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

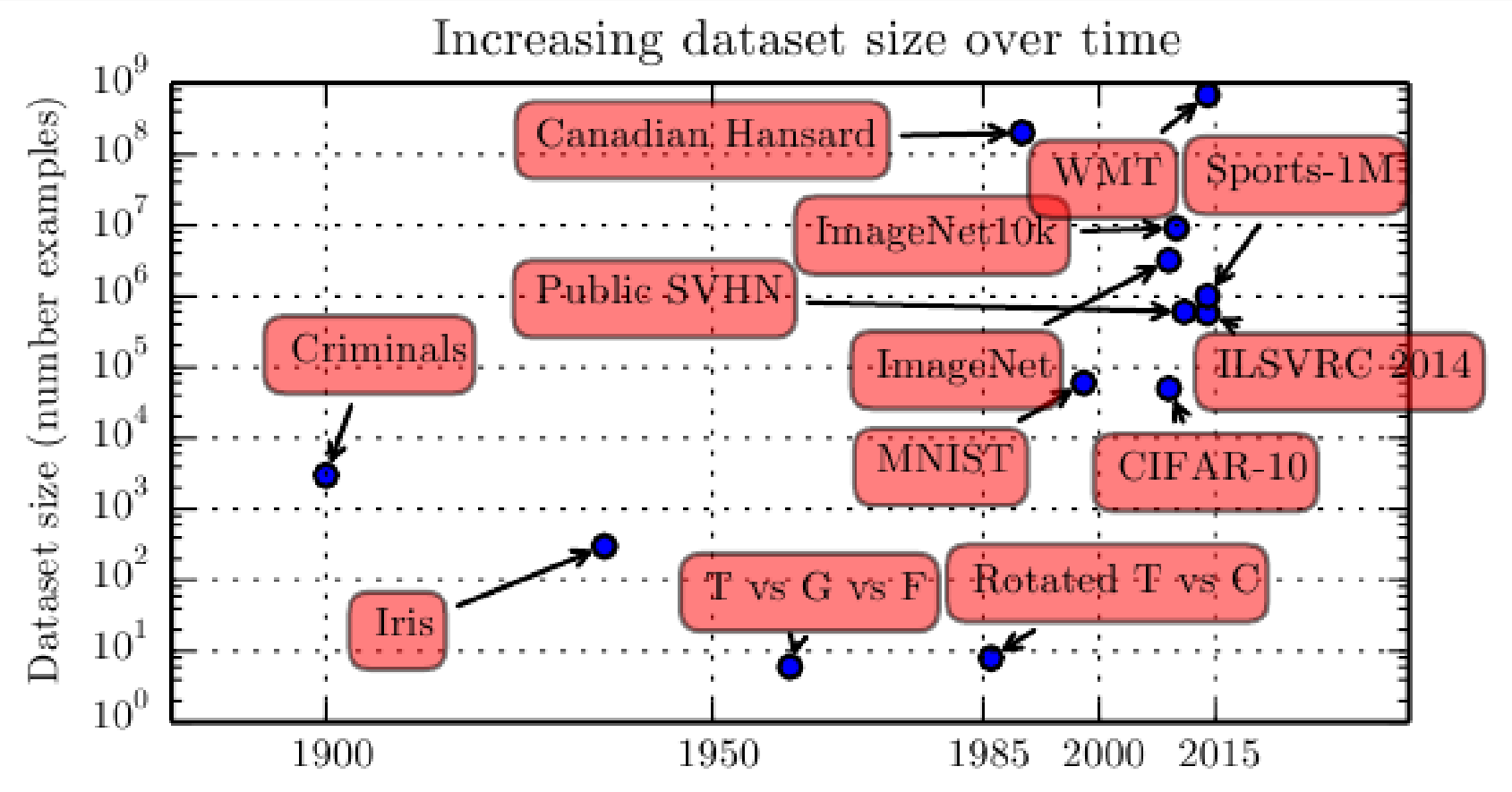
Σκοπός διπλωματικής εργασίας

Το πρόβλημα

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

Η εξέλιξη των προβλημάτων



Σχήμα 1: Θα αλλαχθεί, το βάζω για αναφορά

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

Η επιστήμη του Automl

Απαρχές	Unica, MarketSwitch, KXEN
Πεδία Εφαρμογής	Προεπεξεργασία, Ρύθμιση αλγορίθμου, Αξιολόγηση και κατανόηση 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.

