Βελτιστοποίηση υπερπαραμέτρων αλγορίθμων μηχανικής μάθησης

Νησιώτη Ελένη
21 Δεκεμβρίου 2016

1 Περιγραφή του προβλήματος

Ένα βασικό στάδιο κατά την εκπαίδευση αλγορίθμων μηχανικής μάθησης είναι αυτό της επιλογής των υπερπαραμέτρων του μοντέλου.

Μαθηματικά το πρόβλημα μπορεί να διατυπωθεί ως εξής: σκοπός ενός πειράματος μηχανικής μάθησης είναι η εκπαίδευση ενός μοντέλου M, το οποίο ελαχιστοποιεί μία προκαθορισμένη συνάρτηση κόστους $L(X^{(te)}; M)$ σε ένα δεδομένο σετ δεδομένων. Το μοντέλο κατασκευάζεται από έναν αλγόριθμο μάθησης , ο οποίος παραμετροποιείται από ένα σύνολο παραμέτρων λ.

Καταληγούμε λοιπόν στον μαθηματικό ορισμό της εύρεσης του συνόλου των υπερπαραμέτρων λ^*, που ορίζουν το βέλτιστο μοντέλο M^*

$$\lambda^* = \arg\min L(X^{(te)}; A(X^{(tr)}; \lambda)) = \arg\min_{\lambda} L(\lambda; A, X^{(tr)}, L) \tag{1.1}$$

όπου $X^{(tr)}$ το σετ δεδομένων και $X^{(te)}$ το σετ ελέγχου

2 Τεχνικές βελτιστοποίησης

Μερικά χαρακτηριστικά της παραπάνω συνάρτησης είναι τα εξής:

- είναι μια συνάρτηση μαύρου κουτιού, δηλαδή περιγράφεται μόνο μέσω εισόδων-εξόδων.
- δεν έχουμε γνώση για τις παραγώγους της, το οποίο είναι άμεσο επακόλουθο της προηγούμενης πρότασης.
- είναι non-convex.
- δεν εξαρτάται εξίσου από όλες τις παραμέτρους.
- ο υπολογισμός της για δεδομένο λ είναι υπολογιστικά και χρονικά απαιτητικός.

Η θεωρία της βελτιστοποίησης συναρτήσεων έχει προσφέρει ποικίλλες επιλογές στην επίλυση του υπό μελέτη προβλήματος. Εξελικτικοί αλγόριθμοι [14], gradient decent [13], αλγόριθμοι [5] βασισμένοι σε ευριστικές [12]. Τα χαρακτηριστικά ωστόσο που αναφέρουμε προσδίδουν στη βελτιστοποίηση 1.1 ιδιαιτερότητες, που συγκεκριμενοποιούν τον κατάλληλο αλγόριθμο βελτιστοποίησης.

2.1 Bayesian βελτιστοποίηση

Η τεχνική αυτή επικεντρώνεται στην απαίτηση βελτιστοποίησης μιας άγνωστης, κοστοβόρας συνάρτησης με μικρό πλήθος evaluations. Προς αυτό το σκοπό αντικαθιστά την L με ένα πιθανοτικό μοντέλο, το οποίο ενσωματώνει πληροφορία για όλα τα προηγούμενα γνωστά evaluated σημεία, και στη συνέχεια το εκμεταλλεύεται για να αποφασίσει το επόμενο σημείο αξιολόγησης (evaluation).

Δύο είναι οι βασικές σχεδιαστικές επιλογές κατά την εφαρμογή αυτής της μεθόδου:

PRIOR OVER FUNCTIONS περιέχει υποθέσεις σχετικά με την άγνωστη συνάρτηση L

Μία συνήθης επιλογή [16] αποτελεί η χρήση Γκαουσιανών διαδικασιών ως priors. Πρόκειται για ένα σύνολο τυχαίων μεταβλητών, για το οποίο ισχύει ότι οποιοδήποτε υποσύνολο έχει γκαουσιανή κατανομή πολλών μεταβλητών.

Η τυχαία μεταβλητή ορίζεται ως εξής:

$$X = \begin{cases} 0 & \textit{case 1} \\ 1 & \textit{case 2} \end{cases} \tag{2.1}$$

Η γκαουσιανή κατανομή ορίζεται ως εξής

$$X \approx N(\mu, \sigma^2) = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \tag{2.2}$$

Η γκαουσιανή κατανομή πολλών μεταβλητών χαρακτηρίζεται από το γεγονός ότι ο γραμμικός συνδυασμός τους δίνει γκαουσιανή κατανομή. Χρησιμοποιείται για την περιγραφή συσχετισμένων τυχαίων μεταβλητών.

