί σε δυσκολίες σε προβλήματα ταξινόμησης, όπου η μεταβλητή – στόχος είναι συνήθως κατηγορική. Στα προβλήματα ταξινόμησης που συναντούμε σε αυτή την εργασία, η μεταβλητή – στόχος είναι ακέραιος,λαμβάνοντας τιμές σε κάποιο σύνολο ακεραίων $\{k_0, k_1, \ldots, k_m\}$. Ένας απλός τρόπος να φέρουμε την έξοδο του μοντέλου στην ίδια μορφή είναι να στρογγυλοποιήσουμε κάθε στοιχείο στον πλησιέστερο ακέραιο. Εναλλακτικά, μπορείτε να ορίσετε ένα δικό σας σχήμα διακριτοποίησης της συνεχούς εξόδου του μοντέλου, αν κρίνετε ότι κάτι τέτοιο οδηγεί σε αποδοτικότερη ταξινόμηση.

2) Εφαρμογή σε dataset με υψηλή διαστασιμότητα: Στη δεύτερη φάση της εργασίας θα ακολουθηθεί μια πιο συστηματική προσέγγιση στο πρόβλημα της χρήσης ασαφών νευρωνικών μοντέλων σε προβλήματα ταξινόμησης. Για το σκοπό αυτό θα επιλεχθεί ένα dataset με υψηλότερο βαθμό διαστασιμότητας. Ένα προφανές πρόβλημα που ανακύπτει

από την επιλογή αυτή, είναι η λεγόμενη “έκρηξη” του πλήθους των IF-THEN κανόνων (rule explosion). Όπως είναι γνωστό από τη θεωρία, για την κλασική περίπτωση του grid partitioning του χώρου εισόδου, ο αριθμός των κανόνων αυξάνεται εκθετικά σε σχέση με το πλήθος των εισόδων, γεγονός που καθιστά πολύ δύσκολη την μοντελοποίηση μέσω ενός TSK μοντέλου ακόμα και για datasets μεσαίας κλίμακας.

Σύμφωνα με τα παραπάνω, μια αρκετά προφανής προσέγγιση θα ήταν να προσπαθήσουμε να μειώσουμε ταυτόχρονα *και* το πλήθος των εισόδων *και* το πλήθος των κανόνων. Οι δύο τεχνικές που παρουσιάζονται παρακάτων στοχεύουν σε αυτόν ακριβώς το διττό σκοπό.

1. *Επιλογή χαρακτηριστικών (feature selection):* Η επιλογή χαρακτηριστικών αποτελεί μια από τις πιο διαδεδομένες και πλούσιες οικογένειες τεχνικών για τη μείωση της διαστασιμότητας σε ένα πρόβλημα μοντελοποίησης. Η βασική ιδέα πίσω από αυτές τις τεχνικές είναι ότι στη συντριπτική πλειοψηφία των περιπτώσεων, ένας σημαντικός αριθμός χαρακτηριστικών/εισόδων των δειγμάτων ενός συνόλου δεδομένων είναι πλεονάζοντα, επομένως θα μπορούσαν να παραλειφθούν από τη διαδικασία μοντελοποίησης. Αυτό μας επιτρέπει να μειώσουμε από τη μία τη διαστασιμότητα του προβλήματος, καθιστώντας το ευκολότερα διαχειρίσιμο, και από την άλλη, να απλοποιήσουμε το τελικό μοντέλο, έτσι ώστε να είναι ευκολότερα ερμηνεύσιμο από τους εκάστοτε ερευνητές/χρήστες. Για τους σκοπούς της συγκεκριμένης εργασίας, ως μέθοδος επιλογής χαρακτηριστικών επιλέγεται ο αλγόριθμος Relief. Μια αναλυτική παρουσίαση του αλγορίθμου αυτού περιέχεται στο paper το οποίο είναι αναρτημένο στο Υλικό Μαθήματος.
2. *Διαμέριση διασκορπισμού (scatter partitioning):* Κατά τη διαδικασία αρχικοποίησης ενός TSK μοντέλου, ένα από τα βασικά βήματα είναι ο διαχωρισμός του χώρου εισόδου και η δημιουργία των αρχικών ασαφών συνόλων. Η απλούστερη μέθοδος προς το σκοπό αυτό είναι το grid partitioning, όπως αναφέρθηκε όμως παραπάνω, η συγκεκριμένη μέθοδος οδηγεί σε εκθετικά αυξανόμενο πλήθος κανόνων σε συνάρτηση με το πλήθος των εισόδων. Μια εναλλακτική επιλογή είναι η χρήση μεθόδων ομαδοποίησης για το διαχωρισμό του χώρου εισόδου και η αρχικοποίηση των ασαφών συνόλων πάνω στις ομάδες που προέκυψαν. Με τον τρόπο αυτό, το πλήθος των κανόνων παύει να εξαρτάται από το πλήθος των εισόδων και εξαρτάται πλέον αποκλειστικά από τον αριθμό των clusters. Δύο δημοφιλείς επιλογές για την ομαδοποίηση των δεδομένων και τη μετέπειτα δημιουργία των κανόνων είναι η αφαιρετική ομαδοποίηση (subtractive clustering) και ο αλγόριθμος Fuzzy C-means (FCM).