Εμπλεκόμενη τεχνολογία

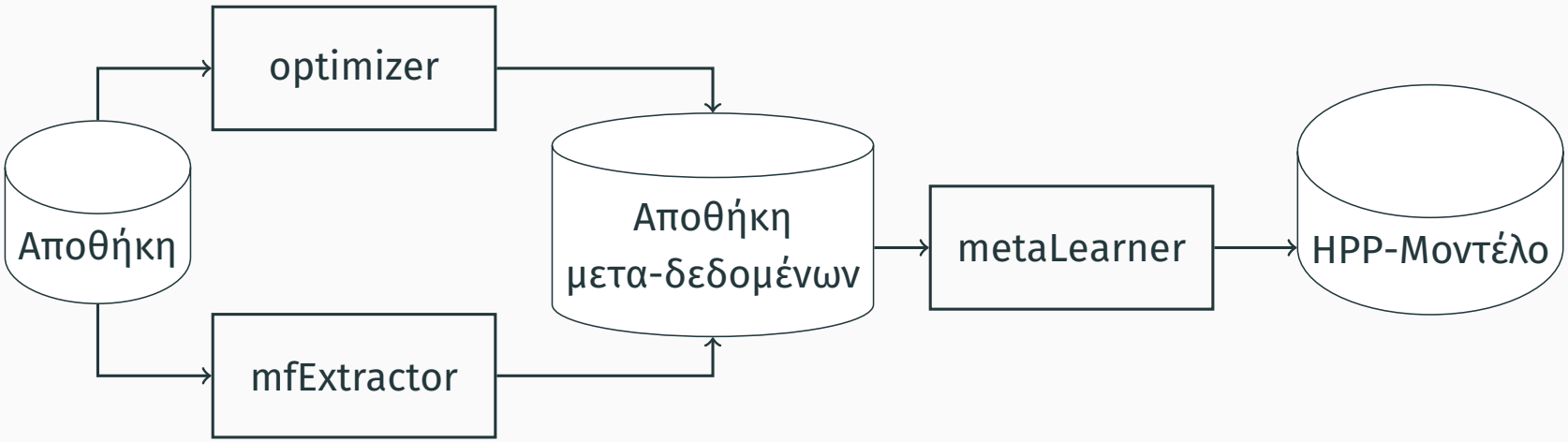
R

Automl

Τεχνολογία Λογισμικού

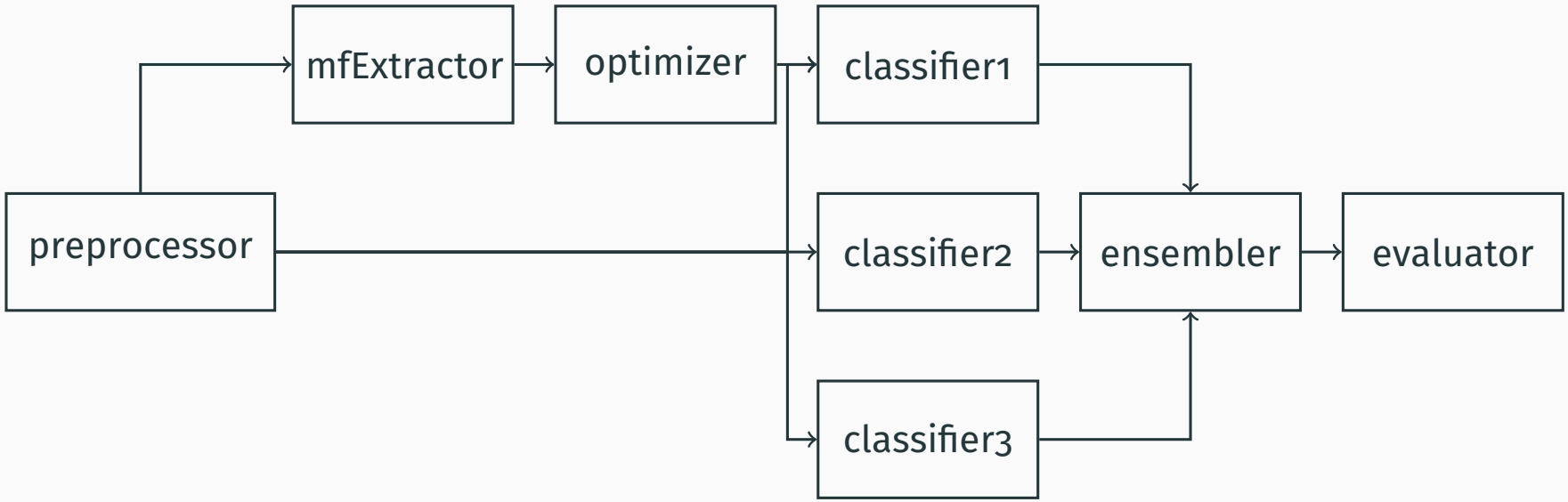
Αρχιτεκτονική Συστήματος

Υποσύστημα εκπαίδευσης



Σχήμα 2: Το υποσύστημα εκπαίδευσης

Υποσύστημα πειράματος



Σχήμα 3: Το υποσύστημα πειράματος

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

Ρύθμιση μοντέλου: τεχνικές βελτιστοποίησης

Πλεγματική αναζήτηση

Τυχαία αναζήτηση

Bayesian βελτιστοποίηση

Ρύθμιση μοντέλου: σχολιασμός-επισκόπηση

