Τεχνικές αξιολόγησης αλγορίθμων ταξινόμησης

Νησιώτη Ελένη
21 Ιανουαρίου 2017

1 Σκοπός και έννοιες αξιολόγησης

Στον τομέα της μηχανικής μάθησης η έννοια της αξιολόγησης σχετίζεται κυρίως με δύο προσπάθειες:

- εξακρίβωση της δυνατότητας ενός μεμονωμένου μοντέλου να ”γενικεύει”, δηλαδή η απόδοση επιτευχθείσα στο σετ εκπαίδευσης να βρίσκεται κοντά στην απόδοση σε άγνωστα δεδομένα
- επιλογή του βέλτιστου μοντέλου μεταξύ διαθέσιμων δεδομένης της απόδοσής τους σε συγκεκριμένο πλήθος σετ δεδομένων

Η κοινότητα της μηχανικής μάθησης καταφεύγει σε στατιστικά εργαλεία για την εξαγωγή έγκυρων, γενικεύσιμων και ανατάξιμων συμπερασμάτων βάσει πειραμάτων. Χρησιμοποιούνται διάφορες τεχνικές, καθώς η καταλληλότητά και εφαρμοσιμότητά τους εξαρτάται από τη φύση του προβλήματος.

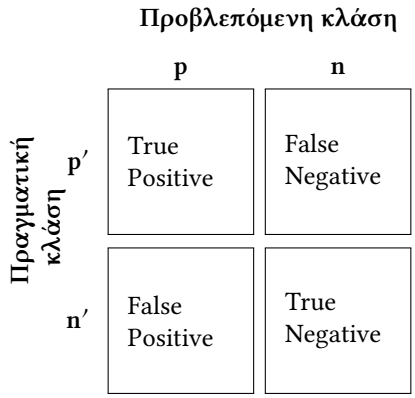
2 Τεχνικές αξιολόγησης

2.1 Μετρικές αξιολόγησης

Η πρώτη απόφαση προς την αξιολόγηση είτε ενός είτε πολλών μοντέλων είναι το κριτήριο αξιολόγησης, δηλαδή η μετρική που αποσκοπεί να βελτιστοποιήσει το πείραμα. Η επιλογή αρχικά περιορίζεται από τους στόχους του ερευνητή και του πειράματος (χρόνος, υπολογιστική πολυπλοκότητα, ακρίβεια μοντέλου). Συχνότερος

στόχος είναι η παραγωγή του ακριβέστερου μοντέλου, οπότε ο ερευνητής επιλέγει μεταξύ:

Μετρικές άμεσα προερχόμενες από τον πίνακα Σύγχυσης Ο πίνακας σύγχυσης για ένα μοντέλο ταξινόμησης συνοψίζει την πληροφορία σχετικά με τις προβλέψεις του και την πραγματική κλάση των παραδειγμάτων



Ακρίβεια (*accuracy*)	$\frac{TP+TN}{TP+TN+FP+FN}$
Ανάκληση(*recall*)	$\frac{TP}{TP+FN}$
Ακρίβεια (*precision*)	$\frac{TP}{TP+FN}$
F-μετρική(*f-measure*)	$\frac{2*(precision+recall)}{precision+recall}$
Matthew συντελεστής συσχέτισης(*Matthew's correlation coefficient*)	$\frac{TP*TN-FP*FN}{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}$

Area under Curve Η μετρική αυτή προέρχεται από τη καμπύλη ROC ενός ταξινομητή και αντιστοιχεί στο εμβαδό της περιοχής κάτω από αυτήν. Με βάση την τιμή της μπορεί να διαπιστωθεί ποιοτικά η λειτουργία του ταξινομητή:

- 0.9 - 1.0 → τέλειος ταξινομητής
- 0.8 - 0.1 → καλός ταξινομητής
- 0.7 - 0.8 → μέτριος ταξινομητής
- 0.6 - 0.7 → κακός ταξινομητής
- 0.5 - 0.6 → τυχαίος ταξινομητής

2.2 Στατιστικά tests

Τα πειράματα ταξινόμησης απαιτούν τη στατιστική ανάλυση πληθυσμών για την εξαγωγή συμπερασμάτων. Αν θεωρήσουμε ένα σύνολο από σετ δεδομένων δυαδικής ταξινόμησης, μερικά ερωτήματα που μπορούν να προκύψουν: υπάρχει κάποια εξάρτηση μεταξύ της κλάσης και κάποιου κατηγορικού χαρακτηριστικού για ένα συγκεκριμένο σετ δεδομένων; Ποιος αλγόριθμος είναι γενικά καλύτερος;

