Automated Data Scientist

Εκπόνηση	Ελένη Νησιώτη
Επίβλεψη	Επίκ. Καθ. Ανδρέας Συμεωνίδης
Συνεπίβλεψη	Δρ. Κυριάκος Χατζηδημητρίου

Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
Πολυτεχνική Σχολή
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Εργαστήριο Επεξεργασίας Πληροφορίας και Υπολογισμών

Σκοπός διπλωματικής εργασίας

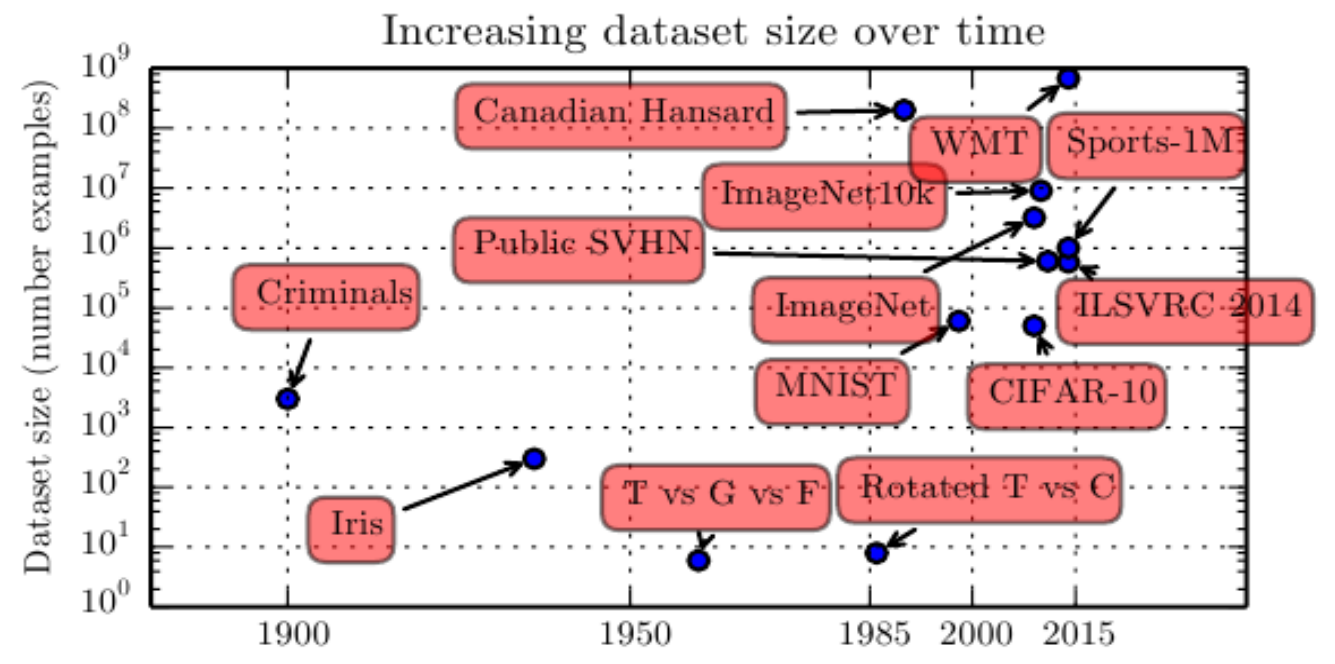
Προβληματισμός

Το 75% ενός πειράματος μηχανικής μάθησης αφιερώνεται στην προετοιμασία της εφαρμογής του αλγορίθμου και το 15% στα βήματα που την ακολουθούν. Το μεγαλύτερο μέρος της έρευνας επικεντρώνεται στο ενδιάμεσο 10% ...

— Rich Caruana, ICML 2015

Ανάγκη για επέκταση της τρέχουσας έρευνας σε χρονοβόρα στάδια της διαδικασίας εφαρμογής μηχανικής μάθησης που μέχρι τώρα γίνονταν χειροκίνητα.

Η εξέλιξη των προβλημάτων



Απαιτητικότητα προβλημάτων

Εμπλεκόμενη τεχνολογία

Η εξέλιξη του ML

Ο προγραμματισμός στοχεύει στην αυτοματοποίηση, η μηχανική μάθηση στην αυτοματοποίηση της αυτοματοποίησης και η αυτοματοποιημένη μηχανική μάθηση στην αυτοματοποίηση του να αυτοματοποιείς την αυτοματοποίηση.

— Matthew Mayo, KDnuggets 2017

Το νέο στάδιο στην εξέλιξη της μηχανικής μάθησης στοχεύει στη δημιουργία μετα-γνώσης για την αυτοματοποίηση της ίδιας της διαδικασίας της μάθησης και όχι μεμονωμένων προβλημάτων.

