Ζητήματα ταξινόμησης (συμπληρωματικά)

Ιωάννης Μαδεμλής

Ταξινομητές κανόνων

- Οι **ταξινομητές κανόνων** είναι ένας μηχανισμός παρεμφερής με τα δένδρα απόφασης, αλλά περισσότερο φορμαλιστικός.
- Αποτελούνται από συλλογές κανόνων οι οποίες κατασκευάζονται με βάση το σύνολο εκπαίδευσης.
  - Κάθε κανόνας έχει την παρακάτω μορφή:
    - «<αν ισχύει το κριτήριο *α* για την τιμή του γνωρίσματος *A* στο τρέχον πρότυπο>
    - ΚΑΙ
    - <αν ισχύει το κριτήριο *β* για την τιμή του γνωρίσματος *B* στο τρέχον πρότυπο>
    - ΚΑΙ...
    - ΤΟΤΕ <το τρέχον πρότυπο ανήκει στην κλάση *C*>».

Ταξινομητές κανόνων

ΣΥΝΘΗΚΕΣ

ΣΥΜΠΕΡΑΣΜΑ

Ταξινομητές κανόνων

- Αν κάποιο πρότυπο/στοιχείο ελέγχου ικανοποιεί όλες τις συνθήκες ενός κανόνα, τότε λέμε ότι **πυροδοτεί** τον εν λόγω κανόνα και του ανατίθεται η αντίστοιχη κλάση.

- Όπως και στην περίπτωση των δένδρων απόφασης, το σύνολο κανόνων δεν χρειάζεται να καλύπτει επακριβώς όλες τις περιπτώσεις του συνόλου εκπαίδευσης.
  - Αν τις καλύπτει καλείται **εξαντλητικό σύνολο κανόνων**.

- Άρα ορίζεται ένας **προεπιλεγμένος κανόνας**, ο οποίος ταξινομεί σε μία προεπιλεγμένη κλάση όσα πρότυπα δεν πυροδοτούν κανέναν από τους βασικούς κανόνες.

Ταξινομητές κανόνων

Ταξινομητές κανόνων

- Τι συμβαίνει αν πολλαπλοί κανόνες πυροδοτούνται ταυτοχρόνως από το ίδιο πρότυπο;
  - Π.χ., ενδέχεται να συμβεί αν οι κανόνες δεν λαμβάνουν καθόλου υπόψη κάποια γνωρίσματα.
- Δύο εναλλακτικές λύσεις:
  - Ορίζεται εκ των προτέρων μία **διάταξη προτίμησης** των κανόνων.
  - Διενεργείται **ψηφοφορία** και επιλέγεται η κλάση την οποία προτείνουν οι περισσότεροι από τους πυροδοτούμενους κανόνες για το τρέχον πρότυπο.

Ταξινομητές κανόνων

Ταξινομητές κανόνων

- Η αξιολόγηση ενός υπό κατασκευή κανόνα γίνεται με δύο μέτρα:
  - **Ακρίβεια.**
  - **Κάλυψη.**
- Ακρίβεια (accuracy) είναι το ποσοστό των ορθών ταξινομήσεων τις οποίες επιτυγχάνει ο κανόνας επί του ολικού πλήθους των προτύπων τα οποία τον πυροδοτούν.
- Κάλυψη (coverage) είναι το ποσοστό των στοιχείων του συνόλου εκπαίδευσης τα οποία πυροδοτούν τον κανόνα επί του ολικού πλήθους των προτύπων εκπαίδευσης.

Ταξινομητές κανόνων

Ταξινομητές κανόνων

Ταξινομητές κανόνων

- Η δημιουργία ενός ταξινομητή κανόνων μπορεί να γίνει *έμμεσα*.
  - Π.χ., συνάγοντας τους κανόνες από κάποιο ήδη κατασκευασμένο **δένδρο απόφασης**.
- Εναλλακτικά μπορεί να γίνει *άμεσα*, μέσω της μεθόδου της **ακολουθιακής κάλυψης**.
  - Εκτελείται ένας βρόχος, ξεχωριστά για κάθε δυνατή κλάση.
  - Σε κάθε επανάληψη του βρόχου εξάγεται ένας κανόνας.
  - Η σειρά με την οποία εκτελούνται οι βρόχοι των διαφορετικών κλάσεων ορίζεται από κάποιες παραμέτρους επιλεγμένες από εμάς.
    - Π.χ., τη σχετική συχνότητα εμφάνισης κάθε κλάσης στο σύνολο εκπαίδευσης.
  - Κάθε βρόχος εκτελείται μέχρι να ικανοποιηθεί κάποιο κριτήριο, επιλεγμένο από εμάς.

