ΠΡΟΣ

1) **Όλα τα μέλη ΔΕΠ του Τμήματος Επιστήμης Υπολογιστών**

2) **Τους εκπροσώπους των Μεταπτυχιακών φοιτητών του Τμήματος Επιστήμης Υπολογιστών**

3) **Την Επταμελή Εξεταστική Επιτροπή**

4) **Όλα τα μέλη της Πανεπιστημιακής Κοινότητας**

Πρόσκληση σε Δημόσια Παρουσίαση της Διδακτορικής Διατριβής του

κ. Μπορμπουδάκη Γεωργίου

Doctoral Dissertation Defense

Mr. Giorgos Borboudakis

Την Τετάρτη, 21 Νοεμβρίου 2018 και ώρα 17:00 στην αίθουσα Τηλεδιάσκεψης Κ206 του Τμήματος Επιστήμης Υπολογιστών του Πανεπιστημίου Κρήτης στο Ηράκλειο, θα γίνει η δημόσια παρουσίαση και υποστήριξη της Διδακτορικής Διατριβής του υποψηφίου διδάκτορα του Τμήματος Επιστήμης Υπολογιστών κ. Γεωργίου Μπουρμπουδάκη με θέμα:

“Αποδοτική και ακριβής επιλογή μεταβλητών, με επεκτάσεις για πολλαπλές λύσεις και για δεδομένα μεγάλου όγκου”

“Efficient and Accurate Feature Selection, with Extensions for Multiple Solutions and to Big Data”

ΠΕΡΙΛΗΨΗ

Το πρόβλημα της επιλογής μεταβλητών μπορεί να οριστεί ως η ανακάλυψη ενός ελάχιστου υποσυνόλου των μεταβλητών εισόδου που είναι βέλτιστα προβλεπτικό για κάποια μεταβλητή ενδιαφέροντος. Η επιλογή μεταβλητών συνηθίζεται να χρησιμοποιείται σε αναλύσεις μηχανικής μάθησης και είναι βασικό εργαλείο όταν ο στόχος της ανάλυσης είναι η ανακάλυψη γνώσης. Αυτό είναι ιδιαίτερα σημαντικό σε τομείς όπως η μοριακή βιολογία και οι επιστήμες της ζωής, όπου ένας ερευνητής ενδιαφέρεται να κατανοήσει το πρόβλημα που μελετάει και όχι απαραίτητα για το προγνωστικό μοντέλο που προκύπτει.

Η επιλογή μεταβλητών είναι δύσκολη: έχει αποδειχθεί ότι είναι NP-σκληρή, και για αυτό οι περισσότεροι αλγόριθμοι είναι προσεγγιστικοί για να είναι υπολογιστικά αποδοτικοί. Υπάρχουν πολλές διαφορετικές προσεγγίσεις στο πρόβλημα επιλογής μεταβλητών, οι οποίες διαφέρουν στο πόσο γενικές είναι (π.χ. τι τύπους δεδομένων και μεταβλητών μπορούν να χειριστούν), στο υπολογιστικό τους κόστος, καθώς και στις θεωρητικές τους ιδιότητες. Οι μέθοδοι βηματικής επιλογής (stepwise selection) είναι αρκετά γενικές και βέλτιστες για μια μεγάλη κατηγορία πιθανοτικών κατανομών, αλλά είναι υπολογιστικά ακριβές. Οι μέθοδοι που βασίζονται σε αραιότητα (sparsity) (π.χ. LASSO) είναι υπολογιστικά αποδοτικές για ορισμένα προβλήματα (π.χ. ταξινόμηση και παλινδρόμηση) και χρονοβόρες για άλλα, (π.χ. δεδομένα χρόνου) και έχουν ισχυρές θεωρητικές εγγυήσεις. Οι προσεγγίσεις βασισμένες στη θεωρίας πληροφορίας είναι υπολογιστικά γρήγορες, αλλά όχι τόσο γενικές (χειρίζονται μόνο διακριτά δεδομένα) και με ασθενέστερες θεωρητικές εγγυήσεις. Μια άλλη πρόκληση είναι να κλιμακωθούν οι μέθοδοι επιλογής μεταβλητών για δεδομένα μεγάλου όγκου, τα οποία μπορεί να περιέχουν εκατομμύρια δείγματα και μεταβλητές. Οι υπάρχουσες προσεγγίσεις είτε είναι πολύ αργές, είτε έχουν κακή απόδοση ως προς την προβλεπτική τους ικανότητα. Τέλος, οι περισσότερες μέθοδοι δεν λαμβάνουν υπόψιν τους την παρουσία πολλαπλών λύσεων, οι οποίες συχνά υπάρχουν σε πραγματικά δεδομένα. Για παράδειγμα, είναι γνωστό πως τα μοριακά δεδομένα συχνά περιέχουν πολλαπλές λύσεις, πιθανώς λόγω του πλεονασμού που υπάρχει στο υποκείμενο βιολογικό σύστημα. Επομένως, παρόλο που ο εντοπισμός μιας λύσης είναι επαρκής για τον σκοπό της πρόβλεψης, δεν αρκεί για την ανακάλυψη γνώσης. Αντιθέτως, η αναφορά μίας και μόνης λύσης και ο ισχυρισμός πως δεν υπάρχουν άλλες λύσεις είναι παραπλανητική.