Συνάρτηση Απόκτησης χρησιμοποιείται για την απόδοση "χρησιμότητας εξερεύνησης" στο χώρο με βάση το πιθανοτικό μοντέλο. Οι συναρτήσεις που χρησιμοποιούνται στη βιβλιογραφία είναι:

- Πιθανότητα βελτίωσης
- Προσδοκώμενη βελτίωση
- Άνω όρια εμπιστοσύνης

Η προσδοκώμενη βελτίωση χρησιμοποιείται σχεδόν καθολικά, καθώς δίνει καλύτερα πειραματικά αποτελέσματα, είναι κατανοητή (intuitive) και δε χρειάζεται ρύθμιση [16], [3].

Η προσέγιση της L που εμφανίζεται στην 1.1 μπορεί να γίνει με τη χρήση μοντέλου, οπότε η βελτιστοποίηση ονομάζεται Sequential Model-Based Optimization (SMBO). Παραδείγματα χωρίς χρήση μοντέλου είναι οι αλγόριθμοι Random Online Adaptive Racing (ROAR) [7], Covariance Matrix Adaptation Evolution Strategy (CMA-ES) [11], Sequential Model-free Optimization (SMFO) [17]

2.1.1 SMBO

Η υπεροχή των SMBO έγκειται στη δυνατότητα παρεμβολής (interpolation) μεταξύ παρατηρούμενων σετ υπερπαραμέτρων και παρέκτασης (extrapolation) σε άγνωστες περιοχές του χώρου παραμέτρων. Επίσης, ποσοτικοποιούν τη σημασία κάθε υπερπαραμέτρου και των αλληλεξαρτήσεων.

Μερικά προβλήματα, τα οποία αντιμετωπίζει η κοινωνία του machine learning οφείλονται σε χαρακτηριστικά των SMBO αλγοριθμών, που προκύπτουν από την ικανότητά τους να λύνουν γενικής φύσης προβλήματα βελτιστοποίησης συναρτήσεων μαύρου κουτιού, εισάγοντας περιορισμούς ανυπόστατους στη μηχανική μάθηση. Τέτοιοι περιορισμοί είναι:

- υπόθεση ντετερμινιστικότητας της συνάρτησης προς βελτιστοποίηση
- ακριβή (costly) σχεδίαση των αρχικών πειραμάτων
- εξάρτηση από υπολογιστικά ακριβά μποντέλα

- υπόθεση ίδιου κόστος για κάθε υπολογισμό (run) του αλγορίθμου-στόχου

Παραδείγματα SMBO αλγορίθμων είναι:

- Sequential Model-based Algorithm Configuration (SMAC) [7]
- Tree Parzen Estimator [3]
- Spearmint, ο οποίος αναλύεται από τους [16], χωρίς ωστόσο να αναφέρεται το όνομά του.

3 Ειδικά Ζητήματα

3.1 Χώρος υπερπαραμέτρων

Το πλήθος των μεταβλητών που αποτελούν άξονες του χώρου διαφέρει αισθητά μεταξύ διαφορετικών αλγορίθμων μηχανικής μάθησης, ενώ το είδος περιλαμβάνει συνεχείς, ακέραιες και κατηγορικές μεταβλητές. Ένα βασικό χαρακτηριστικό του χώρου είναι η δενδρική μορφή, η οποία υποδηλώνει υποθετική παρουσία μεταβλητών δεδομένων προγόνων τους, για παράδειγμα το πλήθος των νευρώνων στο δεύτερο επίπεδο ενός νευρωνικού λαμβάνει υπόσταση όταν το πλήθος των επιπέδων είναι τουλάχιστον 2.

3.2 Συνάρτηση κόστους

Η επιλογή της συνάρτησης κόστους περιλαμβάνει τις απαιτήσεις του προβλήματος βελτιστοποίησης και τις παραδοχές που έχουν γίνει. Τεχνικές που συναντώνται στη σύγχρονη βιβλιογραφία είναι:

- ελαχιστοποίηση σφάλματος του μοντέλου
- ελαχιστοποίηση των evaluations, τεχνική η οποία υποθέτει ίδιο χρόνο για κάθε evaluation
- ελαχιστοποίηση του χρόνου διαδικασίας βελτιστοποίησης
- συνδυασμός μετρικών, όπως το Adjusted Ratios of Ratio (ARR) [1]
- οι Hutter et al. [9] εισάγουν περιορισμούς και μηχανισμούς στον αλγόριθμο Sequential Parameter Optimization [2], αναγκάζοντας τον αλγόριθμο να λειτουργεί με συγκεκριμένο χρονικό budget. Αυτό έχει ως αποτέλεσμα τη μείωση του χρόνου, χωρίς όμως αυτό να είναι απαραίτητητα το κριτήριο βελτιστοποίησης.

