ΔΗΜΙΟΥΡΓΙΑ ΒΑΣΕΩΝ ΚΑΝΟΝΩΝ ΑΠΟ ΔΕΔΟΜΕΝΑ
Μέρος Β': Εξαγωγή Κανόνων

Διδάσκων:

Ι. ΧΑΤΖΗΛΥΓΕΡΟΥΔΗΣ

Πανεπιστήμιο Πατρών, Τμήμα Μηχ/κών Η/Υ και Πληροφορικής

ΕΞΑΓΩΓΗ ΚΑΝΟΝΩΝ

- Η εξαγωγή κανόνων ουσιαστικά αποτελεί ένα πρόβλημα εξαγωγής ενός μοντέλου κατηγοριοποίησης (classification) δεδομένων (υπό μορφή κανόνων) .
- Υπάρχουν δύο βασικές διαδικασίες
  - **Μέθοδος αναμονής (holdout method)**
  - **Μέθοδος διασταυρωμένης επικύρωσης (cross validation method)**

ΜΕΘΟΔΟΣ ΑΝΑΜΟΝΗΣ

Η διαδικασία έχει ως εξής:

1. Χωρισμός του συνόλου δεδομένων σε δύο σύνολα: σύνολο εκπαίδευσης (ΣΕΚ), σύνολο ελέγχου (ΣΕΛ) (συνήθης σχέση μεγέθους 2:1 ή 3:1)
2. Εξαγωγή κανόνων με βάση το ΣΕΚ.
3. Αξιολόγηση των κανόνων με βάση το ΣΕΛ (υπολογισμός μετρικών).

ΜΕΘΟΔΟΣ ΕΠΙΚΥΡΩΜΕΝΗΣ ΔΙΑΣΤΑΥΡΩΣΗΣ (CROSS VALIDATION METHOD)

- Η πιο διαδεδομένη μορφή είναι η **k-fold cross validation**
- Η διαδικασία έχει ως εξής:

1. Χωρισμός του συνόλου δεδομένων σε k υποσύνολα (D1, D2, …, Dk)
2. Για i=1,k

   2.1 Ορίζεται ως σύνολο εκπαίδευσης (ΣΕΚ) το D-Di και ως σύνολο ελέγχου (ΣΕΛ) το Di

   2.2 Εξαγωγή κανόνων με βάση το ΣΕΚ

   2.3 Αξιολόγηση των κανόνων με βάση το ΣΕΛ
3. Υπολογισμός μέσων τιμών μετρικών στα k περάσματα

ΜΕΘΟΔΟΙ ΕΞΑΓΩΓΗΣ ΚΑΝΟΝΩΝ

- Δένδρα Απόφασης (ID3, C4.5)
- Μέθοδοι επαγωγής κανόνων
- Βασισμένες σε SVMs
- Βασισμένες σε Νευρωνικά Δίκτυα
- Βασισμένες σε Γενετικούς Αλγορίθμους

ΜΕΘΟΔΟΙ ΕΞΑΓΩΓΗΣ ΚΑΝΟΝΩΝ

- **Δένδρα Απόφασης (ID3, C4.5)**
- Μέθοδοι επαγωγής κανόνων
- Βασισμένες σε SVMs
- Βασισμένες σε Νευρωνικά Δίκτυα
- Βασισμένες σε Γενετικούς Αλγορίθμους

ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ (DECISION TREES)

- Τα δέντρα απόφασης είναι μια μέθοδος δημιουργίας προτασιακών κανόνων-με άλλες λέξεις ενός μοντέλου κατηγοριοποίησης, το οποίο έχει μορφή δέντρου, από δεδομένα.
- Χρήση της τεχνικής «διαίρει και βασίλευε» για διαίρεση του χώρου αναζήτησης σε υποσύνολα (ορθογώνιες περιοχές).
- Ένα παράδειγμα κατηγοριοποιείται με βάση την περιοχή στην οποία ανήκει.

ΟΡΙΣΜΟΣ

- **Δέντρο Απόφασης** (ΔΑ) ή Δέντρο Κατηγοριοποίησης είναι ένα δέντρο με τις ακόλουθες ιδιότητες:
  - Κάθε εσωτερικός κόμβος και η ρίζα ονοματίζεται με το όνομα ενός χαρακτηριστικού.
  - Κάθε κλάδος ονοματίζεται με ένα κατηγόρημα διάσπασης του χαρακτηριστικού που αποτελεί το όνομα του κόμβου-πατέρα.
  - Κάθε φύλλο ονοματίζεται με το όνομα μιας κλάσης

ΒΑΣΙΚΟΣ ΑΛΓΟΡΙΘΜΟΣ

Input: *D* //σύνολο εκπαίδευσης

Output: *T* //ζητούμενο δέντρο απόφασης

Algorithm: DTBuild

T = ∅;