Η επιστήμη του Automl

Απαρχές	Unica, MarketSwitch, KXEN
Πεδία Εφαρμογής	Προ-επεξεργασία, Επιλογή αλγορίθμου, Ρύθμιση μοντέλου
Σύγχρονα Εργαλεία	AutoWeka, Microsoft Azure, caret, HPOlib, Google Automl

Προτεινόμενο σύστημα

Ένας αυτοματοποιημένος αναλυτής δεδομένων για προβλήματα δυαδικής ταξινόμησης με εμπειρία παλαιότερων πειραμάτων και κατανοητή έξοδο.

- Αυτόματος σχηματισμός βέλτιστου ensemble
- Ενσωμάτωση μετα-μάθησης
- Παραγωγή επεξηγηματικού report για το χρήστη

Μεθοδολογία

No free lunch theorem

Αν λάβουμε υπόψιν όλες τις πιθανές κατανομές δημιουργίας δεδομένων, τότε κάθε αλγόριθμος μηχανικής μάθησης επιδεικνύει κατά μέσο όρο το ίδιο σφάλμα στην πρόβλεψη άγνωστων παραδειγμάτων.

— David Wolpert, 1996

Είναι αδύνατη η εύρεση ενός γενικά βέλτιστου αλγορίθμου. Ένα εργαλείο βελτιστοποίησης γενικής φύσης προβλημάτων οφείλει να εξερευνήσει το χώρο των πιθανών μηχανισμών δημιουργίας δεδομένων και να προσφέρει προσαρμοσμένες λύσεις.

Ρύθμιση μοντέλου: σύγχρονες τεχνικές βελτιστοποίησης

Πλεγματική	Τυχαία	Bayesian
χρονοβόρα	naive	χρονοβόρα
ισοπίθανη αντιμετώπιση		intuitive

Ρύθμιση μοντέλου: επισκόπηση

Οι σύγχρονες τεχνικές:

- είναι χρονοβόρες
- δεν μπορεί να αποδειχθούν γενικά βέλτιστες
- είναι ad-hoc

Μετα-μάθηση

Σκοπός	Δημιουργία μετα-γνώσης από πειράματα μηχανικής μάθησης
Τρόπος	Εξαγωγή μετα-χαρακτηριστικών των σετ δεδομένων, τα οποία περιέχουν ουσιώδη πληροφορία
Εφαρμογές	Πρόβλεψη βέλτιστου αλγορίθμου, υπερ-παραμέτρων

Μετα-χαρακτηριστικά

Πίνακας 1: Λίστα μετα-χαρακτηριστικών μετά από εφαρμογή φιλτραρίσματος

Άθροισμα αθροισμάτων	Τυπική απόκλιση επιπέδων
Άθροισμα μέγιστων τιμών	Κυρτότητα επιπέδων
Μέση τιμή τυπικών αποκλίσεων	Λοξότητα επιπέδων
Μέση τιμή ελαχίστων τιμών	Πλήθος χαρακτηριστικών
Μέση τιμή κυρτοτήτων	Λογάριθμος πλήθους χαρακτηριστικών
Μέση τιμή λοξοτήτων	Πλήθος παραδειγμάτων
Τυπική απόκλιση ελαχίστων τιμών	Λογάριθμος πλήθους παραδειγμάτων
Ελάχιστη τιμή μέσων τιμών	Ποσοστό αγνώστων τιμών
Ελάχιστη τιμή τυπικών αποκλίσεων	Πλήθος αριθμητικών χαρακτηριστικών
Ελάχιστη τιμή ελαχίστων τιμών	Πλήθος κατηγορικών χαρακτηριστικών
Ελάχιστη τιμή μεγίστων τιμών	Μέγιστη πιθανότητα κλάσης
Ελάχιστη τιμή λοξοτήτων	Μέση τιμή πιθανοτήτων κλάσης
Κυρτότητα ελαχίστων τιμών	Ποσοστό PC για 95% διακύμανση
Κυρτότητα μεγίστων τιμών	Κυρτότητα πρώτης PC
Λοξότητα λοξοτήτων	Λοξότητα PC
Άθροισμα επιπέδων	

Ensembles

Μεταξύ ανταγωνιζομένων υποθέσεων πρέπει να επιλέγεται η απλούστερη.

— Το ξυράφι του Occam

Ο συνδυασμός σωστών λύσεων σε ένα πρόβλημα, δε μπορεί παρά να λύνει το πρόβλημα τουλάχιστον εξίσου καλά.