Ταξινομητές κανόνων

- Σε κάθε επανάληψη κάθε βρόχου αναζητείται ένας κανόνας ο οποίος να πυροδοτείται από:
  - Όσο το δυνατόν περισσότερα στοιχεία τα οποία ανήκουν στην τρέχουσα κλάση.
  - Όσο το δυνατόν λιγότερα στοιχεία που ανήκουν σε άλλες κλάσεις.
- Μόλις βρεθεί ένας κανόνας, αφαιρούνται από το σύνολο εκπαίδευσης τα στοιχεία που τον πυροδοτούν.
  - Ακολούθως, ο βρόχος μεταβαίνει στην επόμενη επανάληψη.
- Η εξαγωγή ενός κανόνα γίνεται με έναν από δύο τρόπους:
  - **Από το ειδικό στο γενικό**.
  - **Από το γενικό στο ειδικό**.

Ταξινομητές κανόνων

Ταξινομητές κανόνων

- Από το ειδικό στο γενικό:
  - Επιλέγεται τυχαία ένα πρότυπο της τρέχουσας κλάσης ως βάση του κανόνα.
  - Σταδιακά, αφαιρούνται από τον κανόνα συνθήκες ώστε αυτός να καλύπτει περισσότερα στοιχεία/πρότυπα με την ίδια ετικέτα κλάσης.
- Από το γενικό στο ειδικό:
  - Εκκινούμε με έναν κενό κανόνα, χωρίς καθόλου συνθήκες, ο οποίος απεικονίζει όλα τα δυνατά πρότυπα στην τρέχουσα κλάση.
  - Σταδιακά, προστίθενται στον κανόνα συνθήκες ώστε να καλύπτει όλο και λιγότερα στοιχεία.

Ταξινομητές κανόνων

Ταξινομητές κανόνων

- Και στις δύο περιπτώσεις απαιτείται ένα μέτρο αξιολόγησης του αποτελέσματος της προσθήκης/αφαίρεσης μίας συνθήκης στον κανόνα.
  - Έτσι ώστε σε κάθε βήμα ανάπτυξης του κανόνα να επιλέγεται η προσθήκη ή αφαίρεση εκείνης της συνθήκης, η οποία **αυξάνει περισσότερο την ακρίβεια ή/και την κάλυψη** του κανόνα.
  - Έτσι ώστε να μπορεί να ληφθεί μία απόφαση **τερματισμού της προσθήκης ή αφαίρεσης συνθηκών**.

Ταξινομητές κανόνων

- Ένα τέτοιο μέτρο είναι το **πληροφοριακό κέρδος FOIL**.

Ταξινομητές κανόνων

Ταξινομητές κανόνων

- Στηρίζεται στην έννοια του **μετρητή υποστήριξης** ενός κανόνα για την τρέχουσα κλάση.
  - Το πλήθος των προτύπων εκπαίδευσης με ετικέτα κλάσης η οποία αντιστοιχεί στην τρέχουσα κλάση (*θετικά πρότυπα*) και τα οποία ο κανόνας καλύπτει.
  - Με p_0 και p_1 συμβολίζεται ο μετρητής υποστήριξης πριν και μετά, αντιστοίχως, την προσθήκη/αφαίρεση της υπό εξέταση συνθήκης.
- Με n_0 και n_1 συμβολίζονται τα **συμπληρώματα του μετρητή υποστήριξης** πριν και μετά, αντιστοίχως, την προσθήκη/αφαίρεση της υπό εξέταση συνθήκης.
  - Συμπλήρωμα του μετρητή υποστήριξης είναι το πλήθος των *αρνητικών προτύπων* τα οποία καλύπτονται από τον κανόνα για την τρέχουσα κλάση.

Ταξινομητές κανόνων

- $FOIL = p_1(log_2 \frac{p_1}{p_1+n_1} - log_2 \frac{p_0}{p_0+n_0})$.