Χρονοβόρα

Δεν υπάρχει βέλτιστος αλγόριθμος (No free lunch)

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	Αρχικοποίηση αλγορίθμων βελτιστοποίησης

Ensembles

Μεταξύ ανταγωνιζομένων υποθέσεων πρέπει να επιλέγεται η απλούστερη.

— Τ Ό

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

— Ε

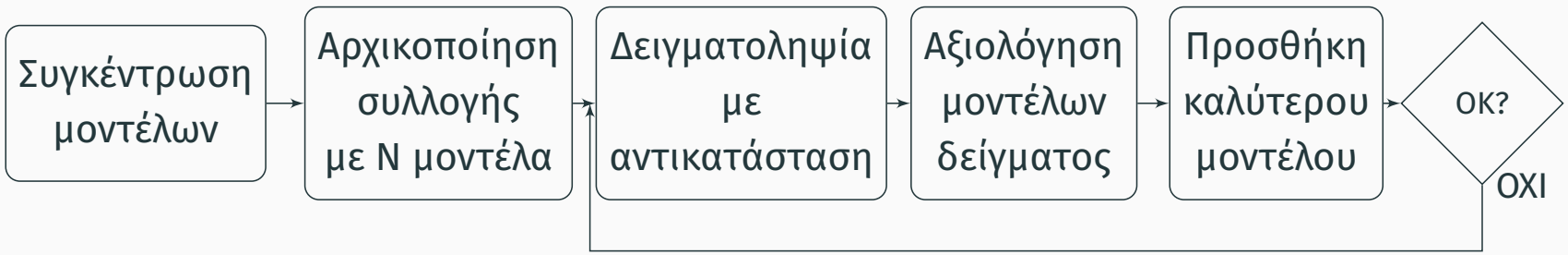
Μία αναγκαία και ικανή συνθήκη για να είναι μία συλλογή μοντέλων πιο ακριβής από τα μοντέλα που την απαρτίζουν είναι αυτά να είναι ακριβή και ετερογενή.