— Επίκουρος

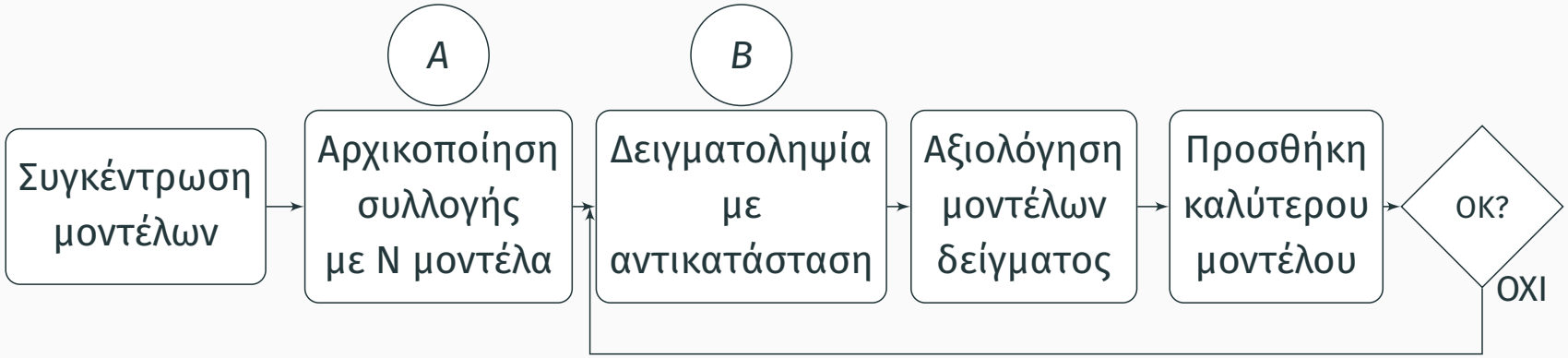
Μία αναγκαία και ικανή συνθήκη για να είναι μία συλλογή μοντέλων πιο ακριβής από τα μοντέλα που την απαρτίζουν είναι αυτά να είναι ακριβή και ετερογενή.

— Dietterich

Ensembles από αποθήκες μοντέλων

Πρόβλημα	Παρουσία πολλών, ετερογενών και ενδεχομένως κακής ποιότητας μοντέλων
Στόχος	Υπολογιστικά εφικτή τεχνική σχηματισμού συλλογής των αποδοτικότερων μοντέλων με αποφυγή υπερπροσαρμογής

Ensemble με προς τα εμπρός επιλογή μοντέλων



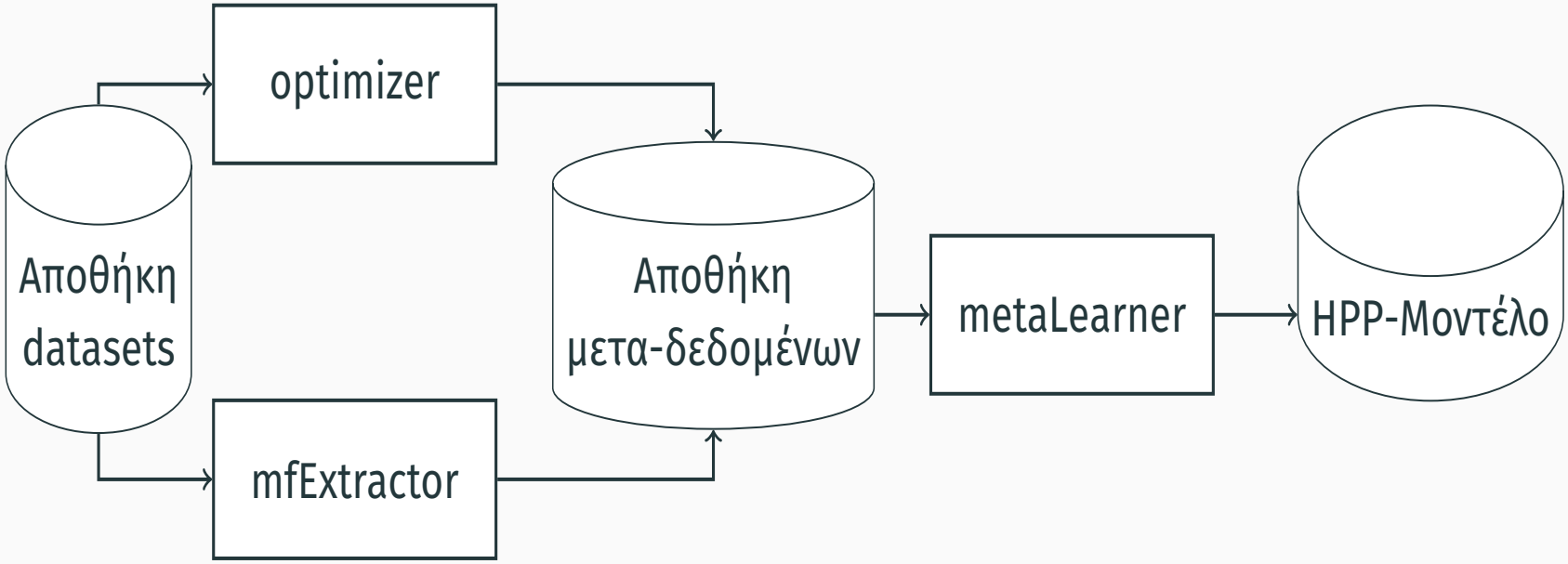
Α: αποφυγή υπερ-προσαρμογής σε μικρές αποθήκες