- Άρα, υψηλό FOIL συνεπάγεται πως η προσθήκη/αφαίρεση της επίμαχης συνθήκης οδηγεί σε υψηλή ακρίβεια, υψηλή υποστήριξη **και** υψηλή διαφορά από την προηγούμενη κατάσταση, **ταυτοχρόνως**.
  - Υψηλό $\frac{p_1}{p_1+n_1}$, υψηλό p_1 και χαμηλό $\frac{p_0}{p_0+n_0}$, αντιστοίχως.

Ταξινομητές κανόνων

Ταξινομητές κανόνων

Ζητήματα στους ταξινομητές κανόνων

- Οι ταξινομητές κανόνων είναι επιρρεπείς στην υπερεκπαίδευση.

- Λύση: το σύνολο κανόνων μπορεί να υποστεί επαναληπτικά κλάδεμα με οδηγό το σφάλμα επικύρωσης.
  - Μικρότερου μεγέθους σύνολο κανόνων συνεπάγεται μικρότερη τάση υπερεκπαίδευσης.
  - Λογική παρόμοια με των δένδρων απόφασης.

Ταξινομητές κανόνων

Ζητήματα στους ταξινομητές κανόνων

- Οι ταξινομητές κανόνων έχουν τις ιδιότητες και την περιγραφική ικανότητα των δένδρων απόφασης.
  - Οι αποφάσεις τους είναι ευκόλως ερμηνεύσιμες.
- Όμως μπορούμε να ενεργοποιήσουμε τη δυνατότητα ταυτόχρονης πυροδότησης πολλαπλών κανόνων για το ίδιο πρότυπο ελέγχου.
  - Έτσι, οι ταξινομητές κανόνων είναι σε θέση να ορίζουν περισσότερο πολύπλοκες περιοχές απόφασης από απλές ορθογώνιες.

Ζητήματα στους ταξινομητές κανόνων

- Αν οριστεί διάταξη προτεραιότητας στις υποψήφιες κλάσεις, οι ταξινομητές κανόνων μπορούν να χειριστούν πολύ εύκολα **ανισοκατανεμημένα σύνολα εκπαίδευσης**.
  - Ανισοκατανομή έχουμε όταν το πλήθος των προτύπων κάποιων κλάσεων είναι πολύ μεγαλύτερο από το πλήθος των στοιχείων άλλων κατηγοριών.

Ομαδική ταξινόμηση

- Έστω μία ομάδα n ανεξάρτητων ταξινομητών, εκπαιδευμένων για το ίδιο πρόβλημα.
- Χαρακτηρίζονται από διαφορετικό σφάλμα γενίκευσης.
- Έχει δειχθεί πως η **συνάθροιση** των προβλέψεών τους για ένα πρότυπο ελέγχου είναι ακριβέστερη από την πρόβλεψη οποιουδήποτε μεμονωμένου ταξινομητή.
  - Η συνάθροιση μπορεί να γίνει απλώς μέσω **ψηφοφορίας**.
  - Προϋπόθεση:το σφάλμα γενίκευσης του καθενός να μην ξεπερνά το 50%.

Ομαδική ταξινόμηση

Ομαδική ταξινόμηση

- Σε κάθε πρότυπο ελέγχου ανατίθεται τελικά η κλάση την οποία προβλέπουν οι περισσότεροι από τους n μεμονωμένους ταξινομητές.
- Η μεθοδολογία αυτή ονομάζεται **ομαδική ταξινόμηση** (ensemble classification).
- Πώς όμως οι επιμέρους ταξινομητές έχουν διαφορετικό σφάλμα γενίκευσης;
  - Ταξινομητές διαφορετικού **τύπου**.
  - Εκπαιδευμένοι ο καθένας σε διαφορετικό **υποσύνολο του ολικού συνόλου εκπαίδευσης**, ή
  - Εκπαιδευμένοι ο καθένας σε διαφορετικό **υποσύνολο των γνωρισμάτων** των προτύπων εκπαίδευσης.