Τα στατιστικά τεστ εφαρμόζονται σε πίνακες ενδεχομένων (*contigency tables*) και έχουν ως στόχο της απόρριψη ή μη της μηδενικής υπόθεσης, η οποία αντιστοιχεί σε ανεξαρτησία των δεδομένων και τυχαιότητα των διαφορών που παρουσιάζονται μεταξύ διαφορετικών πληθυσμών. Οι πίνακες αυτοι είναι 1-way, 2-way ή 3-way, ενώ για την εμπλοκή περισσότερων πληθυσμών οι ερευνητές καταφεύγουν σε γενικευμένα γραμμικά μοντέλα. Presnell [3]

Μερικές έννοιες που σχετίζονται με τα στατιστικά τεστ είναι:

- υπόθεση. Προτείνεται από τον ερευνητή και χαρακτηρίζει τη στατιστική σχέση μεταξύ των δύο πληθυσμών υπό σύγκριση. Συγκρίνεται ως η εναλλακτική μιας ιδανικής μηδενικής υπόθεσης, η οποία αποκλείει οποιαδήποτε σχέση μεταξύ των δύο δειγμάτων. Στόχος του πειράματος είναι η απόρριψη της μηδενικής υπόθεσης, ενώ σε περίπτωση αποτυχίας το συμπέρασμα είναι η αδυναμία απόρριψης της μηδενικής υπόθεσης και όχι η επιβεβαίωση της εναλλακτικής.
- στατιστική σημασία. Το πείραμα έχει στατιστική σημασία, όταν η σχέση μεταξύ των δειγμάτων είναι απίθανο να προκύψει από τη μηδενική υπόθεση με βάση ένα κατώφλι πιθανότητας α.
- διάστημα εμπιστοσύνης (*confidence interval*). Προκύπτει από το κατώφλι πιθανότητας ως $1 - \alpha$ και ερμηνεύεται ως εξής: Αν έχουμε 95% διάστημα εμπιστοσύνης τότε είμαστε κατά ίση πιθανότητα σίγουροι ότι η μέση τιμή του πληθυσμού (και όχι των δειγμάτων) θα κινείται σε συγκεκριμένα πλαίσια (που προκύπτουν από την κατανομή του test statistic).
- p-value. Είναι η τιμή που οδηγεί στο αποτέλεσμα του στατιστικού πειράματος. Αποτελεί απόδειξη κατά της μηδενική υπόθεσης και όσο χαμηλότερη είναι η τιμή του τόσο ισχυρότερη η απόρριψή της. Για δεδομένο κατώφλι πιθανότητας αρκεί να είναι μικρότερο από αυτό.
- σφάλματα τύπου I/II. Η απόρριψη μιας έγκυρης μηδενικής υπόθεσης χαρακτηρίζεται ως σφάλμα τύπου I, ενώ η αδυναμία απόρριψης μιας άκυρης σφάλμα τύπου II.
- ισχύς. Πρόκειται για τη πιθανότητα το τεστ να απορρίψει μια λανθασμένη μηδενική υπόθεση.

2.2.1 Pearson's chi-squared test

Πρόκειται για ένα στατιστικό τεστ μεταξύ δύο συνόλων κατηγορικών δεδομένων που εξετάζει αν οι διαφορές τους προκλήθηκαν τυχαία. Είναι κατάλληλο για unpaired δεδομένα από μεγάλα δείματα. Προέρχεται από την ευρύτερη οικογένεια των τεστς που αξιολογούνται με αναφορά στην κατανομή chi-squared, για την οποία όταν η μηδενική υπόθεση είναι αληθής η κατανομή του test statistic είναι chi-squared.

Yate's correction for continuity Η τεχνική αυτή χρησιμοποιείται για διόρθωση του εξής προβλήματος: κατά την εφαρμογή του Pearson's chi-square τεστ γίνεται η υπόθεση πως η διακριτή πιθανότητα των παρατηρούμενων συχνοτήτων στον πίνακα ενδεχομένων μπορεί να προσεγγιστεί από μία συνεχή chi-squared κατανομή.

2.2.2 Friedman test

Πρόκειται για ένα μη-παραμετρικό τεστ για την ανίχνευση διαφορών μεταξύ πολλών αλγορίθμων σε πολλά σετ δεδομένων. Θεωρείται μια μη-παραμετρική εκδοχή της ANOVA, με απόρροια την απεμπλοκή από τις υποθέσεις της κανονικής κατανομής και των ίσων διακυμάνσεων των residuals και την απώλεια ισχύος.