Β: αποφυγή υπερ-προσαρμογής σε μεγάλες αποθήκες και αναγκαστικής συμπερίληψης κακών μοντέλων

Αρχιτεκτονική Συστήματος

Υποσύστημα εκπαίδευσης

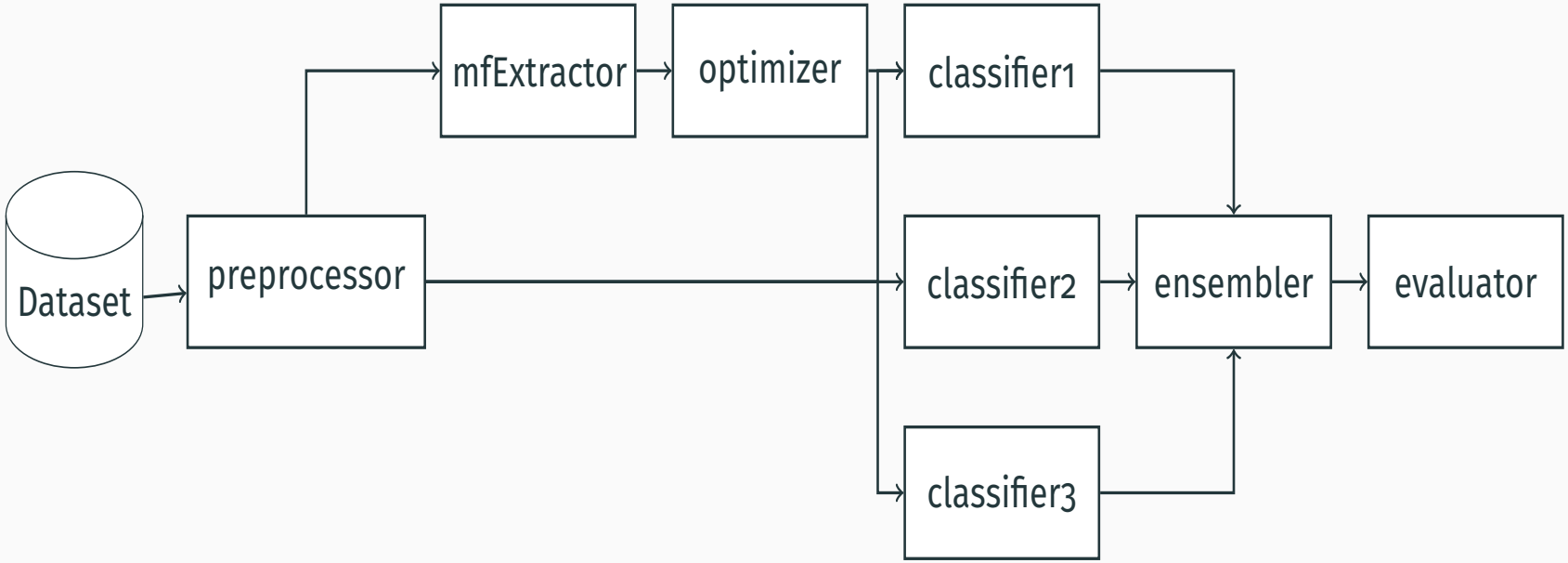
Σκοπός: Εκπαίδευση HyperParameterPrediction (HPP) μοντέλων



1. optimizer: Βελτιστοποίηση υπερ-παραμέτρων
2. mfExtractor: Εξαγωγή μετα-χαρακτηριστικών
3. metaLearner: Εκπαίδευση HPP μοντέλων