Για τη διπλωματική εργασία εστιάζουμε σε άπληστες μεθόδους επιλογής μεταβλητών τύπου «forward-backward» και προτείνουμε διάφορες επεκτάσεις για την αντιμετώπιση των παραπάνω προκλήσεων. Επιλέξαμε αυτή την κατηγορία μεθόδων λόγω των θεωρητικών ιδιοτήτων και της γενικότητάς τους. Δείχνουμε πως αλγόριθμοι διαφόρων κατηγοριών, όπως αυτοί που βασίζονται στην αραιότητα, στη θεωρία πληροφορίας, στη στατιστική ή στη θεωρία αιτιότητας, είναι ειδικές περιπτώσεις ή προσεγγίσεις μεθόδων βηματικής επιλογής. Αυτό επιτρέπει την μετάφραση και χρήση τεχνικών (όπως αυτές που προτείνονται σε αυτή τη διατριβή) μεταξύ διαφορετικών κατηγοριών αλγορίθμων. Στη συνέχεια, προτείνουμε ένα ευριστικό, εμπνευσμένο από αιτιατή μοντελοποίηση, για να επιταχύνουμε τον αλγόριθμο επιλογής forward-backward selection, διατηρώντας τις θεωρητικές ιδιότητές του. Σε υπολογιστικά πειράματα δείχνουμε ότι αυτό οδηγεί σε επιτάχυνση 1-2 τάξεων μεγέθους, διατηρώντας παράλληλα την προβλεπτική του ικανότητα. Στη συνέχεια, επεκτείνουμε τον αλγόριθμο για τις δεδομένα μεγάλου όγκου, επιτρέποντάς του να χειριστεί δεδομένα με δεκάδες εκατομμύρια δείγματα και μεταβλητές. Σε μια σύγκριση με αλγορίθμους από την ίδια αλγοριθμική οικογένεια, δείχνουμε ότι η προτεινόμενη μέθοδος περνά σημαντικά τον ανταγωνισμό όσον αφορά το χρόνο λειτουργίας, έχει την ίδια προβλεπτική ικανότητα, και είναι η μόνη μέθοδος που μπορεί να τερματίσει σε όλα τα σύνολα δεδομένων. Επιπλέον, σε μια σύγκριση με μεθόδους βασισμένες στη θεωρίας πληροφορίας, δείχνουμε ότι, αν και υπολογιστικά βραδύτερη, είναι σε θέση να παράγει σημαντικά καλύτερα

