Μετασχηματισμοί Δεδομένων: Επιλογή Χαρακτηριστικών

ECE-TEL830 ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ

Αθανάσιος Κούτρας

Αναπληρωτής Καθηγητής

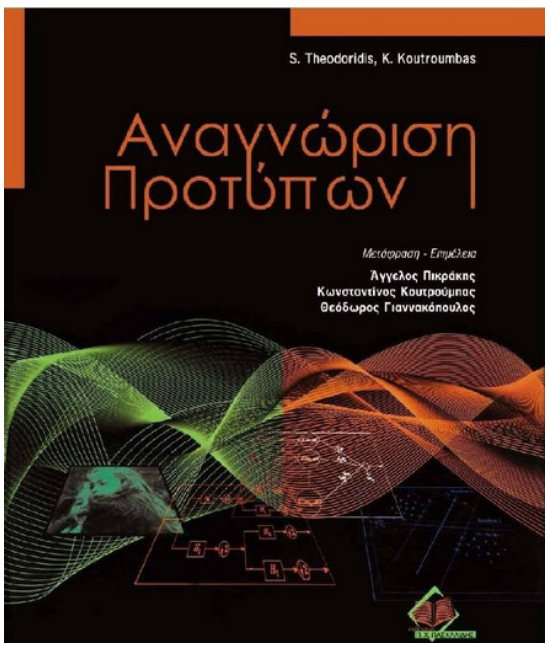
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών,
Παν. Πελοποννήσου

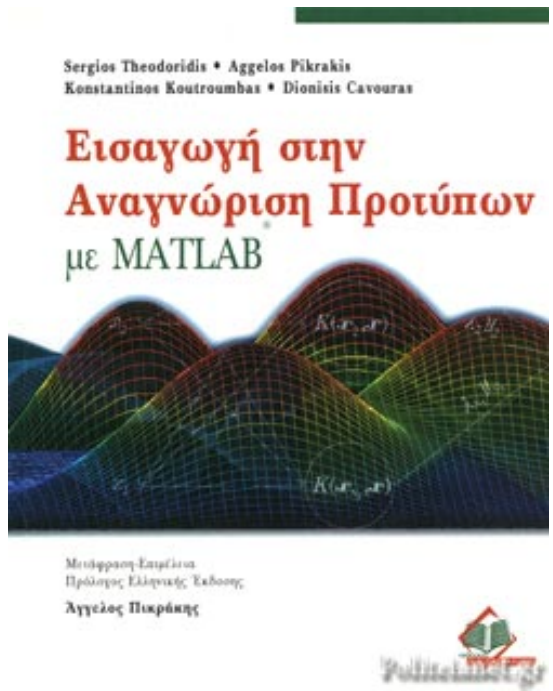
28 Μαΐου 2023

Περίγραμμα διάλεξης

Υλικό μελέτης

- Theodoridis S., Pikrakis, A., Koutroumbas K., Cavouras, D., **"Εισαγωγη στην αναγνώριση προτύπων με MATLAB",ΚΕΦΑΛΑΙΟ** 4
- Theodoridis S., Koutroumbas K., **"Αναγνώριση προτύπων",ΚΕΦΑΛΑΙΟ** 5





Εισαγωγή

- Στην διάλεξη αυτή θα παρουσιαστούν τεχνικές που ασχολούνται με την επιλογή ενός υποσυνόλου χαρακτηριστικών μέσα από ένα μεγαλύτερο σύνολο χαρακτηριστικών
- Ο στόχος είναι να επιλέξουμε εκείνα τα χαρακτηριστικά που παρουσιάζουν μεγάλη διακριτική ικανότητα μεταξύ των κατηγοριών στο πρόβλημα ταξινόμησης που αντιμετωπίζουμε.
- Η επιλογή των χαρακτηριστικών που είναι πλούσια σε πληροφορία έχει ως στόχο:
  - να απομακρύνει τις κλάσεις στον χώρο των χαρακτηριστικών όσο το δυνατό περισσότερο
  - να τοποθετήσει τα σημεία της ίδιας κλάσης όσο το δυνατό πιο κοντά (μικρή διασπορά κλάσης)
- η μείωση των χαρακτηριστικών έχει ως αποτέλεσμα την αποφυγή του προβλήματος overfitting, και την βελτίωση απόδοσης του ταξινομητή όταν βρεθεί αντιμέτωπος με δεδομένα εκτος συνόλου εκπαίδευσης (γενίκευση)
- πριν την επιλογή χαρακτηριστικών, πρέπει να εφαρμόσουμε ένα στάδιο προεπεξεργασίας με εργασίες όπως αποκοπή ακραίων τιμών και κανονικοποίηση δεδομένων.