Υποσύστημα πειράματος

Σκοπός: Παραγωγή βέλτιστου ensemble για δεδομένο πρόβλημα

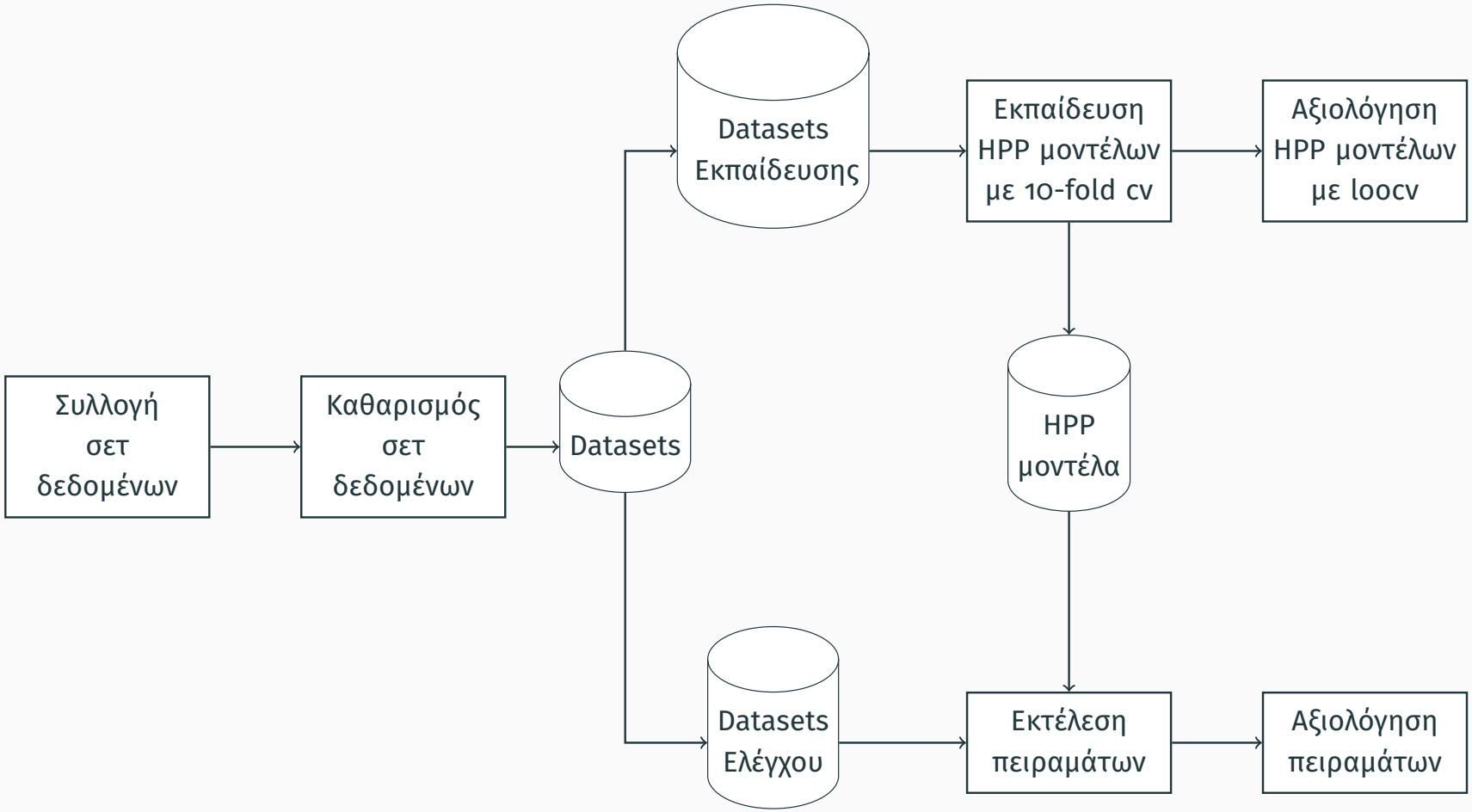


1. preprocessor: Προεπεξεργασία
2. mfExtractor: Εξαγωγή μετα-χαρακτηριστικών
3. optimizer: Ρύθμιση μοντέλων

5. classifier$_i$: Εκπαίδευση μοντέλου
6. ensembler: Σχηματισμός ensemble
7. evaluator: Αξιολόγηση