Το dataset που θα επιλεγεί για την επίδειξη των παραπάνω μεθόδων είναι το Waveform Generation Dataset από το UCI repository. Το συγκεκριμένο dataset, περιλαμβάνει 5000 δείγματα, καθένα από τα οποία περιγράφεται από 40 μεταβλητές/χαρακτηριστικά. Τα δεδομένα αποτελούνται από τρία βασικά σήματα, τα οποία συνδυαζόμενα ανά δύο δίνουν τρεις κυματομορφές, οι οποίες αποτελούν και τις κλάσεις του προβλήματος. Στόχος είναι η ταυτοποίηση της κυματομορφής. Σημειώνεται ότι τα μερικά από τα χαρακτηριστικά αποτελούνται αποκλειστικά από θόρυβο. Είναι φανερό ότι το μέγεθος του dataset καθιστά απαγορευτική μια απλή εφαρμογή ενός TSK μοντέλου. Για να αντιμετωπίσουμε το πρόβλημα που εισάγεται από τα θορυβώδη χαρακτηριστικά, αλλά και από το πλήθος των χαρακτηριστικών γενικότερα θα καταφύγουμε σε μεθόδους επιλογής χαρακτηριστικών και διαμέρισης διασκορπισμού αντίστοιχα. Οι δύο αυτές μέθοδοι όμως, παρά τη ελάττωση της πολυπλοκότητας που επιφέρουν, εισάγουν στο πρόβλημα δύο ελεύθερες παραμέτρους, συγκεκριμένα, τον αριθμό των χαρακτηριστικών προς επιλογή και τον αριθμό των ομάδων που θα δημιουργηθούν. Η επιλογή των δύο αυτών παραμέτρων επαφίεται στον εκάστοτε χρήστη και είναι ουσιαστική όσον αφορά την τελική απόδοση του μοντέλου. Στην παρούσα εργασία, θα υλοποιηθεί η μέθοδος αναζήτησης πλέγματος για την εύρεση των βέλτιστων

τιμών των παραμέτρων. Αναλυτικά, η μοντελοποίηση του προβλήματος θα ακολουθήσει λοιπόν τα εξής βήματα:

1. *Διαχωρισμός σε σύνολα εκπαίδευσης – επικύρωσης – ελέγχου:* Όπως και στο πρώτο κομμάτι της εργασίας, είναι απαραίτητη η διαμέριση του συνόλου δεδομένων σε τρία υποσύνολα, $\{D_{trn}, D_{val}, D_{chk}\}$. Όπως και στο πρώτο ερώτημα, η κατανομή των δειγμάτων που ανήκουν στις εκάστοτε κλάσεις πρέπει να ακολουθούν όσο το δυνατόν παρόμοια κατανομή και στα τρία υποσύνολα.
2. *Επιλογή των βέλτιστων παραμέτρων:* Όπως αναφέρθηκε παραπάνω, το σύστημά μας περιλαμβάνει δύο ελεύθερες παραμέτρους την τιμή των οποίων πρέπει να επιλέξουμε εμείς. Η δημοφιλέστερη μέθοδος μέσω της οποίας επιτυγχάνεται αυτό είναι η αναζήτηση πλέγματος. Συγκεκριμένα, αφού λάβουμε ένα σύνολο τιμών για κάθε παράμετρο, δημιουργούμε ένα n-διάστατο πλέγμα (στην περίπτωσή μας $n=2$), όπου κάθε σημείο αντιστοιχεί σε μια n-άδα τιμών για τις εν λόγω παραμέτρους, και σε κάθε σημείο χρησιμοποιούμε μια μέθοδο αξιολόγησης για ελέγξουμε την ορθότητα των συγκεκριμένων τιμών. Μια καθιερωμένη επιλογή για την αξιολόγηση αυτή αποτελεί η διασταυρωμένη επικύρωση (cross validation). Σύμφωνα με τη μέθοδο αυτή ,και για επιλεγμένες τιμές των παραμέτρων, χωρίζουμε το σύνολο εκπαίδευσης σε δύο υποσύνολα, από τα οποία το ένα θα χρησιμοποιηθεί για την εκπαίδευση ενός μοντέλου και το δεύτερο για την αξιολόγησή του. Η διαδικασία αυτή επαναλαμβάνεται – συνήθως πέντε ή δέκα φορές – όπου κάθε φορά χρησιμοποιείται διαφορετικός διαχωριμός του συνόλου εκπαίδευσης, και στο τέλος λαμβάνουμε τον μέσο όρο του σφάλματος του μοντέλου. Η λογική πίσω από τις πολλαπλές εκπαιδεύσεις και ελέγχους έγκειται στο ότι με αυτό τον τρόπο, αποκτούμε μια αρκετά καλή εκτίμηση της απόδοσης του μοντέλου, και έμεσσα των τιμών των παραμέτρων με βάση τις οποίες χτίστηκε το μοντέλο. Όταν η παραπάνω διαδικασία εκτελεστεί για κάθε σημείο του πλέγματος, λαμβάνουμε ως βέλτιστες τιμές των παραμέτρων, τις τιμές που αντιστοιχούν στο μοντέλο που παρουσίασε το ελάχιστο μέσο σφάλμα. Οι τιμές αυτές χρησιμοποιούνται για την εκπαίδευση του τελικού μας μοντέλου. Η όλη διαδικασία γίνεται ευκολότερα κατανοητή με το παρακάτω κομμάτι ψευδοκώδικα:

```
Define sets of model parameter values to evaluate
for each parameter set do
    for each resampling iteration do
        Hold–out specific samples
        [Optional] Pre–process the data
        Fit the model on the remainder
        Predict the hold–out samples
    end
    Calculate the average performance across hold–out predictions
end
Determine the optimal parameter set
Fit the final model to all the training data using the optimal parameter set
```

Για τους σκοπούς αυτής της εργασίας, ορίζουμε τις εξής παραμέτρους:

- Αριθμός χαρακτηριστικών: Οι τιμές των χαρακτηριστικών που θα χρησιμοποιηθούν δίνονται από το σύνολο $NF=\{4,8,12,16\}$.
- Αριθμός κανόνων: Ο αριθμός κανόνων (αριθμός των clusters), θα λαμβάνει τιμές από το σύνολο $NR=\{3,9,15,21,\}$.

3. Με βάση τις βέλτιστες τιμές των παραμέτρων που επιλέχθηκαν από το προηγούμενο βήμα, εκπαιδεύουμε ένα τελικό TSK μοντέλο και ελέγχουμε την απόδοσή του στο σύνολο ελέγχου.