Determine best splitting criterion;

T = Create root node and label with splitting attribute;

T = Add arc to root node for each splitting predicate and label;

for each arc do

D = database created by applying splitting predicate to *D*;

if stopping point reached for this path

then *T'* = create leaf mode and label with appropriate class;

else *T'* = DTBuild (*D*);

T = Add *T'* to arc;

ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ

- Χαρακτηριστικά διάσπασης (splitting features)

  Τα χαρακτηριστικά των παραδειγμάτων στη βάση D που χρησιμοποιούνται σαν ονόματα κόμβων του δέντρου, δηλ. επιλέχτηκαν ως καλύτερα χαρακτηριστικά.

- Χαρακτηριστικό στόχου (target feature)

  Το χαρακτηριστικό που οι τιμές του αντιπροσωπεύουν τις κλάσεις κατηγοριοποίησης.

- Κατηγορήματα διάσπασης (splitting predicates)

  Τα κατηγορήματα που χρησιμοποιούνται σαν ονόματα των κλάδων του δέντρου.

ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ

- Κριτήριο διάσπασης (splitting criterion).

  Το κριτήριο με βάση το οποίο επιλέγεται το καλύτερο χαρακτηριστικό διάσπασης κάθε φορά.

- Κριτήριο τερματισμού (stopping criterion).

  Το κριτήριο με βάση το οποίο τερματίζεται ο αλγόριθμος.

Παραλλαγές των δύο αυτών κριτηρίων δημιουργούν μια ποικιλία αλγορίθμων.

ΒΑΣΙΚΑ ΘΕΜΑΤΑ

- **Επιλογή χαρακτηριστικών διάσπασης**
  - ✓ Διαφορετικά σύνολα χαρακτηριστικών διάσπασης έχουν σαν αποτέλεσμα διαφορετικά ΔΑ με διαφορετική απόδοση.
  - ✓ Η επιλογή τους στηρίζεται όχι μόνο στο σύνολο εκπαίδευσης, αλλά και στη γνώμη του εμπειρογνώμονα.
- **Διάταξη των χαρακτηριστικών διάσπασης**
  - ✓ Η σειρά επιλογής των χαρακτηριστικών διάσπασης παίζει σημαντικό ρόλο στην απόδοση ενός ΔΑ.
  - ✓ Ο αριθμός διασπάσεων συνδέεται με τη διάταξη των χαρακτηριστικών διάσπασης. Ο αριθμός διασπάσεων μπορεί εύκολα να προσδιοριστεί όταν το πεδίο είναι μικρό (λίγα χαρακτηριστικά, λίγες και διακριτές τιμές), αλλιώς (πολλά χαρακτηριστικά ή πολλές/συνεχείς τιμές) τα πράγματα δυσκολεύουν.

ΒΑΣΙΚΑ ΘΕΜΑΤΑ

- **Δομή του δέντρου**
  - ✓ Επιθυμητό είναι να δημιουργούνται δέντρα που είναι ισορροπημένα και με τα λιγότερα επίπεδα (μικρότερο βάθος). Αυτό όμως δεν είναι πάντα εφικτό ούτε το υπολογιστικά φτηνότερο.
  - ✓ Μερικοί αλγόριθμοι δημιουργούν μόνο δυαδικά δέντρα.
- **Κριτήρια τερματισμού**
  - ✓ Η δημιουργία ενός δέντρου σταματά οπωσδήποτε όταν όλα τα δεδομένα του (εναπομείναντος) συνόλου εκπαίδευσης κατηγοριοποιούνται πλήρως.
  - ✓ Μπορεί όμως να είναι απαραίτητο να σταματήσει νωρίτερα για να αποφευχθούν π.χ. μεγάλα δέντρα. Το πότε ή πού θα σταματήσει είναι θέμα συναλλαγής (trade-off) μεταξύ ακρίβειας (accuracy) και απόδοσης (performance) του αλγορίθμου.

ΒΑΣΙΚΑ ΘΕΜΑΤΑ

- **Κριτήρια τερματισμού (συν.)**
  - ✓ Επίσης, πρώιμος τερματισμός μπορεί να γίνει για αποφυγή του φαινομένου της υπερπροσαρμογής (overfitting).
  - ✓ Τέλος, μπορεί να προχωρήσει σε μεγαλύτερα δέντρα αν είναι γνωστό ότι υπάρχουν κατηγορίες δεδομένων που δεν αντιπροσωπεύονται στο σύνολο εκπαίδευσης.
- **Δεδομένα εκπαίδευσης**
  - ✓ Η δομή ενός ΔΑ εξαρτάται από τα δεδομένα εκπαίδευσης. Αν το σύνολο εκπαίδευσης είναι πολύ μικρό, τότε το δέντρο μπορεί να μην είναι τόσο λεπτομερές, ώστε να ταξινομεί γενικότερα δεδομένα. Αν είναι πολύ μεγάλο, το δέντρο πιθανόν να υπερπροσαρμόζεται (overfits).