— Dietterich

Ensembles από αποθήκες μοντέλων

Πρόβλημα	Στόχος
Τεχνικές	Αποφυγή υπερ-προσαρμογής

Ensemble με προς τα εμπρός επιλογή μοντέλων



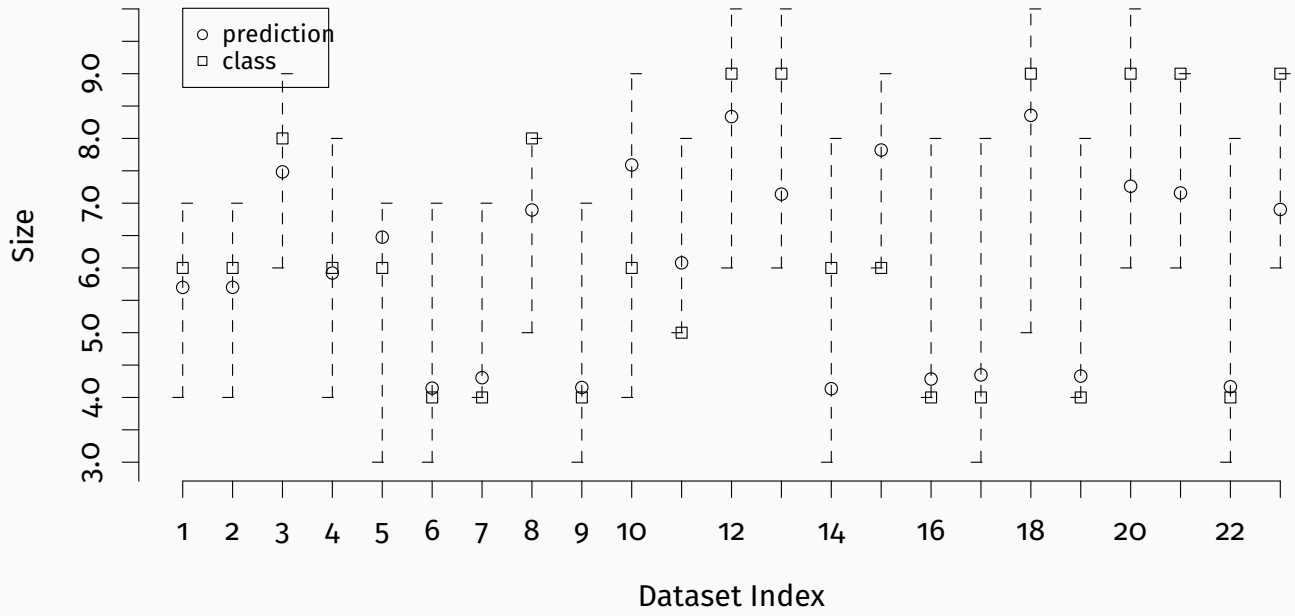
Σχήμα 4: Διάγραμμα ροής της διαδικασίας σχηματισμού μίας συλλογής μοντέλων με την τεχνική της προς τα εμπρός επιλογής μοντέλων