Ομαδική ταξινόμηση

Ομαδική ταξινόμηση

Bagging και Boosting

- Μία δημοφιλής μέθοδος είναι το **bagging**.
  - Για την κατασκευή κάθε στοιχειώδους/επιμέρους ταξινομητή χρησιμοποιείται σύνολο εκπαίδευσης ίσου μεγέθους με το ολικό σύνολο εκπαίδευσης.
  - Έχει όμως προκύψει από ομοιόμορφη **δειγματοληψία με επανατοποθέτηση** στο αρχικό σύνολο εκπαίδευσης.
    - Επομένως το ίδιο αρχικό πρότυπο μπορεί να εμφανίζεται 0, 1, 2 ή πολλαπλές φορές.
- Παρεμφερής λύση είναι το **boosting**.
  - Η δειγματοληψία προτύπων από το ολικό σύνολο εκπαίδευσης, ώστε να σχηματιστεί το επιμέρους σύνολο εκπαίδευσης κάθε μεμονωμένου ταξινομητή, γίνεται **βεβαρυμμένα**.
  - Σε κάθε πρότυπο ανατίθεται βάρος ώστε η πιθανότητα επιλογής του να είναι *ανάλογη της δυσκολίας στην ταξινόμησή του*.

Ομαδική ταξινόμηση

Αλγόριθμος Boosting

- Αρχικώς, όλα τα βάρη στο boosting είναι ίσα.
- Εκτελείται επαναληπτικά δειγματοληψία με επανατοποθέτηση από το ολικό σύνολο εκπαίδευσης, η οποία λαμβάνει υπόψη τα βάρη των προτύπων.
- Σε κάθε επανάληψη εκπαιδεύεται κάποιος ταξινομητής με βάση το τελευταίο δείγμα.
  - Όσα πρότυπα του επιμέρους συνόλου εκπαίδευσής του δεν ταξινομούνται σωστά από τον ίδιο, αυξάνουν το βάρος τους.
- Στην επόμενη επανάληψη δημιουργείται νέος ταξινομητής, με νέα δειγματοληψία με βάση τα ενημερωμένα βάρη.
- Και ούτω καθεξής…

Ομαδική ταξινόμηση

Αλγόριθμος Boosting

- Εν τέλει, οι n ταξινομητές που προέκυψαν από τις n επαναλήψεις συναποτελούν την τελική ομάδα.

- Κατά τη συναγωγή της τελικής **πρόβλεψης για ένα πρότυπο ελέγχου**, η ψηφοφορία μπορεί να είναι **βεβαρυμμένη**.
  - Στην απόφαση κάθε στοιχειώδους ταξινομητή ανατίθεται ένα *βάρος ανάλογο της ακρίβειάς* του.

- Μία ομάδα ταξινομητών αποτελούμενη από n επιμέρους δένδρα απόφασης καλείται **τυχαίο δάσος**.
  - Είναι όμως εφαρμογή του bagging και όχι του boosting.

Ανισοκατανεμημένα δεδομένα

Ανισοκατανομή Κλάσεων

- Σε προβλήματα ταξινόμησης με σημαντική ανισοκατανομή μεταξύ των κλάσεων δεν είναι πάντα βέλτιστη η χρήση της **ακρίβειας** (accuracy) ως μετρικής αξιολόγησης.
  - Η ακρίβεια συμπεριφέρεται σε όλες τις κλάσεις ισότιμα.

- Συνήθης περίπτωση είναι η ανισοκατανομή σε προβλήματα **δυαδικής ταξινόμησης**.
  - Π.χ., σε προβλήματα επιβλεπόμενης ανίχνευσης ανωμαλιών τα *θετικά πρότυπα* (ανώμαλα) είναι εξ ορισμού ελάχιστα σε σχέση με τα *αρνητικά* (κανονικά).
  - Άρα έχουμε μία *σπάνια κλάση* και μία *συνήθη κλάση*.
  - Στόχος είναι η κατασκευή ταξινομητών οι οποίοι εντοπίζουν ορθώς τα πρότυπα ελέγχου της σπάνιας κλάσης (θετικά).

Ανισοκατανομή Κλάσεων

Ανισοκατανεμημένα δεδομένα

- Σε αυτές τις περιπτώσεις χρησιμοποιείται η εξής ορολογία:
  - **Αληθώς θετικά** (True Positive, *TP*): το πλήθος των ορθώς ταξινομούμενων θετικών προτύπων.
  - **Ψευδώς αρνητικά** (False Negative, *FN*): το πλήθος των θετικών προτύπων τα οποία εσφαλμένα ταξινομούνται ως αρνητικά.
  - **Ψευδώς θετικά** (False Positive, *TP*): το πλήθος των αρνητικών προτύπων τα οποία εσφαλμένα ταξινομούνται ως θετικά.
  - **Αληθώς αρνητικά** (True Negative, *TN*): το πλήθος των ορθώς ταξινομούμενων αρνητικών προτύπων.