Αποκοπή ακραίων τιμών

- **Ακραία τιμή**: ένα σημείο που βρίσκεται πολύ μακριά από τη μέση τιμή μιας τυχαίας μεταβλητής.
- επειδή αυτά τα σημεία μπορεί να οδηγήσουν σε μεγάλα σφάλματα ταξινόμησης, πρέπει να τα απομακρύνουμε.
- συνήθως είναι αποτέλεσμα θορύβου κατά τη λήψη των μετρήσεων.
- αν τα δεδομένα μας ακολουθούν την κανονική κατανομή, αποκόπτουμε τα σημεία τα οποία απέχουν από τη μέση τιμή περισσότερο από $1, 2$ ή 3 φορές από την τυπική απόκλιση
- αν η κατανομή δεν είναι κανονική, εφαρμόζονται άλλες πιο σύνθετες τεχνικές (με συναρτήσεις κόστους).

Κανονικοποίηση δεδομένων

- επειδή πολύ συχνά τυχαίνει το εύρος των τιμών να διαφέρει σημαντικά μεταξύ των χαρακτηριστικών, προκύπτουν ταξινομητές με χαμηλή απόδοση.
- χωρίς κανονικοποίηση, τα χαρακτηριστικά που παίρνουν μεγάλες τιμές, έχουν και μεγαλύτερη επίδραση στην συνάρτηση κόστους κατά τη σχεδίαση του ταξινομητή.
- με την διαδικασία της κανονικοποίησης, περιορίζουμε τις τιμές κάθε χαρακτηριστικού σε κάποια προκαθορισμένη έκταση.
- μια συνηθισμένη τεχνική είναι η κανονικοποίηση που οδηγεί σε μηδενική μέση τιμή και μοναδιαία διασπορά στην νέα, κανονικοποιημένη τιμή $\hat{x}_i$ σύμφωνα με:

$$\hat{x}_i = \frac{x_i - \bar{x}}{\sigma}, i = 1, 2, \cdots, N$$

Κανονικοποίηση δεδομένων

- μια άλλη τεχνική είναι ο περιορισμός της έκτασης των τιμών ενός χαρακτηριστικού μεταξύ ενός ελάχιστου και μέγιστου, π.χ. $[0, 1]$ ή $[-1.1]$
- μια τρίτη τεχνική περιλαμβάνει την χρήση μη γραμμικών μεθόδων (π.χ. softmax) που συμπιέζουν τα δεδομένα με μη γραμμικό τρόπο στο διάστημα $[0, 1]$:

$$\boxed{\hat{x_i} = \frac{1}{1 + exp(-y)}}$$

όπου $y = \frac{x_i - \bar{x}}{r\sigma}$, με την παράμετρο r να παίρνει τιμή από τον χρήστη

- η τεχνική αυτή εφαρμοόζεται όταν τα δεδομένα δεν είναι συμμετρικώς κατανεμημένα γύρω από τη μέση τιμή.

Έλεγχος υποθέσεων: t-TEST

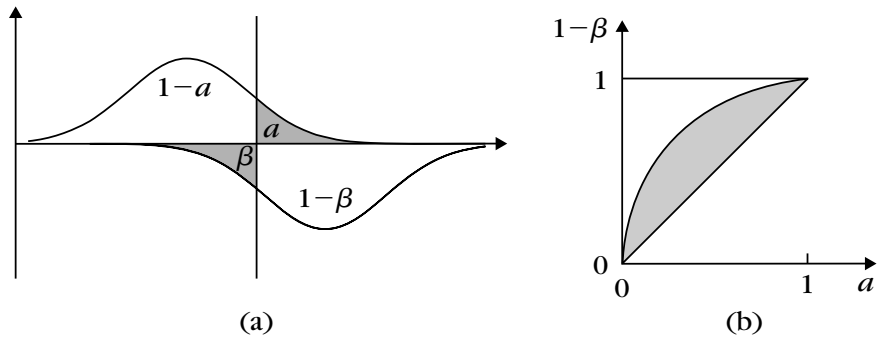
- το πρώτο βήμα στην επιλογή είναι να εξετάσουμε κάθε χαρακτηριστικό ξεχωριστά και να ελέγξουμε αν μεταφέρει αρκετή πληροφορία.
- σε περίπτωση που δεν ισχύει αυτό, τότε μπορούμε να αποκόψουμε το χαρακτηριστικό αυτό.
- μια κατηγορία τεχνικών αυτής της περίπτωσης είναι οι στατιστικοί έλεγχοι.
- μπορούμε να ελεγξουμε αν η μέση τιμή του χαρακτηριστικού διαφέρει σημαντικά μεταξύ των κλάσεων. Αν έχουμε περισσότερες από 2 κλάσεις, ο έλεγχος μπορεί να εφαρμοστεί σε κάθε ζεύγος κλάσεων.
- σε περίπτωση που έχουμε δεδομένα στις δύο κατηγορίες κανονικώς κατανεμημένα, χρησιμοποιείται ο έλεγχος υπόθεσης t-test.