ΒΑΣΙΚΑ ΘΕΜΑΤΑ

- **Κλάδεμα (Pruning)**
  - ✓ Μετά τη δημιουργία ενός ΔΑ μπορεί να χρειάζονται τροποποιήσεις για να βελτιώσουν την απόδοσή του, όπως π.χ. το κλάδεμα πλεοναζόντων συγκρίσεων ή υποδέντρων.

ΠΟΛΥΠΛΟΚΟΤΗΤΑ

Η πολυπλοκότητα χρόνου και χώρου των αλγορίθμων ΔΑ εξαρτώνται από το μέγεθος του συνόλου εκπαίδευσης *k*, τον αριθμό των χαρακτηριστικών διάσπασης *n* και το σχήμα του ΔΑ. Στη χειρότερη περίπτωση το ΔΑ είναι βαθύ και μη ισορροπημένο.

- ✓ Η πολυπλοκότητα χρόνου για τη δημιουργία ενός ΔΑ είναι $O(n*k*\log k)$
- ✓ Η πολυπλοκότητα χρόνου κατηγοριοποίησης μιας βάσης n παραδειγμάτων εξαρτάται από το ύψος του ΔΑ και είναι $O(n*\log k)$, υποθέτοντας πολυπλοκότητα για το ύψος $O(\log k)$.

ΑΛΓΟΡΙΘΜΟΣ ID3

- Χρησιμοποιεί σαν κριτήριο για τον προσδιορισμό του «καλύτερου χαρακτηριστικού διάσπασης» το «κέρδος πληροφορίας» (information gain).
- Το «κέρδος πληροφορίας» μετριέται ποσοτικά με την εντροπία (entropy).
  - ✓ Η εντροπία εν γένει εκφράζει το μέγεθος της ανομοιογένειας σε ένα σύνολο δεδομένων. Π.χ. αν όλα τα δεδομένα ανήκουν σε μια κλάση, τότε δεν υπάρχει ανομοιογένεια: η εντροπία είναι μηδέν.
  - ✓ Το ζητούμενο σ' ένα ΔΑ είναι ο διαχωρισμός του συνόλου εκπαίδευσης , μ' ένα επαναληπτικό τρόπο, σε υποσύνολα μηδενικής εντροπίας.
  - ✓ Αν p η πιθανότητα να συμβεί ένα γεγονός, τότε log(1/p) παριστάνει το ποσό της τυχαιότητας με βάση την πιθανότητα.
  - ✓ Η αναμενόμενη πληροφορία με βάση την p ορίζεται: p log(1/p)
  - ✓ Αν έχω δύο συμπληρωματικά γεγονότα e, e' με p, p', τότε η αναμενόμενη πληροφορία είναι: p log(1/p) + p' log(1/p')

ΑΛΓΟΡΙΘΜΟΣ ID3

Ορισμός (εντροπία)

Δεδομένων των πιθανοτήτων $p_1, p_2, \ldots p_k$, όπου $\Sigma p_i = 1$, η εντροπία Ε ορίζεται ως εξής:

$$E(S) = E(p_1, p_2, \ldots, p_k) = \Sigma(p_i \log(1/p_i))$$

Ορισμός (κέρδος πληροφορίας)

$$G(S, X) = E(S) - \Sigma(|S_i|/|S|)\, E(S_i)$$

όπου S_i υποσύνολο του S, που περιέχει τα παραδείγματα του S με τιμή x_i για το χαρακτηριστικό X.

ΑΛΓΟΡΙΘΜΟΣ ID3-ΠΑΡΑΔΕΙΓΜΑ

No	Outlook	Temp.	Humid.	Wind	PlayTennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

ΑΛΓΟΡΙΘΜΟΣ ID3-ΠΑΡΑΔΕΙΓΜΑ

Επιλογή ρίζας

$G(S, Outlook) = E(S) - (|S_{sunny}|/|S|)\, E(S_{sunny}))$

$- (|S_{overcast}|/|S|)\, E(S_{overcast}))$

$- (|S_{rain}|/|S|)\, E(S_{rain}))$

Χαρακτηριστικό-στόχος: PlayTennis:PT (yes, no)