Ανισοκατανομή Κλάσεων

Ανισοκατανεμημένα δεδομένα

- Με βάση τα προηγούμενα, ορίζονται τα παρακάτω:
  - **Ρυθμός αληθώς θετικών** (True Positive Rate, *TPR*), ή **ευαισθησία** (sensitivity), ή **ανάκληση** (recall): το ποσοστό μεταξύ των πράγματι θετικών προτύπων τα οποία ταξινομούνται ορθώς από το μοντέλο.
    - $TPR = \frac{TP}{TP+FN}$.
  - **Ρυθμός αληθώς αρνητικών** (True Negative Rate, *TNR*), ή **εξειδίκευση** (specificity): το ποσοστό των πράγματι αρνητικών προτύπων τα οποία ταξινομούνται ορθώς από το μοντέλο.
    - $TNR = \frac{TN}{TN+FP}$.
  - **Ρυθμός ψευδώς θετικών** (False Positive Rate, *FPR*): το ποσοστό των αρνητικών προτύπων τα οποία ταξινομούνται εσφαλμένα από το μοντέλο.
    - $FPR = \frac{FP}{TN+FP}$.

Ανισοκατανομή Κλάσεων

Ανισοκατανεμημένα δεδομένα

- Με βάση τα προηγούμενα, ορίζονται τα παρακάτω:
  - **Ρυθμός ψευδώς αρνητικών** (False Negative Rate, *FNR*): το ποσοστό των θετικών προτύπων τα οποία ταξινομούνται εσφαλμένα από το μοντέλο.
    - $FNR = \frac{FN}{TP+FN}$.
  - **Ορθότητα** (Precision, p): το ποσοστό των πράγματι θετικών προτύπων μεταξύ όσων προβλέπονται ως θετικά.
    - $p = \frac{TP}{TP+FP}$.
    - Όσο μεγαλύτερη η ορθότητα, τόσο λιγότερα το ψευδώς θετικά.
- Η πρόκληση στα εν λόγω προβλήματα είναι η κατασκευή ταξινομητών με **υψηλή ανάκληση** και **υψηλή ορθότητα** *ταυτοχρόνως*.
  - Π.χ., ένας τετριμμένος ταξινομητής ο οποίος προβλέπει όλα τα πρότυπα ως θετικά έχει τέλεια ανάκληση, αλλά μικρή ορθότητα.

Ανισοκατανεμημένα δεδομένα

- Έτσι, ως μέτρο επίδοσης συνήθως χρησιμοποιείται ο **αρμονικός μέσος της ανάκλησης και της ορθότητας**.
  - Καλείται **μέτρο F_1** (F_1-measure, ή F-measure).
  - $F_1 = \frac{2*TP}{2*TP+FP+FN}$.
  - Κατ' αντιδιαστολή, η συνήθης ακρίβεια ισούται με:
    - $acc = \frac{TP+TN}{TP+TN+FP+FN}$.

Ανισοκατανομή Κλάσεων

- Ο αρμονικός μέσος δύο τιμών τείνει να είναι εγγύτερα στη μικρότερη από τις δύο.
  - Έτσι, ένα υψηλό F_1 συνεπάγεται πως είναι υψηλή και η ανάκληση και η ορθότητα.

Ανισοκατανεμημένα δεδομένα

- Τι γίνεται όμως την εκπαίδευση σε ανισοκατανεμημένα δεδομένα δύο κλάσεων;

- Απλές λύσεις: τροποποίηση του συνόλου εκπαίδευσης μέσω κατάλληλης δειγματοληψίας.
  - Π.χ., μπορεί να γίνει υπερδειγματοληψία της σπάνιας κλάσης, υποδειγματοληψία της συνήθους ή και τα δύο.
  - Έτσι, το τελικό σύνολο δεδομένων όπου εκπαιδεύεται ο ταξινομητής παρουσιάζει ομοιόμορφη κατανομή δεδομένων σε κλάσεις.

Ανισοκατανομή Κλάσεων

Thank you for your attention!

Q & A

Contact: imademlis@aueb.gr