το t-test

- Ο στόχος που έχουμε είναι να ελέγξουμε ποια από τις δύο παρακάτω υποθέσεις ισχύει:
  - H_1: *το χαρακτηριστικό έχει διαφορετική μέση τιμή σε κάθε κλάση*
  - H_0: *το χαρακτηριστικό έχει την ίδια μέση τιμή σε κάθε κλάση*
- η πρώτη υπόθεση είναι γνωστή ως εναλλακτική υπόθεση και ερμηνεύεται ως ένδειξη ότι οι τιμές του χαρακτηριστικού διαφέρουν σημαντικά μεταξύ των κλάσεων
- η δεύτερη υπόθεση είναι γνωστή ως μηδενική υπόθεση και ερμηνεύεται ως ένδειξη ότι οι τιμές του χαρακτηριστικού δε διαφέρουν σημαντικά.
- ένα χαρακτηριστικό απορρίπτεται, όταν ισχύει η μηδενική υπόθεση, σε διαφορετική περίπτωση επιλέγεται.
- το t-test γίνεται ως προς το λεγόμενο βαθμό σπουδαιότητας ρ που αντιστοιχεί στην πιθανότητα εσφαλμένης απόφασης. Συνήθως χρησιμοποιούνται τιμές $\rho = 0.05$ ή $\rho = 0.001$

η καμπύλη ROC

- Η καμπύλη Receiver Operating Characteristics (ROC) είναι ένα μέτρο της διαχωριστικής ικανότητας ενός χαρακτηριστικού μεταξύ των κλάσεων.
- μετρά την επικάλυψη μεταξύ των συναρτήσεων πυκνότητας πιθανότητας που περιγράφουν την κατανομή των τιμών ενός χαρακτηριστικού σε κάθε μια από τις δύο κλάσεις.
- η επικάλυψη ποσοτικοποιείται μέσω μιας περιοχή που ορίζουν οι δύο καμπύλες. Η περιοχή αυτή ονομάζεται Area Under the ROC - AUC).
- όταν έχουμε πλήρη επικάλυψη, η τιμή της AUC είναι κνοτά στο μηδέν, όταν οι κλάσεις διαχωρίζονται πλήρως, ισούται με 0.5



Σχήμα: (α) Επικαλυπτόμενες σ.π.π. του ίδιου χαρακτηριστικού σε δύο κλάσεις (σχεδιάστηκαν ανάποδα για λόγους καλύτερης απεικόνισης) και (β) η προκύπτουσα καμπύλη ROC.

Ο λόγος διάκρισης Fisher

- Ο λόγος διάκρισης Fisher (Fisher's Discriminant Ratio - FDR) χρησιμοποιείται για να ποσοτικοποιήσουμε τη διακριτική ικανότητα ενός χαρακτηριστικού μεταξύ δύο κλάσεων.
- ένα πλεονέκτημα που παρουσιάζει είναι ότι είναι ανεξάρτητος του είδους κατανομής που ακολουθεί η κλάση
- ο λόγος για δύο κλάσεις με μέσες τιμές και διασπορές m_1, m_2 και $\sigma_1 \sigma_2$ αντίστοιχα ορίζεται ως:

$$FDR = \frac{(m_1 - m_2)^2}{(\sigma_1^2 + \sigma_2^2)}$$

Μέτρα διαχωριστικής Ικανότητας

- οι προηγούμενες τεχνικές αναφέρονται στις διακριτικές ικανότητες των μεμονωμένων χαρακτηριστικών.
- αυτές όμως αγνοούν τη συσχέτιση που υπάρχει ανάμεσα σε διάφορα χαρακτηριστικά και επηρεάζει τις δυνατότητες ταξινόμησης των διανυσμάτων χαρακτηριστικών που προκύπτουν.
- στην ενότητα αυτή μελετάμε την διακριτική αποτελεσματικότητα διανυσμάτων χαρακτηριστικών.
- σκοπός μας είναι να συνδυάσουμε χαρακτηριστικά με κατάλληλο τρόπο και να καταλύξουμε στο καλύτερο διάνυσμα χαρακτηριστικών για δεδομένη διάσταση l
- για τον λόγο αυτό θα παρουσιαστούν μέτρα διαχωρισιμότητας κλάσεων που θα αξιοποιηθούν από διαδικασίες επιλογής χαρακτηριστικών.

Απόκλιση

- Για δύο κανονικώς κατανεμημένες κλάσεις στον $l-$διάστατο χώρο, η απόκλιση τους υπολογίζεται από:

$$d_{1,2} = \frac{1}{2} trace\{S_1^{-1} S_2 + S_2^{-1} S_1 - 2I\} + \frac{1}{2}(m_1 - m_2)^T (S_1^{-1} + S_2^{-1})(m_1 - m_2)$$

- το μέτρο της απόκλισης εξαρτάται σε σημαντικό βαθμό από τις διασπορές.
- ακόμα και για ίσες μέσες τιμές, η απόκλιση μπορεί να λάβει μεγάλες τιμές, αν διαφέρουν σημαντικά οι διασπορές των δύο κλάσεων.

Απόσταση Bhattacharyya και Φράγμα Chernoff

- για κανονικές κατανομές και στις δύο κλάσεις, η απόσταση Bhattacharyya ορίζεται ως:

$$\boxed{B_{1,2} = \frac{1}{8}(m_1 - m_2)^T(\frac{S_1 + S_2}{2})(m_1 - m_2) + A}$$

όπου

$$A = 0.5ln\left(\frac{0.5(|S_1 + S_2|)}{\sqrt{|S_1||S_2|}}\right)$$

με το σύμβολο $|\cdot|$ να δηλώνει την ορίζουσα του πίνακα.

- το φράγμα Shernoff είναι ένα άνω όριο του Bayesian σφάλματος και δίνεται από την εξίσωση:

$$\boxed{e_{CB} = exp(-B_{1,2})\sqrt{P(\omega_1)P(\omega_2)}}$$

όπου $P(\omega_1), P(\omega_2)$ είναι οι εκ των προτέρων πιθανότητες των κλάσεων.

Μέτρα βασισμένα σε Μητρώα Σκέδασης (Scatter Matrices)

- τα μητρώα αυτά αποτελούν την πιο δημοφιλή μέθοδο για την ποσοτικοποίηση του τρόπου που τα διανύσματα χαρακτηριστικών είναι διασκορπισμένα στον χώρο χαρακτηριστικών.
- χάρις στην πλούσια φυσική τους σημασία, έχουν αναπτυχθεί πολλά και διαφορετικά μέτρα διαχωρισιμότητας μεταξύ των κλάσεων:

$$\boxed{J_1 = \frac{trace\{S_m\}}{trace\{S_w\}}}$$

$$\boxed{J_2 = \frac{|S_m|}{|S_w|}}$$

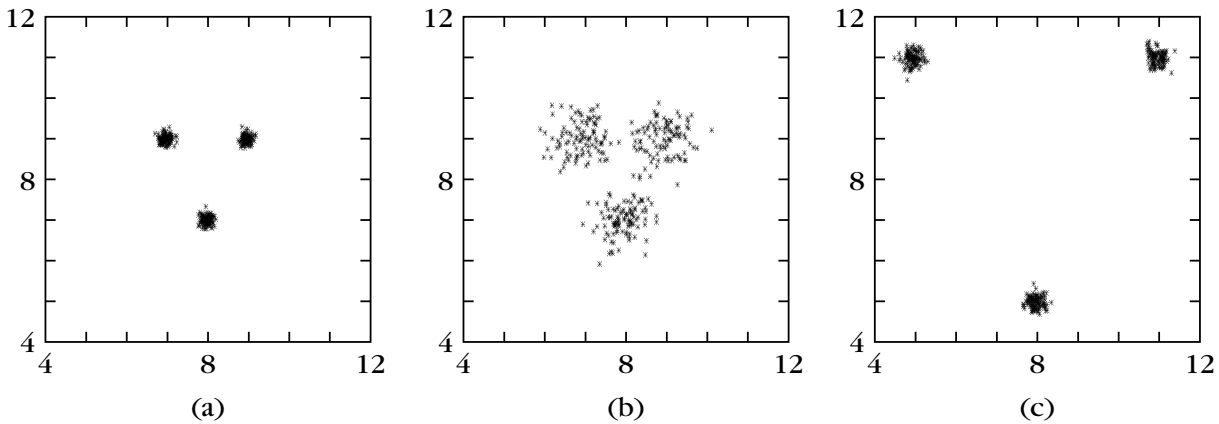
$$\boxed{J_3 = trace\{S_w^{-1} S_b\}}$$

Μέτρα βασισμένα σε Μητρώα Σκέδασης (Scatter Matrices)

- **το μητρώο σκέδασης εντός κλάσης** $\boxed{S_w = \sum_{i=1}^{C} P_i S_i}$, P_i, η a-priori πιθανότητα της κλάσης $i = 1, 2, \cdots, c$ και S_i το μητρώο συνδιασποράς της κλάσης
- **το μητρώο σκέδασης μεταξύ κλάσεων** $\boxed{S_b = \sum_{i=1}^{C} P_i (m_i - m_0)(m_i - m_0)^T}$ με $m_0 = \sum_{i=1}^{c} P_i m_i$ το καθολικό μέσο διάνυσμα (από τα δεδομένα όλων των κλάσεων)
- **το μητρώο μείξης σκέδασης** $\boxed{Sm = S_w + S_b}$
- μεγάλες τιμές των J_1, J_2, J_3 δηλώνουν ότι τα σημεία στον αντίστοιχο χώρο δεδομένων παρουσιάζουν μικρή διασπορά εντός κάθε κλάσης και μεγάλη απόσταση μεταξύ των κλάσεων.

Παράδειγμα

$J_3 = 164.7$ $J_3 = 12.5$ $J_3 = 620.9$



Σχήμα: Κλάσεις με (α) μικρή διασπορά εντός κλάσης και μικρές αποστάσεις μεταξύ κλάσεων (β) μεγάλη διασπορά εντός κλάσης και μικρές αποστάσεις μεταξύ κλάσεων (γ) μικρή διασπορά εντός κλάσης και μεγάλες αποστάσεις μεταξύ κλάσεων

Βαθμωτή επιλογή χαρακτηριστικών (Scalar Feature Selection)

- Για να μειώσουμε τον αριθμό των χαρακτηριστικών, μπορούμε να εξετάσουμε κάθε ένα ξεχωριστά και να χρησιμοποιήσουμε κάποιο από τα κριτήρια που αναφέρθηκαν (t-test, FDR, ROC) για να κρατήσουμε ή να απορρίψουμε ορισμένα από αυτά, ή για να τα ιεραρχήσουμε σε φθίνουσα τάξη και να επιλέξουμε τα l κορυφαία.
- επειδή αυτό είναι χρονοβόρο, χρησιμοποιούμε τον συντελεστή ετεροσυσχέτισης μεταξύ των χαρακτηριστικών

Βαθμωτή επιλογή χαρακτηριστικών (Scalar Feature Selection)

- **ταξινομούμε τα χαρακτηριστικά** σε φθίνουσα τάξη σύμφωνα με κάποιο κριτήριο C με τον δείκτη i_1 να δείχνει στο καλύτερο χαρακτηριστικό.
- **υπολογίζουμε την ετεροσυσχέτιση** μεταξύ του πρώτου στην ιεραρχία και καθενός από τα υπόλοιπα για να βρούμε το δεύτερο καλύτερο χαρακτηριστικό i_2:

$$\boxed{i_2 = \max_i \{\alpha_1 C_j - \alpha_2 |\rho_{i_1,j}|\}, j \neq i_1}$$

με C η τιμή του κριτηρίου που χρησιμοποιούμε για το $j-$στό χαρακτηριστικό και $\rho_{i_1,j}$ είναι η ετεροσυσχέτιση μεταξύ του καλύτερου χαρακτηριστικού i_1 και κάθε χαρακτηριστικού $j \neq i$. Οι παράμετροι α_1 και α_2 καθορίζονται από τον χρήστη.