$p_1 = p(PT=yes) = 9/14$, $p_2 = p(PT=no) = 5/14$

$E(S) = p_1 \log(1/p_1) + p_2 \log(1/p_2) = -p_1 \log(p_1) – p_2 \log(p_2) =$

$-(9/14) \log(9/14) – (5/14) \log(5/14) = 0{,}283$

ΑΛΓΟΡΙΘΜΟΣ ID3-ΠΑΡΑΔΕΙΓΜΑ

$E(S_{sunny}) = -(2/5)\log(2/5) – (3/5)\log(3/5) = 0,292$

$E(S_{rain}) = -(3/5)\log(3/5) – (2/5)\log(2/5) = 0,292$

$E(S_{overcast}) = -(4/4)\log(4/4) – (0/4)\log(0/4) = 0$

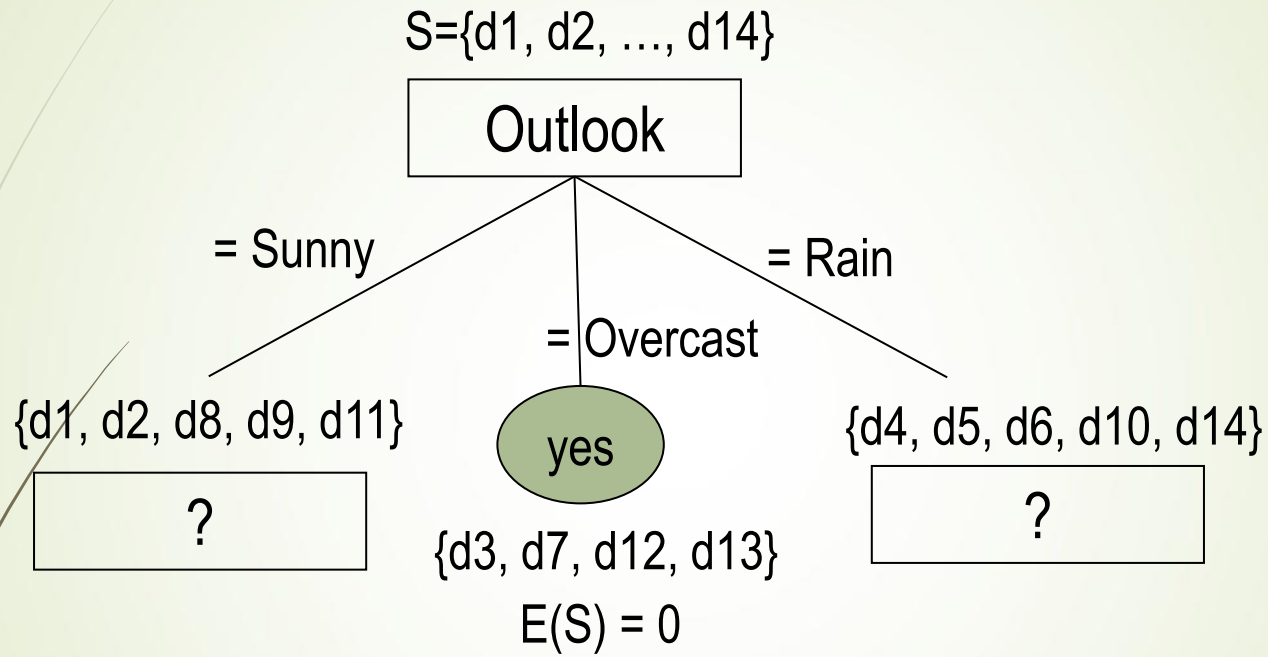
$|S| = 14, |S_{sunny}| = 5, |S_{rain}| = 5, |S_{overcast}| = 4$