Πειραματικά Αποτελέσματα

Περιγραφή πειραμάτων

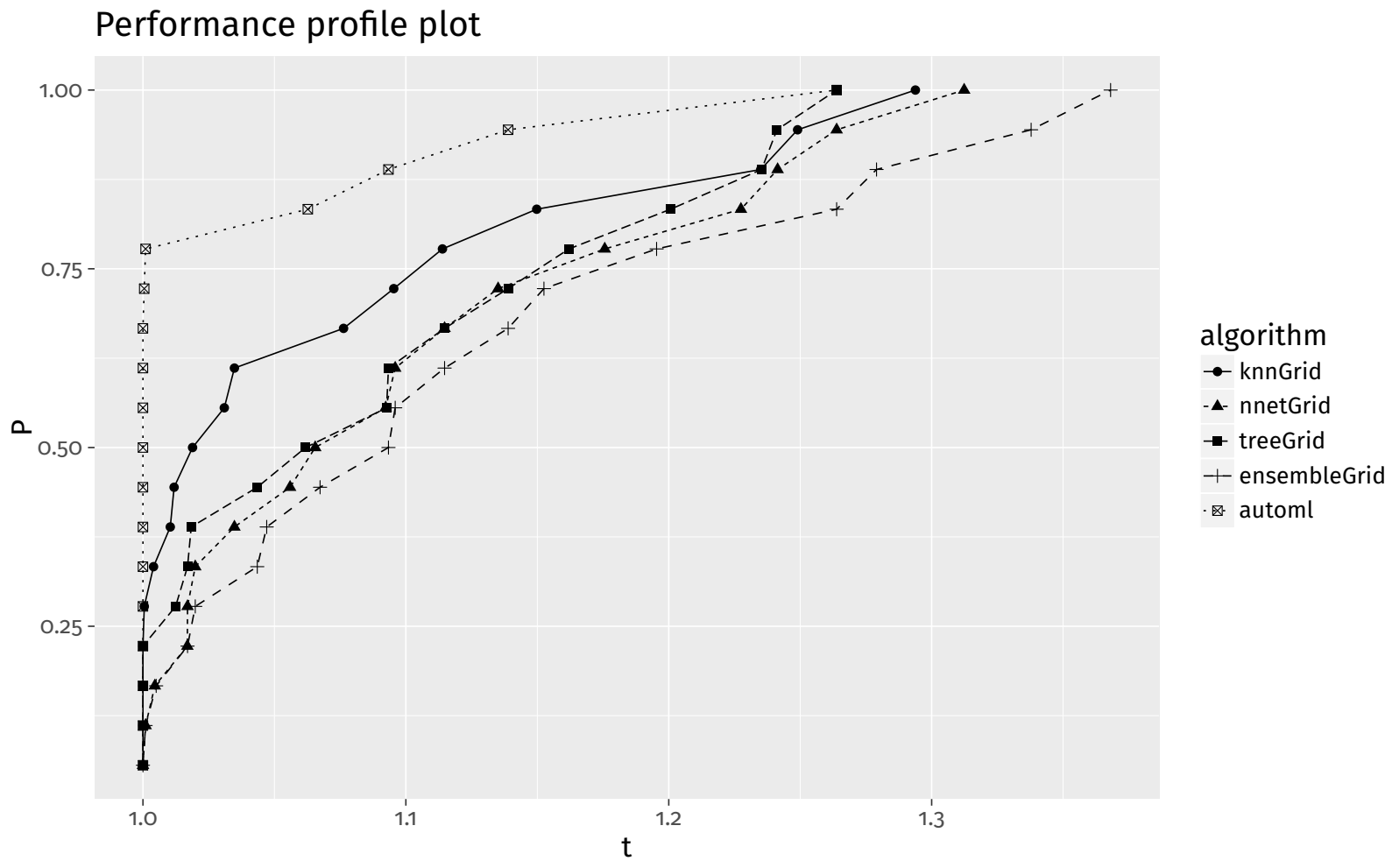
Σετ δεδομένων	50 σετ δυαδικής ταξινόμησης από το UCI Repository.
Off-line	Βελτιστοποίηση των αλγορίθμων SVM, NN, Tree, Bayes της caret με χρήση της βιβλιοθήκης HPOlib με hold-out validation, εξαγωγή μετα-χαρακτηριστικών και εκπαίδευση νευρωνικού με 10-fold CV. [Χρόνος και hardware]