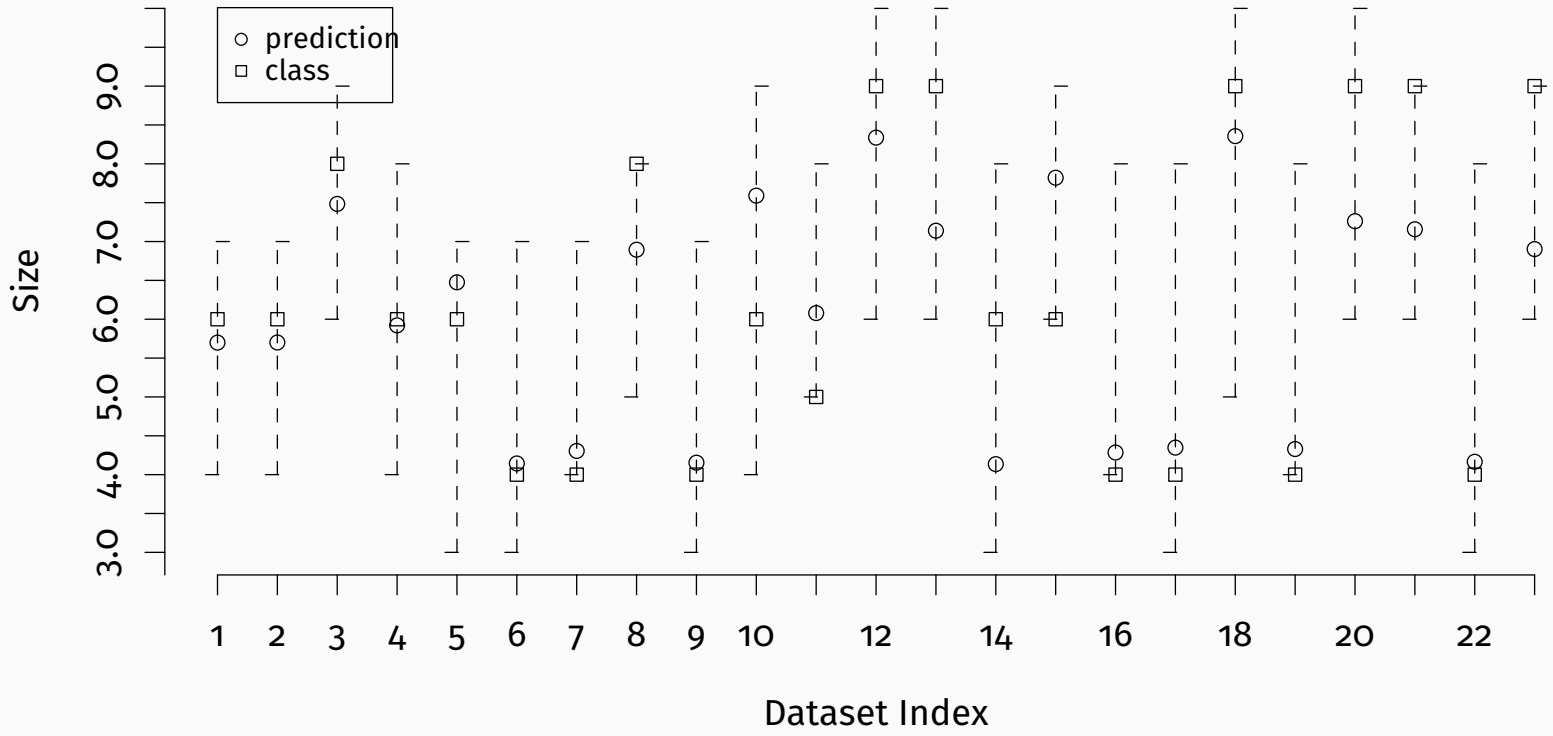
Πειραματικά Αποτελέσματα

Περιγραφή πειραμάτων



Αξιολόγηση HPP μοντέλων

Αξιολόγηση HPP μοντέλου για πρόβλεψη υπερ-παραμέτρου size για το ANN μοντέλο