G(S, Outlook) = 0,074

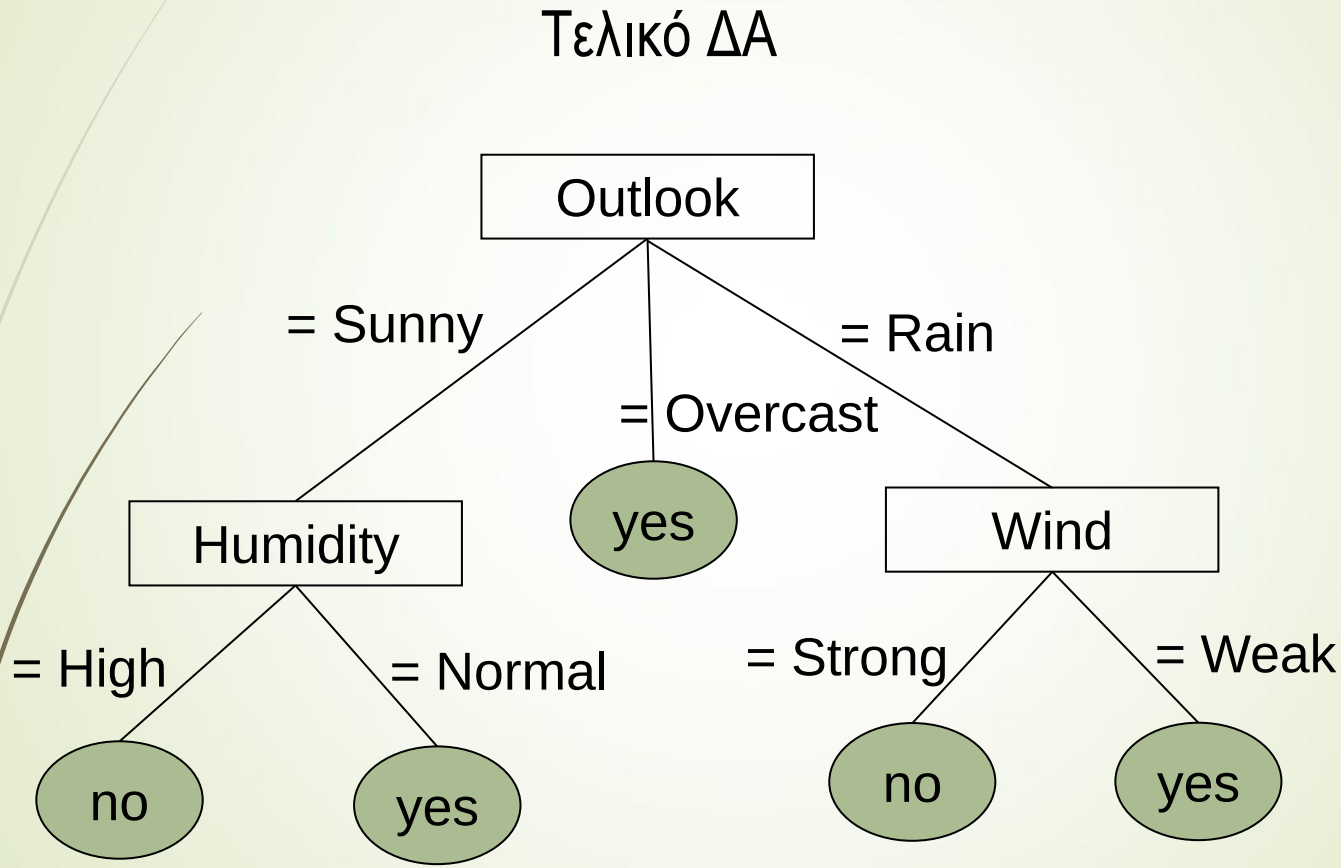
Ομοίως G(S, Humidity) = 0,04565, G(S, Wind) = 0,0144

και G(S, Temperature) = 0,0087

ΑΛΓΟΡΙΘΜΟΣ ID3-ΠΑΡΑΔΕΙΓΜΑ



ΑΛΓΟΡΙΘΜΟΣ ID3-ΠΑΡΑΔΕΙΓΜΑ



ΑΛΓΟΡΙΘΜΟΣ ID3-ΠΑΡΑΔΕΙΓΜΑ

Εξαγόμενοι Κανόνες

if outlook is sunny and
and humidity is high
then playtennis is no

if outlook is rain and
and wind is strong
then playtennis is no

if outlook is sunny and
and humidity is normal
then playtennis is yes

if outlook is rain and
and wind is weak
then playtennis is yes

if outlook is overcast
then playtennis is yes

ΙΔΙΟΤΗΤΕΣ ID3

- Προτιμά
  - ✓ τα μικρότερα δέντρα από τα μεγαλύτερα
  - ✓ τοποθετεί χαρακτηριστικά με υψηλό κέρδος πληροφορίας κοντύτερα στη ρίζα
- Είναι αλγόριθμος αναζήτησης τύπου Hill Climbing, που
  - ✓ Προχωρά από τα απλά στα σύνθετα ξεκινώντας από το κενό δέντρο
  - ✓ Ψάχνει στον πλήρη χώρο των υποθέσεων (όλων των πιθανών δέντρων)
  - ✓ Διατηρεί μόνο μια υπόθεση κάθε φορά
  - ✓ Δεν κάνει οπισθοδρόμηση (backtracking), δηλ. δεν αναθεωρεί προηγούμενη απόφαση/επιλογή (κίνδυνος τοπικού βέλτιστου)
  - ✓ Χρησιμοποιεί όλα τα δεδομένα εκπαίδευσης (λιγότερο ευαίσθητος σε λάθη)
  - ✓ Δεν φτάνει σε αποφάσεις αυξητικά, δηλ. βασιζόμενος σε ατομικά δεδομένα

ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ-ΔΥΝΑΤΟΤΗΤΕΣ ID3

- Τα παραδείγματα (δεδομένα) αναφέρονται σε ένα συγκεκριμένο σύνολο χαρακτηριστικών και τις τιμές τους, που είναι διακριτές και, κατά προτίμηση, λίγες. Χειρισμός μεταβλητών με πραγματικές τιμές απαιτεί επέκταση του βασικού αλγορίθμου
- Η μεταβλητή (ή συνάρτηση) στόχου έχει διακριτές τιμές, συνήθως δύο (π.χ. PlayTennis → yes, no) (boolean classification). Η επέκταση για έξοδο με περισσότερες από δύο τιμές είναι εύκολη. Δυσκολότερη η επέκταση για χειρισμό εξόδου με συνεχείς (πραγματικές) τιμές (πράγμα όχι σύνηθες όμως)

ΑΛΓΟΡΙΘΜΟΣ C4.5

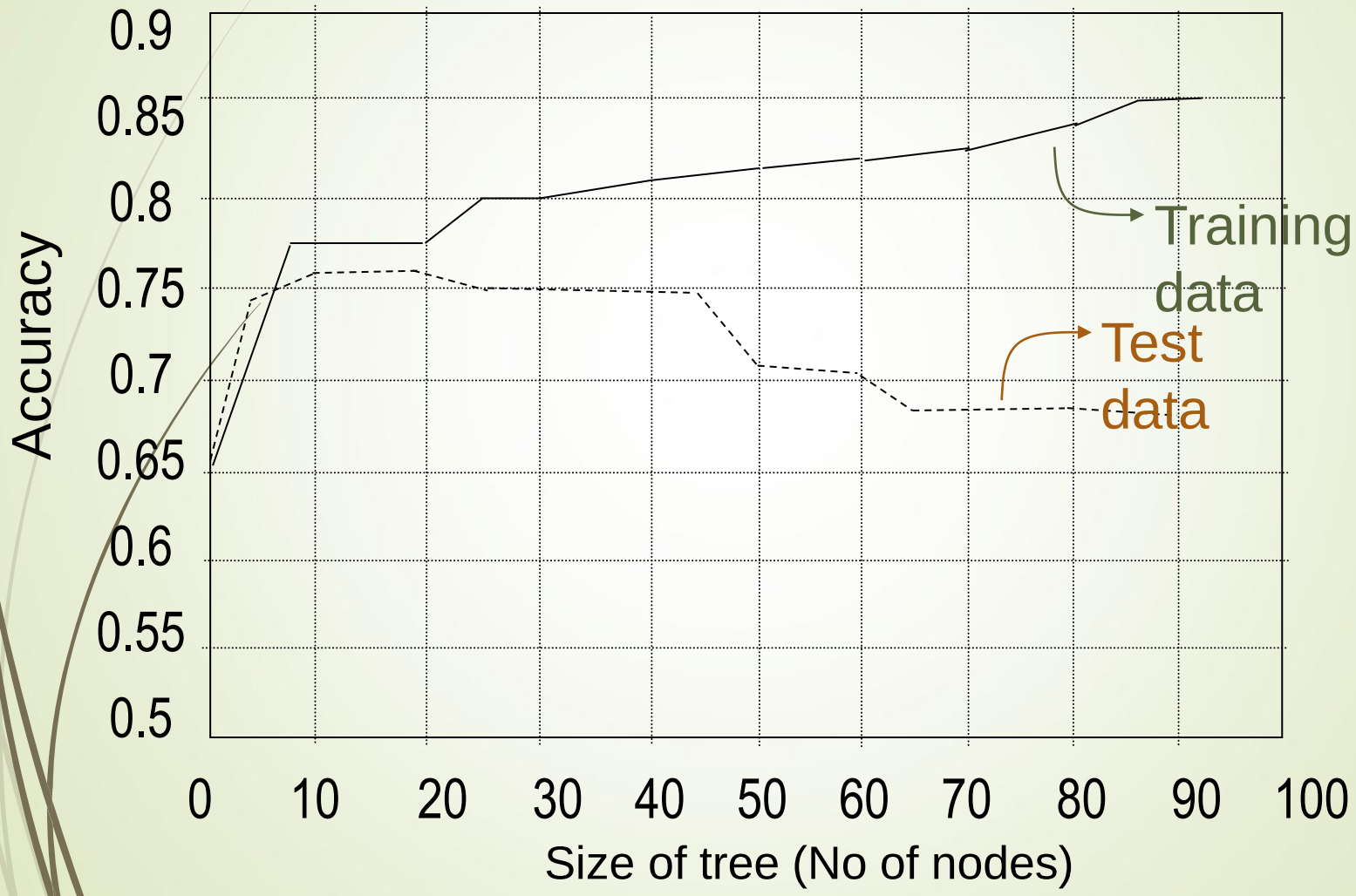
- Αποφυγή υπερπροσαρμογής (overfitting)
  - ✓ Reduced error pruning
  - ✓ Rule post-pruning
- Χειρισμός χαρακτηριστικών συνεχών τιμών
- Επιλογή καλύτερης μετρικής για την επιλογή των χαρακτηριστικών διάσπασης
- Χειρισμός συνόλου εκπαίδευσης με ελλιπείς τιμές
- Χειρισμός χαρακτηριστικών με διαφορετικά κόστη
- Βελτίωση υπολογιστικής απόδοσης