προγνωστικά μοντέλα. Τέλος, ασχολούμαστε με το πρόβλημα ανακάλυψης πολλαπλών λύσεων. Δείχνουμε ότι η υπάρχουσα ταξινόμηση χαρακτηριστικών είναι παραπλανητική όταν υπάρχουν πολλές λύσεις και προτείνουμε μια εναλλακτική ταξινόμηση που λαμβάνει υπόψη την ύπαρξη πολλαπλών λύσεων. Στη συνέχεια, εξετάζουμε αρκετούς ορισμούς της στατιστικής ισοδυναμίας συνόλων μεταβλητών, καθώς και μεθόδους ελέγχου της ισοδυναμίας συνόλων μεταβλητών. Έπειτα, προτείνουμε μια γενική λύση για την επέκταση των μεθόδων τύπου forward-backward για τον εντοπισμό πολλαπλών, στατιστικά ισοδύναμων λύσεων και παρέχουμε συνθήκες υπό τις οποίες είναι σε θέση να ανακαλύψει όλες τις ισοδύναμες λύσεις. Σε μια σύγκριση με τη μόνη εναλλακτική μέθοδο με τις ίδιες θεωρητικές εγγυήσεις, δείχνουμε ότι παράγει παρόμοια αποτελέσματα ενώ είναι υπολογιστικά ταχύτερη.

Επιβλέπων: Καθηγητής, Ιωάννης Τσαμαρδίνος

ABSTRACT

The problem of feature selection can be defined as identifying a minimal subset of the input variables that is optimally predictive for an outcome variable of interest. Feature selection is a common component in supervised machine learning pipelines, and an essential tool when the goal of the analysis is knowledge discovery. The latter is especially important in domains such as molecular biology and life sciences, where a researcher is interested in understanding the problem under study and not necessarily in the resulting predictive model.

Feature selection is challenging: it has been shown to be NP-hard, and thus most approaches rely on approximations to solve it efficiently. There exist many different approaches to the feature selection problem, trading off generality (e.g., what types of data and outcomes they can handle), computational cost, and theoretical properties of optimality. Stepwise selection methods (e.g., forward-backward selection) are quite general and optimal for a large class of distributions, but are computationally expensive. Sparsity-based methods (e.g., LASSO) are computationally efficient for some problems (e.g., classification and regression) and slow for others (e.g., time-course data), and have strong theoretical guarantees. Information theoretic approaches are computationally fast, but not as general (they only handle discrete data) and with weaker theoretical guarantees. Another challenge is to scale feature selection methods to very large datasets, which may contain millions of samples and variables. Existing approaches are either too slow or perform poorly in terms of predictive performance. Finally, most methods do not deal with the presence of multiple solutions to the feature selection problem, which often exist in real data. For example, it has been shown molecular data often contain multiple solutions, possibly due to the redundancy present in the underlying biological system. Therefore, although identifying a single solution is adequate for predictive purposes, it is not

sufficient when the focus is on knowledge discovery. On the contrary, reporting a single solution and claiming that there are no other solutions is misleading.

In this thesis, we focus on forward-backward selection and propose several extensions to tackle the above challenges. Forward-backward selection was chosen because of its theoretical properties and generality, as it is applicable to different predictive tasks and data types. We provide a unified view of several classes of feature selection methods, such as sparsity-based, information theoretic, statistical and causal-based methods, and show that they are instances or approximations of stepwise selection methods. This allows one to translate and use techniques (such as the ones proposed in this thesis) between different approaches to the feature selection problem. Then, we propose a heuristic inspired by causal modeling to speed-up the forward-backward selection algorithm, while preserving its theoretical properties. In experiments we show that this leads to a speed-up of 1-2 orders of magnitude over the standard forward-backward selection algorithm, while retaining its predictive performance. Afterwards, we extend the algorithm for Big Data settings, enabling it to scale to data with tens of millions of samples and variables.
In a comparison with alternative methods from the same algorithmic family, we show that the proposed method significantly outperforms all competitors in terms of running time, being the only method that is able to terminate on all datasets, and without sacrificing predictive performance. Furthermore, in a comparison with information theoretic methods we show that, although computationally slower, it is able to produce significantly better predictive models. Finally, we deal with the multiple feature selection problem. We show that the existing taxonomy of features is misleading when multiple solutions are present, and propose an alternative taxonomy that takes multiplicity into account. Then, we consider several definitions of statistical equivalence of feature sets, as well as methods to test for equivalence of feature sets. Afterwards, we propose a general strategy to extend forward-backward type methods for identifying multiple, statistically equivalent solutions. We provide conditions under which it is able to identify all equivalent solutions. In a comparison with the only alternative method with the same theoretical guarantees, we show that it produces similar results while being computationally faster.

Supervisor: Professor, Ioannis Tsamardinos