Αξιολόγηση HyperParameterPrediction μοντέλων



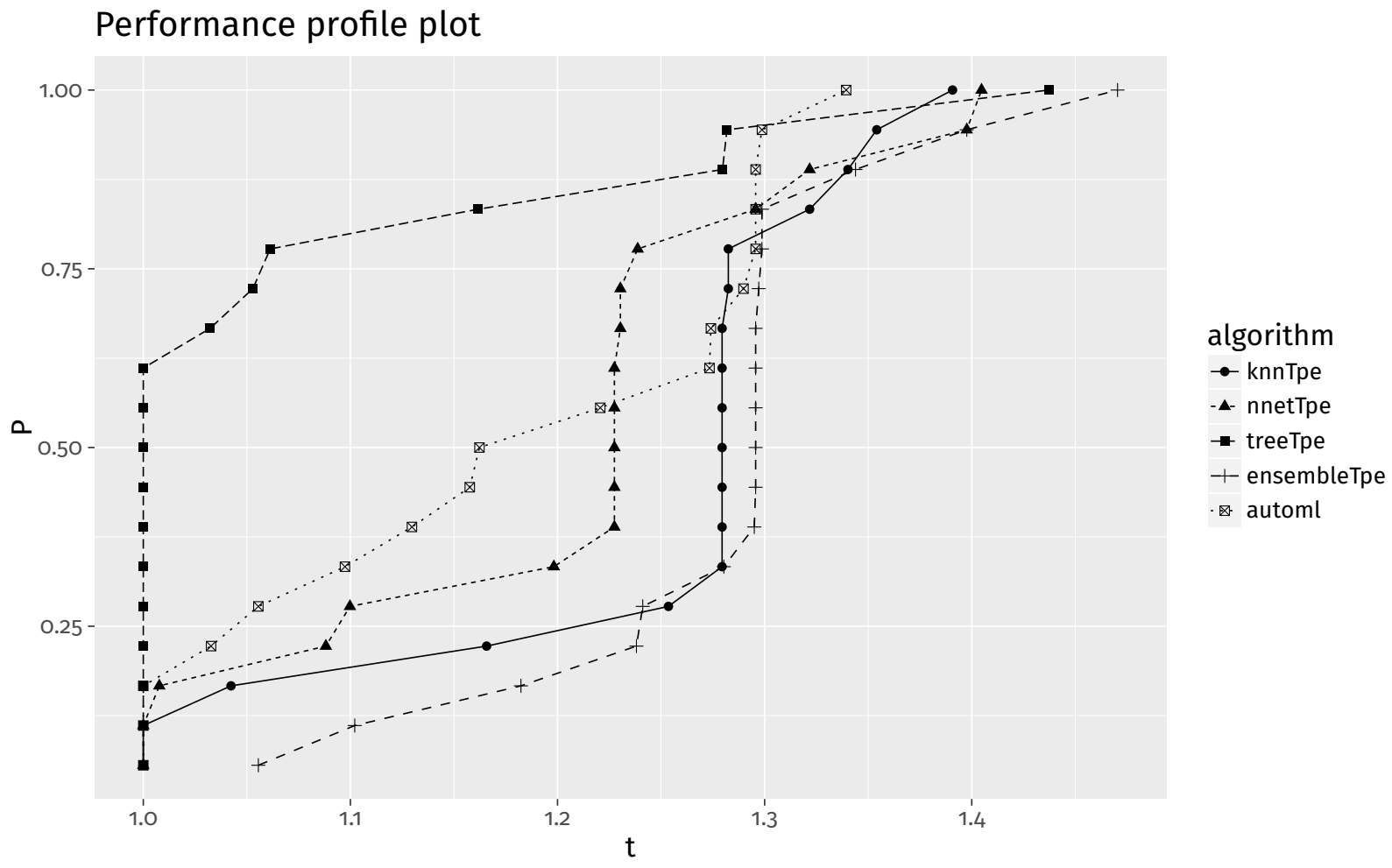
Σχήμα 5: Διάγραμμα διαστημάτων πρόβλεψης για υπερ-παράμετρο size.

Αξιολόγηση συστήματος Α



Σχήμα 6: Διάγραμμα προφίλ απόδοσης συστήματος ADS: σύγκριση του συστήματός μας με τη μέθοδο της πλεγματικής αναζήτησης

Αξιολόγηση συστήματος Β



Σχήμα 7: Διάγραμμα προφίλ απόδοσης συστήματος ADS: σύγκριση του συστήματός μας με τη μέθοδο της TPE βελτιστοποίησης.

Συμπεράσματα

Η επιστήμη του automl επιδιώκει ένα συμβιβασμό μεταξύ αυτοματοποίησης και κατανοησιμότητας (intuition). Το λογισμικό που σχεδιάσαμε συνδυάζει τη λογική (reasoning) ενός αναλυτή δεδομένων με τεχνικές αυτοματοποίησης. Η αρχιτεκτονική εξασφαλίζει την εύκολη ενσωμάτωση state of the art τεχνικών και πρόσβαση (interface) στη χρήσιμη πληροφορία.

Μελλοντικές επεκτάσεις

Βελτίωση μοντέλων μετα-μάθησης

Παραλληλοποίηση

Ενσωμάτωση διεπαφών αυτοματοποίησης για συλλογή σετ δεδομένων, χρήση ευριστικών,

Ερωτήσεις

Βιβλιογραφία(να μπει??)

 Rich Caruana. “Research Opportunities in Automl”. In: *Automl Workshop, ICML 2015* (2015).