ΥΠΕΡΠΡΟΣΑΡΜΟΓΗ (OVERFITTING)

Ορισμός

Δεδομένου ενός χώρου υποθέσεων Η, μια υπόθεση $h \in H$ λέγεται ότι υπερπροσαρμόζει τα δεδομένα εκπαίδευσης, αν υπάρχει κάποια εναλλακτική υπόθεση $h' \in H$, τέτοια ώστε η h έχει μικρότερο λάθος (δηλ. καλύτερα αποτελέσματα) κατηγοριοποίησης από την h' ως προς τα δεδομένα εκπαίδευσης, αλλά η h' έχει μικρότερο λάθος από την h ως προς το σύνολο των δεδομένων. (Mitchell, 1997)

ΥΠΕΡΠΡΟΣΑΡΜΟΓΗ (OVERFITTING)



ΥΠΕΡΠΡΟΣΑΡΜΟΓΗ (OVERFITTING)

Πιθανοί λόγοι

- ✓ Τα δεδομένα εκπαίδευσης περιέχουν τυχαία λάθη ή θόρυβο
- ✓ Κάποια φύλλα αντιπροσωπεύουν μικρό αριθμό δεδομένων

Πιθανές λύσεις

- ✓ Πρόωρο σταμάτημα, πριν την πλήρη κατηγοριοποίηση (Δυσκολία στον προσδιορισμό του σημείου σταματήματος)
- ✓ Εκ των υστέρων κλάδεμα του δέντρου (reduced error pruning)

  (rule post-pruning)

  (Πιο αποτελεσματική λύση)

REDUCED ERROR PRUNING (ΚΛΑΔΕΜΑ ΕΛΑΤΤΩΜΕΝΟΥ ΛΑΘΟΥΣ)

- Χρησιμοποιεί σύνολο εγκυροποίησης
- Το κλάδεμα ξεκινά αφού δημιουργηθεί το δέντρο
- Κάθε κόμβος θεωρείται υποψήφιος για κλάδεμα
- Κλάδεμα ενός κόμβου σημαίνει διαγραφή του υποδέντρου που τον έχει ως ρίζα και μετατροπή του σε φύλλο, στο οποίο προσάπτεται η πιο κοινή κλάση των παραδειγμάτων εκπαίδευσης που σχετίζονται με τον κόμβο
- Ένας κόμβος κλαδεύεται μόνο αν προκύπτει δέντρο που δεν είναι χειρότερο από το αρχικό (με βάση την ακρίβεια ως προς το σύνολο εγκυροποίησης)
- Το κλάδεμα συνεχίζεται μέχρις ότου διαπιστωθεί μείωση της ακρίβειας

RULE POST-PRUNING

- Μια παραλλαγή του χρησιμοποιείται στον C4.5
- Διαδικασία
  1. Δημιούργησε το ΔΑ από το training set αναπτύσσοντάς το μέχρις ότου ικανοποιούνται όσο το δυνατόν περισσότερα παραδείγματα, επιτρέποντας υπερπροσαρμογή
  2. Μετάτρεψε το ΔΑ σ' ένα ισοδύναμο σύνολο κανόνων δημιουργώντας ένα κανόνα για κάθε μονοπάτι από τη ρίζα σε φύλλο
  3. Κλάδεψε/Γενίκευσε κάθε κανόνα διαγράφοντας κάθε συνθήκη που έχει σαν αποτέλεσμα τη βελτίωση της εκτιμώμενης ακρίβειας
  4. Διάταξε τους κλαδεμένους κανόνες με βάση την εκτιμώμενη ακρίβεια και θεώρησέ τους μ' αυτή τη σειρά όταν κατηγοριοποιείς παραδείγματα

RULE POST-PRUNING

Π.χ. Από το ΔΑ του παραδείγματος PlayTennis έχουμε:

IF (Outlook=Sunny) $\wedge$ (Humidity=High)

