Δέντρα Απόφασης (Decision Trees)

- Το μοντέλο που δημιουργείται είναι ένα δέντρο
- 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

Ορισμός

- Δέντρο Απόφασης (ΔΑ) ή Δέντρο Κατηγοριοποίησης είναι ένα δέντρο με τις ακόλουθες ιδιότητες:
  - Κάθε εσωτερικός κόμβος ονοματίζεται με το όνομα ενός χαρακτηριστικού X_i.
  - Κάθε κλαδί/σύνδεση ονοματίζεται με ένα κατηγόρημα που μπορεί να εφαρμοστεί στο χαρακτηριστικό που αποτελεί το όνομα του κόμβου-πατέρα.
  - Κάθε φύλλο ονοματίζεται με το όνομα μιας κλάσης

Βασικός Αλγόριθμος

Input: *D* //δεδομένα 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;

Χαρακτηριστικές Έννοιες (1)

- Χαρακτηριστικά διάσπασης (splitting attributes)

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

- Χαρακτηριστικό στόχου (target attribute)

  Το χαρακτηριστικό που οι τιμές του αντιπροσωπεύουν τις κλάσεις κατηγοριοποίησης.

- Κατηγορήματα διάσπασης (splitting predicates)

  Τα κατηγορήματα που χρησιμοποιούνται σαν ονόματα των κλαδιών/συνδέσεων του δέντρου.

Χαρακτηριστικές Έννοιες (2)

- 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.

Βασικά Θέματα (1)

- 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-Διασπάσεις
  - 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.

Βασικά Θέματα (2)

- 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.
  - 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

Βασικά Θέματα (3)

- 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).
- 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.

Πολυπλοκότητα

- 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*logk)
  - Η πολυπλοκότητα χρόνου κατηγοριοποίησης μιας βάσης n παραδειγμάτων εξαρτάται από το ύψος του ΔΑ και είναι O(n*logk), υποθέτοντας πολυπλοκότητα για το ύψος O(logk).

Ιδιότητες 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

Χαρακτηριστικά-Δυνατότητες ID3 (1)

- 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)

Χαρακτηριστικά-Δυνατότητες ID3 (2)

- Απαιτούνται διαζευκτικές (disjunctive) περιγραφές. Ένα δέντρο απόφασης αναπαριστά μια διάζευξη συζεύξεων περιορισμών στις τιμές των χαρακτηριστικών των παραδειγμάτων. Κάθε μονοπάτι από τη ρίζα σ' ένα φύλλο αντιστοιχεί σε μια σύζευξη τεστ των χαρακτηριστικών, και το ίδιο το δέντρο μια διάζευξη αυτών των συζεύξεων:

  (Outlook=Sunny $\wedge$ Humidity=Normal) $\vee$

  (Outlook=Overcast) $\vee$

  (Outlook=Rain $\wedge$ Wind=Weak)
- Τα δεδομένα εκπαίδευσης μπορεί να περιέχουν λάθη. Ο ID3 είναι ανεκτικός στα λάθη (και στην κατηγοριοποίηση και στις τιμές των χαρακτηριστικών)
- Από τα δεδομένα εκπαίδευσης μπορεί να λείπουν τιμές για ορισμένα χαρακτηριστικά. Αυτό μπορεί να αντιμετωπιστεί.