Τα παραπάνω βήματα συνοψίζουν πλήρως τη διαδικασία μοντελοποίησης που θα ακολουθηθεί. Σημειώνεται ότι τα σύνολα παραμέτρων έτσι όπως παρουσιάζονται παραπάνω είναι προαιρετικά, και μπορεί κανείς να αντικαταστήσει τις τιμές τους, ειδικά αν η διαδικασία της διασταυρωμένης επικύρωσης αποδειχθεί ιδιαίτερα χρονοβόρα. Ζητούνται τα εξής:

1. Ο διαχωρισμός του συνόλου δεδομένων να γίνει όπως και στο πρώτο κομμάτι, με τα σύνολα εκπαίδευσης-επικύρωσης-ελέγχου να περιλαμβάνουν αντίστοιχα το 60% - 20% - 20% του συνόλου.
2. Να εκτελεστεί αναζήτηση πλέγματος (grid search) και αξιολόγηση μέσω 5-πτυχης διασταυρωμένης επικύρωσης (5-fold cross validation) για την επιλογή των βέλτιστων τιμών των παραμέτρων. Σε κάθε επανάληψη να αποθηκεύεται το μέσο σφάλμα. Ο διαχωρισμός των δεδομένων να γίνει έτσι ώστε σε κάθε επανάληψη, το 80% των δεδομένων να χρησιμοποιείται για εκπαίδευση και το υπόλοιπο 20% για επικύρωση (ως είσοδοι στη συνάρτηση anfis() του MATLAB). Μια συνάρτηση που μπορεί να βοηθήσει σε αυτό το έργο είναι η cvpartition(). Θα πρέπει κι εδώ να δοθεί προσοχή έτσι ώστε η κατανομή των κλάσεων να διατηρηθεί και στα δύο υποσύνολα. Ως μέθοδος επιλογής χαρακτηρηστικών επιλέγεται ο αλγόριθμος Relief και ως μέθοδος διαμέρισης διασκορπισμού ο αλγόριθμος Fuzzy C-Means (FCM - genfis3). Να εφαρμοστεί προεπεξεργασία των δεδομένων αν αυτό κριθεί απαραίτητο. Μετά το πέρας της διαδικασίας, να σχολιαστούν τα αποτελέσματα όσον αφορά το μέσο σφάλμα σε συνάρτηση με τις τιμές των παραμέτρων. Να δοθούν διαγράμματα τα οποία να απεικονίζουν την καμπύλη αυτού του σφάλματος σε σχέση με τον αριθμό των κανόνων και σε σχέση με τον αριθμό των επιλεχθέντων χαρακτηριστικών. Ποιά συμπεράσματα μπορούν να βγουν;
3. Να εκπαιδευτεί το τελικό TSK μοντέλο με τις βέλτιστες τιμές των παραμέτρων. Να δοθούν τα εξής διαγράμματα:
    - Διαγράμματα όπου να αποτυπώνονται οι προβλέψεις του τελικού μοντέλου καθώς και οι πραγματικές τιμές.
    - Διαγράμματα εκμάθησης όπου να απεικονίζεται το σφάλμα συναρτήσει του αριθμού επαναλήψεων.
    - Να δοθούν ενδεικτικά μερικά ασαφή σύνολα στην αρχική και τελική τους μορφή.
    - Να δοθεί ο πίνακας σφαλμάτων ταξινόμησης και να εξαχθούν από αυτόν τιμές των δεικτών απόδοσης $OA, PA, UA, \hat{K}$.
    - Τέλος, να σχολιαστούν τα αποτελέσματα όσον αφορά τα χαρακτηριστικά που επιλέχθηκαν και τον αριθμό IF-THEN κανόνων του ασαφούς συστήματος συμπερασμού. Να γίνει σύγκριση με τον αντίστοιχο αριθμό κανόνων αν για το ίδιο πλήθος χαρακτηριστικών, είχαμε επιλέξει grid partitioning με δύο ή τρία ασαφή σύνολα ανά είσοδο. Ποιά είναι τα συμπεράσματα; Τέλος, να γίνουν αντίστοιχα σχόλια όπως και στο πρώτο τμήμα, σχετικά με την επικάλυχη των προβολών των ασαφών συνόλων στο χώρο των μεταβλητών εισόδου και την επίδραση του διαμερισμού του συνολικού χώρου εισόδου στο ποσοστό των ενεργών κανόνων.