- για την ιεράρχιση των υπόλοιπων χαρακτηριστικών, υπολογίζουμε με παρόμοιο με τον παραπάνω τρόπο την:

$$\boxed{i_k = \max_j \{\alpha_1 C_j - \frac{\alpha_2}{k-1} \sum_{r=1}^{k-1} |\rho_{i_r,j}|\}, j \neq i_r, r = 1, 2, \cdots, k-1}$$

Διανυσματική επιλογή Χαρακτηριστικών (Vector Feature Selection)

- η αντιμετώπιση του κάθε χαρακτηριστικού ξεχωριστά έχει το πλεονέκτημα της απλότητας, αλλά δεν είναι αποδοτική σε σύνθετα προβλήματα και χαρακτηριστικά που παρουσιάζουν μεγάλη αμοιβαία συσχέτιση.
- για τον λόγο αυτό εξετάζουμε τεχνικές που μετρούν τις ικανότητες ταξινόμησης διανυσμάτων χαρακτηριστικών.
- αν και παρουσιάζουν μεγαλύτερη πολυπλοκότητα και υπολογιστικό φόρτο, χρησιμοποιούνται πιο συχνά.
- ανάλογα με την βελτιστοποίηση που θα χρησιμοποιηθεί, η διαδικασία επιλογής χαρακτηριστικών ταξινομείται σε δύο κατηγορίες:
  1. Προσέγγιση φίλτρου (filter approach)
  2. Προσέγγιση Συνολικότητας (wrapper approach)

Κατηγορίες επιλογής χαρακτηριστικών

- **Προσέγγιση φίλτρου**
  - ο κανόνας βελτιστοποίησης για την επιλογή χαρακτηριστικών είναι ανεξάρτητος του ταξινομητή που θα χρησιμοποιηθεί στο στάδιο σχεδίασης του ταξινομητή
  - για κάθε συνδυασμό χαρακτηριστικών που εξετάζουμε, χρησιμοποιούμε κάποιο από τα προηγούμενα κριτήρια διαχωρισμού (J3, απόσταση Bhattacharrya)
  - ο συνολικός αριθμός συνδυασμών είναι πολύ μεγάλος

$$\boxed{\binom{m}{l} = \frac{m!}{l!(m-l)!}}$$

- **Προσέγγιση Συνολικότητας**
  - σε αυτή την περίπτωση η επιλογή γίνεται αξιοποιώντας τα αποτελέσματα που παίρνουμε πάνω στην απόδοση του ταξινομητή αυτού καθ' εαυτού.
  - η προσέγγιση αυτή αυξάνει την πολυπλοκότητα ακόμα περισσότερο, ανάλογα και με τον τύπο του ταξινομητή που χρησιμοποιούμε.
- και για τις δύο παραπάνω περιπτώσεις έχουν προταθεί πλήθος αποτελεσματικών τεχνικών αναζήτησης για μείωση πολυπλοκότητας (βέλτιστες - υποβέλτιστες).

Διανυσματική επιλογή Χαρακτηριστικών (Vector Feature Selection)

- Υποθέτουμε ότι έχουν επιβιώσει m χαρακτηριστικά από το προηγούμενο βήμα. Αυτό που θέλουμε τώρα είναι να βρούμε τον "καλύτερο" συνδυασμό αυτών των χαρακτηριστικών.
- συνήθως ξέρουμε εκ των προτέρων την τελική διάσταση l του χώρου των χαρακτηριστικών ή ένα άνω όριο αυτής.
- θα μπορούσαμε να εξετάσουμε όλους τους πιθανούς συνδυασμούς, να σχεδιάσουμε σε κάθε περίπτωση έναν ταξινομητή, να μετρήσουμε την απόδοση του με κάποια μέθοδο και μετά να επιλέξουμε τον καλύτερο συνδυασμό.
- αυτό είναι απαγορευτικό στην πράξη και για αυτό χρησιμοποιούνται:
  - εξαντλητική αναζήτηση (exhaustive)
  - σειριακή προς τα εμπρός (sequential forward) και σειριακή προς τα πίσω (sequential backward) επιλογή
  - επιλογή της προς τα εμπρός κινητής αναζήτησης (forward floating search selection)