3.3 Αξιολόγηση και έλεγχος

Στο τέλος της βελτιστοποίησης πρέπει να αξιολογηθούν οι υπερπαράμετροι που επιλέχθηκαν ως προς την επίτευξη του στόχου, αλλά και να συγκριθούν σετ υπερπαραμέτρων που παρήχθησαν με διαφορετικές τεχνικές, ώστε και αυτές να αξιολογηθούν.

3.3.1 Αξιολόγηση αλγορίθμου βελτιστοποίησης

Ο αλγόριθμος βελτιστοποίησης οφείλει να αξιολογηθεί ως προς την ικανότητά του να αποδίδει στις υπερπαραμέτρους το σωστό κόστος, να βρίσκει το βέλτιστο συνδυασμό υπερπαραμέτρων και ενδεχομένως να ελαχιστοποιεί τους καταναλώσιμους πόρους. Μερικοί βασικοί δείκτες για την αξιολόγηση ενός αλγορίθμου βελτιστοποίησης σύμφωνα με τους Hutter et al. [9] είναι οι εξής:

- η ρίζα της μέσης τιμής της διαφοράς μεταξύ προβλεφθείσας και πραγματικής απόδοσης των υπερπαραμέτρων(RMSE)
- Quality of predictive ranks, η οποία υπολογίζεται μέσω του συντελεστή συσχέτισης Spearman. Πρόκειται για μία μη-παραμετρική μετρική συσχέτισης κατάταξης, η οποία εκτιμά τη δυνατότητα περιγραφής της σχέσης μεταξύ δύο μεταβλητών μέσω μίας μονότονης συνάρτησης.
- EIC quality. Πρόκειται για το συντελεστή συσχέτισης Spearman μεταξύ της πραγματικής απόδοσης και του Expected Improvement Criterion, το οποίο χρησιμοποιείται στην acquisition function με βάση το πιθανοτικό μοντέλο. Ο δείκτης αυτός είναι χρήσιμος για την αξιολόγηση του μοντέλου σε έναν SMBO αλγόριθμο, καθώς η βασική χρήση του μοντέλου είναι στην επιλογή υποσχόμενων υπερπαραμέτρων με βάση το EIC.
- χρόνος διεξαγωγής βελτιστοποίησης σε μορφή Box plots του $\log_{10}(time)$

Συχνή είναι και η χρήση διαγραμμάτων:

Generalization performance plots Σύμφωνα με τους Pedregosa [13] τα διαγράμματα γενίκευσης απόδοσης ως προς το χρόνο είναι καταλληλότερα σε σχέση με τα sub-optimality plots που χρησιμοποιούνται συχνά στην αξιολόγηση βελτιστοποίησης συναρτήσεων.

AUC vs generation Μία εναλλακτική των generalization performance plots, στην οποία φαίνεται το εμβαδό της περιοχής κάτω από τη καμπύλη Receiver Operating Characteristic (Area under Curve) σε συνάρτηση των γενεών του αλγορίθμου βελτιστοποίησης χρησιμοποίεται από τους Rojas and Fernandez-Reyes [14]

3.3.2 Σύγκριση μεταξύ διαφορετικών τεχνικών

Σε περίπτωση που έχουμε διεξάγει πειράματα για δύο διαφορετικές τεχνικές βελτιστοποίησης, τα οποία έχουν δημιουργήσει δύο σετ λύσεων, που το καθένα περιλαμβάνει πειράματα για τη δεδομένη τεχνική και ποικίλλα σετ δεδομένων και θέλουμε να διαπιστώσουμε στατιστικά σημαντική διαφορά μεταξύ των δύο τεχνικών.