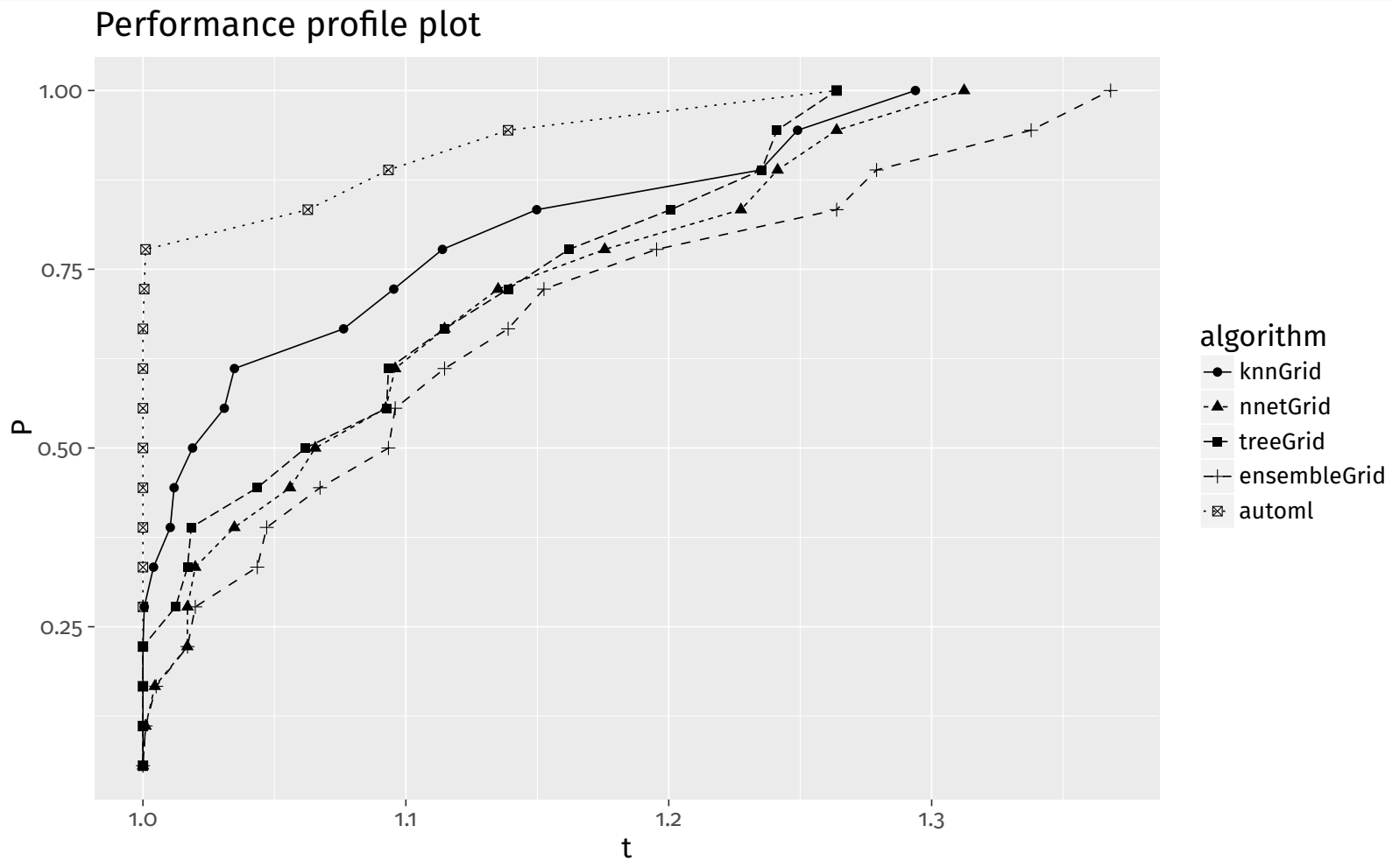


Αξιολόγηση HPP μοντέλων: συμπεράσματα

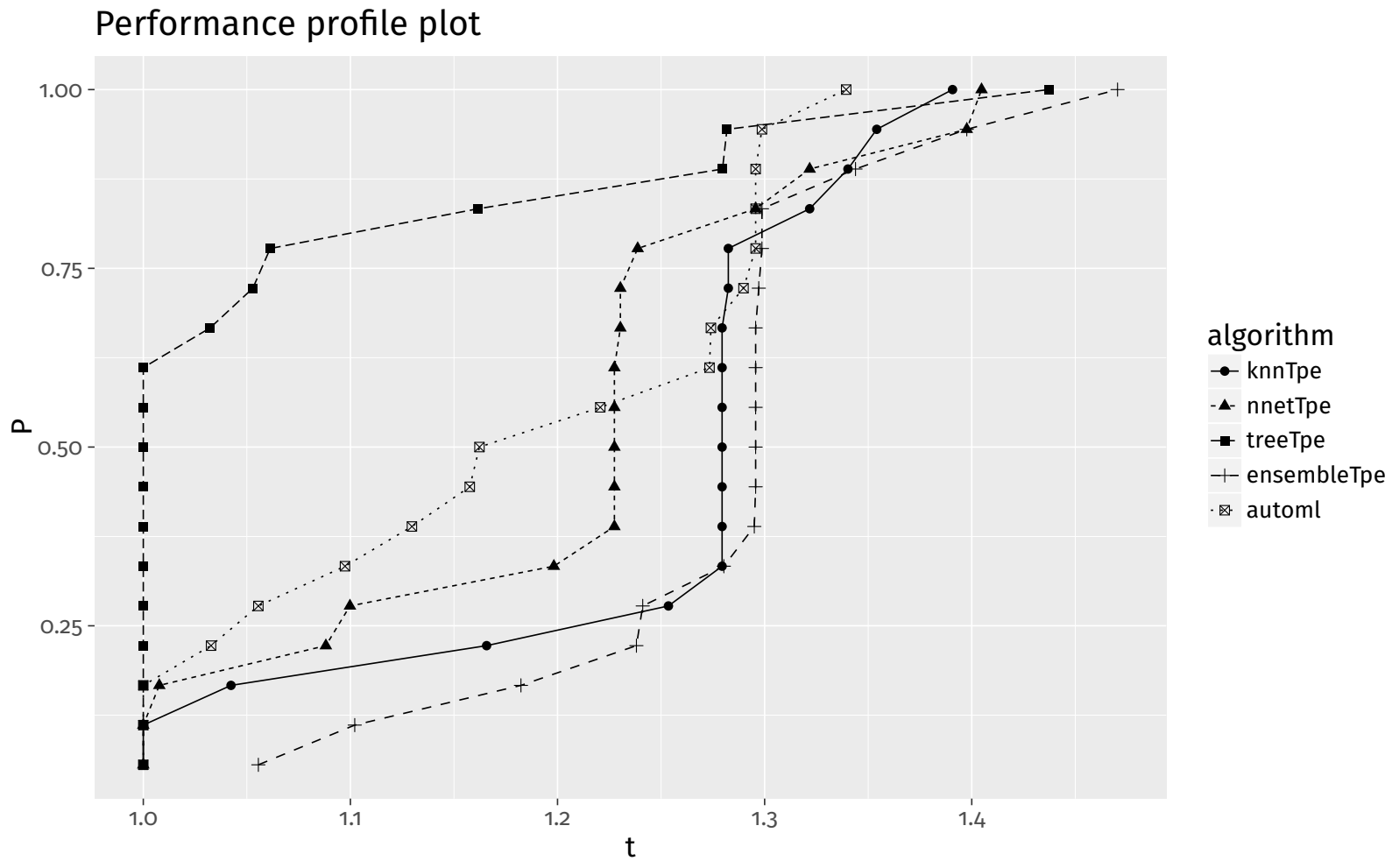
Από τα πειράματά μας προκύπτει ότι:

- τα HPP μοντέλα δεν προβλέπουν επακριβώς τις βέλτιστες υπερ-παραμέτρους, γεγονός που μάλλον οφείλεται στην αδυναμία των μετα-χαρακτηριστικών να περιγράψουν τον μηχανισμό επιρροής των υπερ-παραμέτρων στην απόδοση ενός μοντέλου.
- η χρήση διαστημάτων πρόβλεψης οδηγεί σχεδόν πάντα στη συμπερίληψη της σωστής τιμής, επομένως ο τελικός ensemble οφείλει να πετύχει απόδοση τουλάχιστον ίση με του βέλτιστου μοντέλου.

Αξιολόγηση συστήματος: σύγκριση με πλεγματική αναζήτηση



Αξιολόγηση συστήματος: σύγκριση με Bayesian βελτιστοποίηση



Αξιολόγηση συστήματος: συμπεράσματα

Με βάση τα διαγράμματα προφίλ απόδοσης και τα στατιστικά τεστ που εφαρμόσαμε μπορούμε να συμπεράνουμε ότι:

- το σύστημά μας είναι αποδοτικότερο από όλα τα μοντέλα που ρυθμίστηκαν με πλεγματική αναζήτηση
- το σύστημά μας είναι αποδοτικότερο από όλα τα μοντέλα που ρυθμίστηκαν με bayesian βελτιστοποίηση, εκτός από τα CART δέντρα. Το γεγονός ότι τα μοντέλα αυτά είναι αποδοτικότερα και από τον τελικό ensemble υποδηλώνει ότι η διαφορά αυτή οφείλεται στο σχηματισμό του ensemble και όχι στα HPP μοντέλα που χρησιμοποιεί το σύστημά μας.

Σύνοψη

Το λογισμικό που σχεδιάσαμε:

- επεκτείνει την τρέχουσα κατάσταση στη ρύθμιση υπερ-παραμέτρων ενσωματώνοντας μετα-μάθηση
- ενσωματώνει την εμπειρία της κοινότητας μέσω ευριστικών κανόνων
- εισάγει την R στις γλώσσες που χρησιμοποιούνται σε AutoML εργαλεία
- εξασφαλίζει κατανοητή και επαναχρησιμοποιήσιμη έξοδο του συστήματος

Μελλοντικές επεκτάσεις

Βελτίωση μοντέλων μετα-μάθησης:

- εύρεση νέων μετα-χαρακτηριστικών
- πειραματισμός με μεταβλητά διαστήματα εμπιστοσύνης

Παραλληλοποίηση των embarassingly parallel διαδικασιών:

- k-fold αξιολόγηση
- βελτιστοποίηση υπερ-παραμέτρων

Ενσωμάτωση διεπαφών αυτοματοποίησης για:

- συλλογή σετ δεδομένων
- εκπαίδευση μετα-μοντέλων
- χρήση ευριστικών κανόνων

?

Βιβλιογραφία

- Rich Caruana. “Research Opportunities in Automl”. In: *Automl Workshop, ICML 2015* (2015).
- Thomas G. Dietterich. “Ensemble Methods in Machine Learning”. In: *Multiple Classifier Systems: First International Workshop, MCS 2000 Cagliari, Italy, June 21–23, 2000 Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, pp. 1–15. ISBN: 978-3-540-45014-6. DOI: 10.1007/3-540-45014-9_1. URL: http://dx.doi.org/10.1007/3-540-45014-9_1.


Yoshua Bengio Ian Goodfellow and Aaron Courville. “Deep Learning”. Book in preparation for MIT Press. 2016. URL: http://www.deeplearningbook.org.


The Current State of Automated Machine Learning. http://www.kdnuggets.com/2017/01/current-state-automated-machine-learning.html. Accessed: 2017-06-14.


David H. Wolpert. “The Lack of a Priori Distinctions Between Learning Algorithms”. In: *Neural Comput.* 8.7 (Oct. 1996), pp. 1341–1390. ISSN: 0899-7667. DOI: 10.1162/neco.1996.8.7.1341. URL: http://dx.doi.org/10.1162/neco.1996.8.7.1341.