Σημαντική προσθήκη αποτελεί η εναλλακτική test statistic που εισήγαγαν οι [], καθώς διαπίστωσαν ότι η βασική ήταν ανεπιθύμητα συντηρητική.

Σε περίπτωση διαπίστωσης σημαντικής στατιστικής διαφοράς στην απόδοση πολλών αλγορίθμων προκύπτει η ανάγκη εξακρίβωσης των ζευγαριών που οδήγησαν σε αυτό το αποτέλεσμα. Προς αυτό το σκοπό εφαρμόζονται τα εξής δύο post-hoc τεστ:

Nemey Χρησιμοποιείται κατά τη σύγκριση όλων των αλγορίθμων μεταξύ τους.

Boferroni Χρησιμοποιείται κατά τη σύγκριση όλων των αλγορίθμων με έναν αλγόριθμο -αναφορά. Αυτή η τεχνική είναι η συχνότερη, καθώς συνηθίζεται να προτείνεται ένας αλγόριθμος ως βελτίωση της τρέχουσας έρευνας.

2.2.3 ANOVA

Το τεστ αυτό αναλύει τη διακύμανση των δεδομένων και απορρίπτει τη μηδενική υπόθεση αν η διακύμανση μεταξύ των ταξινομητών είναι σημαντικά μεγαλύτερη από τη διακύμανση σφάλματος. Βασικές προϋποθέσεις της είναι η κανονικότητα των δειγμάτων και η σφαιρικότητα (οι τυχαίες μεταβλητές έχουν ίση διακύμανση). Απαιτούνται επίσης post-hoc τεστ για την εξακρίβωση σημαντικά διαφορετικών ζευγών αλγορίθμων.

Tukey test

Dunnett

2.2.4 Fisher's exact test

Λέγεται ακριβές επειδή για μικρά δείγματα η σημασία της διακύμανσης από τη μηδενική υπόθεση (p-value) μπορεί να υπολογιστεί ακριβώς αντί να βασίζεται σε μια προσέγγιση που γίνεται ακριβής καθώς το μέγεθος του δείγματος πλησιάζει το άπειρο.

2.2.5 Wilcoxon rank test (Mann-Whitney)

2.2.6 Cochran-Mantel-Haenszel

Χρησιμοποιείται για την εξέταση της σχέσης μεταξύ ενός δυαδικού προβλέπτη και της κλάσης λαμβάνοντας υπόψην του stratification κατά τη συλλογή των δεδομένων. Είναι γενίκευση του τεστ McNemar.

2.2.7 McNemar

2.3 Διαγράμματα

Roc καμπύλη

Διαγράμματα performance profile [2]

Graphs of post-hoc tests

3 Αξιολόγηση στη πράξη - Ευριστικές

[1]

3.1 πολλά μοντέλα - 1 σετ δεδομένων

- το accuracy δεν είναι κατάλληλο σε μη-σταθμισμένα σετ δεδομένων
- το f-measure δεν είναι κατάλληλη μετρική για το συνδυασμό recall και accuracy
- το recall είναι η κατάλληλη μετρική σε προβλήματα που υπάρχει κλάση ενδιαφέροντος
- μια γενικά αποδεκτή μετρική είναι το AUC

- η τεχνική του 5-2 crossvalidation t-test είναι προτιμότερη από το paired t-test σε k-fold cross validation, καθώς το δεύτερο ενέχει το πρόβλημα της υποτίμησης της διακύμανσης και αυξημένου σφάλματος τύπου I. Το McNnemar τεστ είναι εξίσου ισχυρό στις περιπτώσεις που δεν ενδείκνυται η εφαρμογή του αλγορίθμου πολλαπλές φορές.

3.2 2 μοντέλα - πολλά σετ δεδομένων

Το πλεονέκτημα της σύγκρισης μοντέλων σε πολλά σετ δεδομένων έναντι ενός είναι πως η πηγή της διακύμανσης εντοπίζεται σε διαφορές της απόδοσης σε ανεξάρτητα σετ δεδομένων και όχι σε τυχαία δείγματα ενός σετ. Έτσι, αποφεύγεται το πρόβλημα της πολωμένης εκτίμησης της διακύμανσης και διάφορες μορφές cross-validation είναι επιτρεπτές.