Wilcoxon signed-rank Πρόκειται για ένα ισχυρό στατιστικό τεστ για τη σύγκριση μεθόδων σε διαφορετικά σετ δεδομένων και είναι μία μη-παραμετρική εκδοχή του Student's t-test, η οποία δεν απαιτεί την παραδοχή της κανονικής κατανομής. Χρησιμοποιείται από τους Levésque, Gagné, and Sabourin [10]

Θηκογράμματα (Box plots) Η χρήση Box plots οπτικοποιεί την κατανομή των λύσεων που έχει δωθεί από μία τεχνική βελτιστοποίησης και η παράθεση διαφορετικών Box plots στο ίδιο διάγραμμα χρησιμοποιείται συχνά [7] για σύγκριση μεταξύ τεχνικών.

Mann-Whitney U test Το Mann-Whitney U τεστ, ή Wilcoxon rank-sum τεστ, είναι ένα μη παραμετρικό τεστ της μηδενικής υπόθεσης ότι είναι εξίσου πιθανό μία τυχαία επιλεχθείσα τιμή από το ένα σύνολο να είναι μικρότερη ή μεγαλύτερη από μία τυχαία τιμή του δεύτερου συνόλου. Σε αντίθεση με το t-test δεν απαιτεί την υπόθεση κανονικής κατανομής και είναι σχεδόν τόσο αποτελεσματικό όσο το πρώτο σε κανονικές κατανομές. [7] [8]

3.3.3 Εκτίμηση σημασίας υπερπαραμέτρων

Η κατανόηση της επιρροής που έχει κάθε υπερπαράμετρος στην απόδοση ενός αλγορίθμου μάθησης, καθώς και η συσχέτιση μεταξύ των υπερπαραμέτρων μπορεί να προσφέρει μία ποιοτική βάση στο πρόβλημα βελτιστοποίησης, αναδεικνύοντας τις σημαντικές υπερπαραμέτρους και γενικότερα υποχώρους του χώρου παραμετροποίησης.

Οι Hutter, Hoos, and Leyton-Brown [6] κατάφεραν με χρήση της τεχνικής ANOVA και ενός αλγορίθμου υπολογισμού των marginal predictions μίας συνάρτησης μαύρου κουτιού με random forests να αναδείξουν σημαντικές υπερπαραμέτρους και συσχετίσεις μεταξύ αυτών.

3.4 Επιτάχυνση βελτιστοποίησης

Οι χρονικοί και υπολογιστικοί πόροι καθορίζουν την ακρίβεια του αποτελέσματος της βελτιστοποίησης, η οποία σε πολλές περιπτώσεις είναι απαγορευτικά ακριβή.

Παραδείγματα της προσπάθειας επιτάχυνσης της διαδικασίας είναι:

- η εισαγωγή χρονικών ορίων που είδαμε [9],
- η προσέγγιση της προβαλλόμενης Διαδικασίας (Projected Process Approximation) [9], η οποία μειώνει την πολυπλοκότητα των Γκαουσιανών Διαδικασιών.
- η χρήση της τεχνικής Cholesky Decomposition [15] κατά τον υπολογισμό του αντίστροφου του πίνακα πυρήνα (που προκύπτει στη ρύθμιση των Γκαουσιανών Διαδικασιών).

4 Χρήση μετα-γνώσης

Ως μετα-γνώση ορίζουμε την πληροφορία που έχει παραχθεί σε παλαιότερα πειράματα με διαφορετικά σετ δεδομένων, την οποία θα θέλαμε να ενσωματώσουμε στο σετ δεδομένων υπό βελτιστοποίηση.

Οι Schilling, Wistuba, and Schmidt-Thieme [15] αναγνωρίζουν δύο άξονες εφαρμογής μετα-γνώσης στο πρόβλημα της βελτιστοποίησης:

- χρήση μετα-χαρακτηριστικών των παρελθοντικών σετ δεδομένων
- χρήση μετα-γνώσης για την αρχικοποίηση του αλγορίθμου βελτιστοποίησης. [4]

Οι ίδιοι, κινούμενοι στον πρώτο άξονα, υπολογίζουν την απόδοση 50 διαφορετικών σετ δεδομένων σε ένα πλέγμα υπερπαραμέτρων, καθώς και μετα-χαρακτηριστικά των σετ δεδομένων δημιουργώντας ένα σετ μετα-δεδομένων, το οποίο χρησιμοποιούν κατά τη βελτιστοποίηση. Είναι σημαντικό πως για κάθε σετ δεδομένων χρησιμοποιούν μία διαφορετική γκαουσιανή διαδικασία, συναθροίζοντας την πληροφορία με χρήση της θεωρίας των γινομένων γκαουσιανών διαδικασιών.

