ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΑΡΟΥΣΙΑΣΗ / ΕΞΕΤΑΣΗ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Γρεασίδου Ελισσάβετ

Μεταπτυχιακή Φοιτήτρια

Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

Επόπτης Μεταπτ. Εργασίας: Αναπλ. Καθηγητής, Ι. Τσαμαρδίνος

Τετάρτη, 8/2/2017, 12:00

Αίθουσα B108, Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

"Διόρθωση της μεροληψίας της εκτίμησης της απόδοσης του πρωτοκόλλου της διασταυρωμένης επικύρωσης και επιτάχυνση της εκτέλεσης του "

ΠΕΡΙΛΗΨΗ

Η μέθοδος διασταυρωμένη επικύρωση (Cross Validation - CV) αποτελεί ένα ντεφάκτο πρότυπο στον τομέα της εφαρμοσμένης στατιστικής και εποπτευόμενης μηχανικής μάθησης (supervised machine learning) τόσο για την επιλογή ενός μοντέλου αλλά και την αξιολόγηση του. Η διαδικασία αυτή εφαρμόζεται σε ένα σύνολο υποψήφιων διαμορφώσεων (configurations) (δηλαδή, ένα σύνολο ακολουθιών βημάτων μοντελοποίησης με καθορισμένους αλγορίθμους και τιμές για τις ύπερ-παραμέτρους τους για κάθε βήμα) και εκείνη με την καλύτερη απόδοση, σύμφωνα με ένα προκαθορισμένο κριτήριο, επιλέγεται. Ωστόσο, η "καλύτερη απόδοση που επιτυγχάνεται κατά τη διαδικασία του CV είναι γνωστό ότι είναι μία αισιόδοξα μεροληπτική (biased) εκτίμηση της γενίκευσης της απόδοσης του τελικού μοντέλου. Μέχρι σήμερα, ένα σχετικά περιορισμένο μέρος της έρευνας έχει αφιερωθεί στη διόρθωση αυτής της μεροληψίας (bias), και όλες οι προτεινόμενες μέθοδοι είτε έχουν την τάση να την διορθώνουν περισσότερο από όσο χρειάζεται ή έχουν περιορισμούς που μπορούν να κάνουν τη χρήση τους ανέφικτη.

Σε αυτή την εργασία, προτείνουμε μια μέθοδο διόρθωσης της μεροληψίας βασισμένη στην μέθοδο του bootstrap (Bootstrap-biased Bias Correction method - BBC) η οποία λειτουργεί ανεξάρτητα από την εργασία ανάλυσης δεδομένων (π.χ. ταξινόμηση, παλινδρόμηση), ή τη δομή των μοντέλων που εμπλέκονται, και απαιτεί μόνο μια μικρή υπολογιστική επιβάρυνση σε σχέση με τη βασική διαδικασία του CV. Η BBC μέθοδος διορθώνει την μεροληψία με συντηρητικό τρόπο παρέχοντας μια σχεδόν αμερόληπτη εκτίμηση της απόδοσης. Η βασική ιδέα είναι να εφαρμοστεί η μέθοδος του bootstrap σε όλη τη διαδικασία της επιλογής της καλύτερης μεθόδου στις εκτός εκπαιδευμένου δείγματος (out-of-sample) προβλέψεις της κάθε διαμόρφωσης (configuration), χωρίς πρόσθετη εκπαίδευση μοντέλων. Σε σύγκριση με τις εναλλακτικές μεθόδους, δηλαδή την εμφωλευμένη διασταυρωμένη επικύρωση (Nested Cross Validation - NCV), και την μέθοδο των Tibshirani και Tibshirani (TT), η BBC μέθοδος είναι υπολογιστικά πιο αποδοτική, είναι εφαρμόσιμη σε οποιαδήποτε διαδικασία CV, και η εκτίμηση της απόδοσης που παρέχει είναι ανταγωνιστική σε σχέση με εκείνη του NCV. Επίσης, χρησιμοποιούμε την ιδέα της εφαρμογής της bootstrap μεθόδου στις εκτός εκπαιδευμένου δείγματος (out-of-sample) προβλέψεις για την επιτάχυνση του χρόνου εκτέλεσης της CV διαδικασίας. Συγκεκριμένα, χρησιμοποιώντας ένα στατιστικό έλεγχο υποθέσεων (hypothesis test) βασισμένο στη μέθοδο του bootstrap, σταματάμε την εκπαίδευση μοντέλων σε καινούργια υποσύνολα των δεδομένων (folds) για στατιστικά-σημαντικά (statistically-signicantly) υποδεέστερες διαμορφώσεις (configurations). Η μέθοδος Bootstrap-based Early Dropping (BED) μειώνει σημαντικά τον υπολογιστικό χρόνο του CV με αμελητέα ή καμία επίδραση στην απόδοση. Οι δύο μέθοδοι μπορούν να συνδυαστούν οδηγώντας στην BED-BBC μέθοδο η οποία είναι αποδοτική και παρέχει ακριβείς εκτιμήσεις της απόδοσης.

Greasidou Elisavet

M.Sc. Thesis

Computer Science Department

University of Crete

Master's Thesis Supervisor: Associate Professor I. Tsamardinos

Wednesday, 8/2/2017, 16:00
Room B108, Computer Science dept., University of Crete

"Bias Correction of the Cross-Validation Performance Estimate and Speed Up of its Execution Time"

ABSTRACT

Cross Validation (CV) is a de-facto standard in applied statistics and supervised machine learning both for model selection and assessment. The procedure is applied on a set of candidate configurations (i.e. a set of sequences of modelling steps with specified algorithms and their hyper-parameter values for each step) for model production, and the one with the best performance, according to a pre-specified criterion, is selected. However, the "best" performance achieved during CV is known to be an optimistically biased estimation of the generalization performance of the final model. To date, a relatively limited amount of research has been devoted to the correction of this bias, and all proposed methods either tend to over-correct or have limitations which can make their use impractical.

In this thesis, we propose a Bootstrap-based Bias Correction method (BBC) which works regardless of the data analysis task (e.g. classification, regression), or the structure of the models involved, and requires only a small computational overhead with respect to the basic CV procedure. BBC corrects the bias in a conservative way, providing an almost unbiased estimate of performance. Its main idea is to bootstrap the whole process of selecting the best-performing configuration on the out-of-sample predictions of each configuration, without additional training of models. In comparison to the alternatives, namely the Nested Cross Validation (NCV), and a method by Tibshirani and Tibshirani (TT), BBC is computationally more efficient, yields performance estimates competitive to those of NCV and is applicable to any CV procedure. Subsequently, we also employ the idea of bootstrapping the out-of-sample predictions in order to speed up the execution time of the CV procedure. Specifically, using a bootstrap-based hypothesis test we stop training of models on new folds of statistically-significantly inferior configurations. The Bootstrap-based Early Dropping (BED) method significantly reduces the computational time of CV with a negligible or no effect on performance. The two methods can be combined leading to the BED-BBC procedure that is both efficient and provides accurate estimates of performance.