- το t-τεστ θεωρείται ακατάλληλο για τους εξής λόγους: έχει νόημα μόνο όταν οι διαφορές στην απόδοση είναι συγκρίσιμες (ενώ στην πραγματικότητα εξαρτώνται από τη φύση του σετ δεδομένων), οι διαφορές πρέπει να έχουν κανονική κατανομή εκτός αν το δείγμα είναι αρκετά μεγάλο (γεγονός που δεν επαληθεύεται σε συνήθη πειράματα με ≈ 30 σετ δεδομένων) και τέλος, είναι ευεπηρέαστο σε ακραίες τιμές. Προτείνεται μόνο σε περιπτώσεις που θεωρούμε ότι έχουμε αρκετά σετ δεδομένων και η απόδοση ακολουθεί κανονική κατανομή.
- καταλληλότερο τεστ είναι το Wilcoxon signed rank τεστ. Απαιτεί ποιοτική και όχι ποσοτική συγκρισιμότητα (commensurability) των διαφορών, δεν προϋποθέτει κανονική κατανομή, επηρεάζεται λιγότερο από εξωκείμενες τιμές. Καθώς χρησιμοποιεί συνεχείς διαφορές η μείωση της ακρίβειας (με χρήση λιγότερων δεκαδικών) οδηγεί σε εξασθένησή του.
- η τεχική του Yate's δεν προτιμάται, καθώς καθιστά το τεστ ιδιαίτερα συντηρητικό.
- το chi-squared τεστ είναι μια καλή προσέγγιση για το τεστ Fisher όταν η κατανομή του test-statistic είναι σχεδόν ίση με την κατανομή chi-squared. Στις περιπτώσεις μικρών δειγμάτων ή άνισα κατανεμημένων δεδομένων στον πίνακα ενδεχομένων αυτό δεν ισχύει. Μία ευριστική επιβεβαιώσης της μη καταλληλότητας της κατανομής chi-squared είναι οι τιμές στα κελιά του πίνακα επιβεβαιώσης να είναι κάτω από 5 ή 10 για ένα βαθμό ελευθερίας (ο κανόνας αυτός έχει αποδειχθεί υπερσυντηρητικός). [4]

3.3 πολλά μοντέλα - πολλά σετ δεδομένων

Η σύγκριση πολλών μοντέλων απαιτεί εξειδικευμένα τεστ, σε αντίθεση με τη συνηθισμένη τεχνική της επέκτασης τεστ σχεδιασμένων για δύο μοντέλα, η οποία κρίνεται ακατάλληλη.

Η παρουσίαση των αποτελεσμάτων σε μορφή πίνακα είναι μη επαρκής, η εξαγωγή του μέσου όρου αποτελεί υπεραπλούστευση, η μέτρηση στατιστικά σημαντικών νικών και ηττών αναξιόπιστη, το sign-test είναι αδύναμο. Η εφαρμογή paired t-test είναι ακατάλληλη για εξαγωγή συμπερασμάτων μεταξύ πολλών σετ, καθώς ένα ποσοστό των μηδενικών υποθέσεων απορρίπτεται τυχαία.

- η τεχνική ANOVA παρουσιάζει προβλήματα, καθώς προϋποθέτει κανονική κατανομή δειγμάτων και σφαιρικότητα.
- προτείνεται το τεστ Friedman με post-hoc το τεστ Holm's (αν και είναι ελάχιστα λιγότερο ισχυρό από το τεστ Hommel προσφέρει τα πλεονεκτήματα της απλότητας και ευκολότερου υπολογισμού).

3.4 Συμπεράσματα

Τα παραμετρικά τεστ είναι πιο ισχυρά από τα μη-παραμετρικά όταν ισχύουν οι προϋποθέσεις τους. Καθώς στη πλειοψηφία των περιπτώσεων πειραμάτων μηχανικής μάθησης οι προϋποθέσεις δεν ικανοποιούνται (ή τουλάχιστον δεν έχουν λόγο να ικανοποιούνται) προτεινόμενα είναι τα μη-παραμετρικά.

Βιβλιογραφία

[1] Janez Demšar. "Statistical Comparisons of Classifiers over Multiple Data Sets". In: *J. Mach. Learn. Res.* 7 (Dec. 2006), pp. 1–30. ISSN: 1532-4435.

[2] Elizabeth D. Dolan and Jorge J. Moré. "Benchmarking optimization software with performance profiles". In: *Mathematical Programming* 91.2 (2002), pp. 201–213. ISSN: 1436-4646. DOI: 10.1007/s101070100263.

[3] Brett Presnell. "An introduction to Categorical Data Analysis Using R". In: ().

[4] Wikipedia. *Fisher's exact test*. 1999. URL: https://en.wikipedia.org/wiki/Fisher's_exact_test (visited on 01/21/2017).