Βιβλιογραφία

[1] Salisu Mamman Abdulrahman and Pavel Brazdil. "Measures for Combining Accuracy and Time for Meta-learning". In: *Proceedings of the 2014 International Conference on Meta-learning and Algorithm Selection - Volume 1201*. MLAS'14. Prague, Czech Republic: CEUR-WS.org, 2014, pp. 49–50. ISBN: 1613-0073.

[2] T. Bartz-Beielstein, C. W. G. Lasarczyk, and M. Preuss. "Sequential parameter optimization". In: *2005 IEEE Congress on Evolutionary Computation*. Vol. 1. Sept. 2005, 773–780 Vol.1. DOI: 10.1109/CEC.2005.1554761.

[3] James Bergstra et al. "Algorithms for Hyper-parameter Optimization". In: *Proceedings of the 24th International Conference on Neural Information Processing Systems*. NIPS'11. Granada, Spain: Curran Associates Inc., 2011, pp. 2546–2554. ISBN: 978-1-61839-599-3.

[4] Matthias Feurer, Jost Springenberg, and Frank Hutter. *Initializing Bayesian Hyperparameter Optimization via Meta-Learning*. 2015.

[5] D. Huang et al. "Global Optimization of Stochastic Black-Box Systems via Sequential Kriging Meta-Models". In: *Journal of Global Optimization* 34.3 (2006), pp. 441–466. ISSN: 1573-2916. DOI: 10.1007/s10898-005-2454-3.

[6] F. Hutter, H. Hoos, and K. Leyton-Brown. "An Efficient Approach for Assessing Hyperparameter Importance". In: *Proc. of ICML-14*. To appear. 2014.

[7] Frank Hutter, Holger H. Hoos, and Kevin Leyton-Brown. "Sequential Model-Based Optimization for General Algorithm Configuration". In: *Learning and Intelligent Optimization: 5th International Conference, LION 5, Rome, Italy, January 17-21, 2011. Selected Papers*. Ed. by Carlos A. Coello Coello. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 507–523. ISBN: 978-3-642-25566-3. DOI: 10.1007/978-3-642-25566-3_40.

[8] Frank Hutter et al. "An Experimental Investigation of Model-based Parameter Optimisation: SPO and Beyond". In: *Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation*. GECCO '09. Montreal, Québec, Canada: ACM, 2009, pp. 271–278. ISBN: 978-1-60558-325-9. DOI: 10.1145/1569901.1569940.

[9] F. Hutter et al. "Time-Bounded Sequential Parameter Optimization". In: *Proceedings of the conference on Learning and Intelligent OptimizatioN (LION 4)*. Jan. 2010.

[10] Julien-Charles Levésque, Christian Gagné, and Robert Sabourin. *Bayesian Hyperparameter Optimization for Ensemble Learning*. Version 1.

[11] Ilya Loshchilov and Frank Hutter. "CMA-ES for Hyperparameter Optimization of Deep Neural Networks". In: *CoRR* abs/1604.07269 (2016).

[12] J. A. Nelder and R. Mead. "A Simplex Method for Function Minimization". In: *The Computer Journal* 7.4 (1965), pp. 308–313. DOI: 10.1093/comjnl/7.4.308.

[13] Fabian Pedregosa. *Hyperparameter optimization with approximate gradient*. Version 1.

[14] S. A. Rojas and D. Fernandez-Reyes. "Adapting multiple kernel parameters for support vector machines using genetic algorithms". In: *2005 IEEE Congress on Evolutionary Computation*. Vol. 1. Sept. 2005, 626–631 Vol.1. DOI: 10.1109/CEC.2005.1554741.

[15] Nicolas Schilling, Martin Wistuba, and Lars Schmidt-Thieme. "Scalable Hyperparameter Optimization with Products of Gaussian Process Experts". In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I*. Ed. by Paolo Frasconi et al. Cham: Springer International Publishing, 2016, pp. 33–48. ISBN: 978-3-319-46128-1. DOI: 10.1007/978-3-319-46128-1_3.

[16] Jasper Snoek, Hugo Larochelle, and P. Ryan Adams. *Practical Bayesian Optimization of Machine Learning Algorithms*. Version 2.

[17] M. Wistuba, N. Schilling, and L.Schmidt-Thieme. "Sequential Model-Free Hyperparameter Tuning". In: *2015 IEEE International Conference on Data Mining*. Nov. 2015, pp. 1033–1038. DOI: 10.1109/ICDM.2015.20.