THEN PlayTennis=yes

1. Δοκιμάζουμε να αφαιρέσουμε τις συνθήκες μια-μια.
2. Για κάθε συνθήκη που αφαιρούμε εκτιμούμε την ακρίβεια του κανόνα. Η αφαίρεση εκτελείται αν οδηγεί σε μεγαλύτερη ακρίβεια.
3. Η εκτίμηση της ακρίβειας γίνεται
    1. Με τη χρήση validation set ξένου προς το training set
    2. Με τη χρήση του training set χρησιμοποιώντας μια απαισιόδοξη εκτίμηση, για να εξισορροπηθεί το γεγονός ότι το training set δίνει μια εκτίμηση που ευνοεί τους κανόνες (C4.5)

RULE POST-PRUNING

- Πιο συγκεκριμένα (βήμα 3.2)
  - ✓ Υπολογίζεται η ακρίβεια του κανόνα με βάση τα παραδείγματα εκπαίδευσης
  - ✓ Υπολογίζεται η τυπική απόκλιση (standard deviation: std) της ακρίβειας υποθέτοντας δυωνιμική κατανομή
  - ✓ Με δεδομένο το επίπεδο βεβαιότητας, σαν το χαμηλότερο όριο της εκτίμησης λαμβάνεται η απόδοση του κανόνα (π.χ. για βεβαιότητα 95%, η ακρίβεια του κανόνα εκτιμάται απαισιόδοξα ως η παρατηρηθείσα ακρίβεια-1,96*std)

ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΜΕ ΣΥΝΕΧΕΙΣ ΤΙΜΕΣ

- Ο ID3 περιορίζεται σε χαρακτηριστικά στόχου και διάσπασης με διακριτές τιμές.
- Όταν έχουμε χαρακτηριστικά με συνεχείς τιμές, τότε το πρόβλημα λύνεται σχετικά απλά για διακριτοποίηση σε δύο τιμές.
  - ✓ Για κάθε χαρακτηριστικό Χ που παίρνει συνεχείς τιμές δημιουργούμε μια δυαδική μεταβλητή Χc που είναι αληθής για Χc < C και ψευδής αλλού. Το πρόβλημα είναι η επιλογή της τιμής C.
  - ✓ Διατάσσουμε τα παραδείγματα με βάση τις τιμές του χαρακτηριστικού Χ και προσδιορίζουμε γειτονικά παραδείγματα, όπου έχουμε αλλαγή στην ταξινόμηση. Τότε παίρνουμε σαν C την ενδιάμεση τιμή της «καλύτερης» περίπτωσης, δηλ. της περίπτωσης με το μεγαλύτερο κέρδος πληροφορίας.

ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΜΕ ΣΥΝΕΧΕΙΣ ΤΙΜΕΣ

Παράδειγμα

Έστω ότι υπάρχει ένα χαρακτηριστικό Temperature που παίρνει συνεχείς τιμές και το S σ' ένα κόμβο περιέχει παραδείγματα με τιμές (διατεταγμένα)

Temperature	40	48	60	72	80	90
PlayTennis	no	no	yes	yes	yes	no

Υποψήφια κατώφλια: (48+60)/2 = 54, (80+90)/2 = 85

Υπολογίζουμε τα $G(S, Temp_{>54})$ και $G(S, Temp_{>85})$

Επειδή $G(S, Temp_{>54}) > G(S, Temp_{>85})$, επιλέγεται το C=54

Η προσέγγιση αυτή μπορεί να επεκταθεί και για διακριτοποίηση με περισσότερες από δύο τιμές.

ΕΝΑΛΛΑΚΤΙΚΕΣ ΜΕΤΡΙΚΕΣ

- Το κέρδος πληροφορίας ευνοεί χαρακτηριστικά με πολλές τιμές σε βάρος αυτών με λίγες.
  - ✓ Π.χ. το χαρακτηριστικό 'ημερομηνία' λόγω των πολλών τιμών θα είχε το μεγαλύτερο κέρδος πληροφορίας → δέντρο με πολλές διασπάσεις (πλατύ δέντρο), αλλά βάθος 1: τέλεια ταξινόμηση, αλλά όχι καλό δέντρο στη συνέχεια.
- Ένας τρόπος να το διορθώσουμε είναι να αλλάξουμε την μετρική:
  - ✓ $GainRatio(S, X) = Gain(S, X)/SplitInfo(S, X)$ (λόγος κέρδους)
  - ✓ $SplitInfo(S, X)= -\sum_{i \in \{x1, x2, \ldots, xk\}} (|Si|/|S|)*\log (|Si|/|S|)$ (πληροφορία διάσπασης)
  - ✓ Τιμωρούνται χαρακτηριστικά με πολλές τιμές μέσω της «πληροφορίας διάσπασης»