Εξαντλητική αναζήτηση

- γίνονται όλοι οι δυνατοί συνδυασμοί χαρακτηριστικών και για κάθε συνδυασμό θα υπολογιστεί κάποιο από τα γνωστά μέτρα διαχωριστικής ικανότητας μεταξύ κλάσεων
- αν και είναι βέλτιστη, έχει μεγάλες υπολογιστικές απαιτήσεις
- τις περισσότερες φορές δεν μπορεί να υλοποιηθεί (μεγάλος αριθμός χαρακτηριστικών)

Υποβέλτιστες τεχνικές Αναζήτησης

- για να ξεπεραστούν τα προβλήματα της εξαντλητικής αναζήτησης, χρησιμοποιούνται υποβέλτιστες τεχνικές, χαμηλού υπολογιστικού κόστους.
- οι κυριότερες από αυτές είναι η Sequential Forward Selection (SFS), Sequential Backward Selection (SBS), Sequential Forward Floating Selection (SFFS)

Sequential Forward Selection

- ξεκινάμε με ένα χαρακτηριστικό και υπολογίζουμε την τιμή του κριτηρίου για κάθε ένα. Επιλέγουμε αυτό που έχει την καλύτερη τιμή.
- προσθέτουμε ένα χαρακτηριστικό και δημιουργούμε τον συνδυασμό όλων των δυνατών.
- υπολογίζουμε τις τιμές του κριτηρίου και επιλέγουμε την καλύτερη.
- στο διάνυσμα που επιλέχθηκε, προσθέτουμε ένα χαρακτηριστικά ξανά και δημιουργούμε πάλι όλους τους συνδυασμούς.
- υπολογίζουμε και πάλι τις τιμές του κριτηρίου και επιλέγουμε την καλύτερη.
- η διαδικασία επαναλαμβάνεται μέχρι να καταλήξουμε σε ένα διάνυσμα l χαρακτηριστικών.
- είναι υποβέλτιστη τεχνική αναζήτησης καθώς δεν μπορεί να εγγυηθεί ότι το βέλτιστο διάνυσμα θα προέρχεται από τα προηγούμενα διανύσματα.

Sequential Backward Selection

- ξεκινάμε με διάνυσμα χαρακτηριστικών αποτελούμενο από όλα τα χαρακτηριστικά.
- αφαιρούμε ένα χαρακτηριστικό και δημιουργούμε τον συνδυασμό όλων των υπολοίπων.
- υπολογίζουμε τις τιμές του κριτηρίου και επιλέγουμε την καλύτερη.
- στο διάνυσμα που επιλέχθηκε, μειώνουμε κατά ένα τα χαρακτηριστικά και δημιουργούμε πάλι όλους τους συνδυασμούς.
- υπολογίζουμε και πάλι τις τιμές του κριτηρίου και επιλέγουμε την καλύτερη.
- η διαδικασία επαναλαμβάνεται μέχρι να καταλήξουμε σε ένα διάνυσμα l χαρακτηριστικών.
- είναι υποβέλτιστη τεχνική αναζήτησης καθώς δεν μπορεί να εγγυηθεί ότι το βέλτιστο διάνυσμα θα προέρχεται από τα προηγούμενα διανύσματα.
- οι συνολικοί συνδυασμοί είναι ίσοι με:

$$\boxed{1 + 1/2((m+1)m - l(l+1))}$$

Floating Search methods

- οι προηγούμενες μέθοδοι υποφέρουν από την επίδραση του εμφωλιασμού (nesting effect)
- σε περίπτωση που ένα χαρακτηριστικό απορριφθεί, δεν υπάρχει περίπτωση να επανεξεταστεί στην SBS
- το ίδιο ισχύει και για την SFS, τη στιγμή που θα επιλεγεί ένα χαρακτηριστικό, δεν μπορεί να απορριφθεί αργότερα.
- για την αντιμετώπιση του προβλήματος αυτού προτάθηκαν η τεχνική κινητής αναζήτησης (floating search method) η οποία δίνει τη δυνατότητα απόρριψης χαρακτηριστικού που έχει επιλεγεί ή προσθήκης χαρακτηριστικού που έχει απορριφθεί σε προηγούμενο βήμα.
- εμφανίζεται σε δύο σχήματα: εμπρόσθιας αναζήτησης και οπισθοδρομικής αναζήτησης.

Signal & Image Processing, Pattern Recognition Group (SIPPRE)

www.sippre-group.com