Υπ. Διδάκτωρ Μαρία Τσιακμάκη

Τίτλος: Νέοι αλγόριθμοι μηχανικής μάθησης για την επαγωγή γνώσης από εκπαιδευτικά/μαθησιακά δεδομένα.

Περίληψη στα ελληνικά

Στις μέρες μας, καθημερινά μεγάλος όγκος εκπαιδευτικών δεδομένων γίνεται όλο και πιο διαθέσιμος. Διάφορα εκπαιδευτικά πληροφοριακά συστήματα συλλέγουν και χρησιμοποιούν τεράστια σύνολα δεδομένων τόσο για τους εκπαιδευόμενους, το εκπαιδευτικό προσωπικό και το υλικό, όσο και για τις μεταξύ τους αλληλεπιδράσεις. Η επεξεργασία και ανάλυση αυτών των δεδομένων προσέλκυσε την προσοχή διαφόρων ερευνητών κυρίως για τη δυνατότητά της να εντοπίσει μοτίβα, τάσεις και προβλέψεις που μπορούν να χρησιμοποιηθούν για τη βελτίωση της μαθησιακής διαδικασίας και των αποτελεσμάτων της. Στον χώρο της Εκπαίδευσης, η Εξόρυξη Γνώσης από εκπαιδευτικά δεδομένα και η Μαθησιακή Αναλυτική συνιστούν πλέον δύο ιδιαίτερα δυναμικά διεπιστημονικά πεδία, τα οποία αναπτύσσονται σήμερα με ταχείς ρυθμούς. Απώτερος κοινός σκοπός τους είναι κατανόηση και βελτίωση της μάθησης και του εκπαιδευτικού περιβάλλοντος. Η πρόβλεψη των μαθησιακών αποτελεσμάτων συνιστά ένα από τα σημαντικότερα προβλήματα των πεδίων της Εξόρυξης Γνώσης από εκπαιδευτικά δεδομένα και της Μαθησιακής Αναλυτικής. Ανεξάρτητα από τον τρόπο προσέγγισης του προβλήματος, η επίλυσή του παρουσιάζει ιδιαίτερη πρακτική αξία, καθώς επιτρέπει την αποτελεσματική πρόβλεψη των μαθησιακών αποτελεσμάτων των σπουδαστών και δίνει την δυνατότητα έγκαιρης παρέμβασης για την παροχή υποστήριξης σε αυτούς. Παράλληλα, αποτελεί συνήθως κανόνα παρά εξαίρεση το γεγονός ότι τα διαθέσιμα δεδομένα του πραγματικού κόσμου ενδέχεται να περιέχουν ανακριβή στοιχεία, να είναι ελλιπή ή/και να χαρακτηρίζονται από υποκειμενικότητα. Ακόμα, κατά την επίλυση προβλημάτων, πολλές φορές είναι εξίσου απαραίτητο τα εξαγόμενα μοντέλα μάθησης να έχουν μία εύκολα κατανοητή αναπαράσταση και διαφάνεια του τρόπου λειτουργίας τους ώστε να επιτρέπεται η συνέργεια ανθρώπου και μηχανής. Την ίδια στιγμή, η ταχεία ανάπτυξη και η μεγάλη ποικιλία εξελιγμένων μεθόδων Μηχανικής Μάθησης απαιτούν εξειδικευμένες γνώσεις και εμπειρία στους σχετικούς τομείς από τους χρήστες τους, εμποδίζοντας, έτσι, τη συμμετοχή μη έμπειρων χρηστών.

Η παρούσα διδακτορική διατριβή πραγματεύεται το γενικότερο πρόβλημα της δημιουργίας προγνωστικών μοντέλων ταξινόμησης και παλινδρόμησης στον χώρο της εκπαίδευσης. Η βασική επιδίωξη είναι η ανάπτυξη νέων μεθοδολογιών υποστήριξης αξιόπιστων μοντέλων πρόβλεψης της απόδοσης των μαθητών σε ακαδημαϊκά μαθήματα εξαμήνου της ανώτατης εκπαίδευσης. Σε αυτό το πλαίσιο, περιγράφονται έξι ερευνητικές μελέτες που πραγματεύονται νέες μεθόδους Εξόρυξης Γνώσης από εκπαιδευτικά δεδομένα και επικεντρώνονται σε δύο κεντρικά ζητήματα του χώρου: τη διαχείριση της ατελούς πληροφορίας που υπάρχει εν γένει στα ψηφιακά δεδομένα (εκπαιδευτικά και μη) και την αξιοποίηση τεχνικών που εξάγουν μοντέλα υψηλής απόδοσης με διαισθητικά ερμηνεύσιμη δομή. Καθώς, στην Εξόρυξη Γνώσης, η πολυπλοκότητα των διαθέσιμων επιλογών που συνοδεύουν τις μεθόδους έχει ως αποτέλεσμα τον αποκλεισμό των μη ειδικών στον χώρο, η υποστήριξη της ερμηνευσιμότητας είναι σημαντική. Για τον σκοπό αυτόν, χρησιμοποιήθηκαν και συνδυάστηκαν προοδευτικά ασαφείς αλγόριθμοι μάθησης, τεχνικές αυτόματης βελτιστοποίησης και στρατηγικές Ενεργούς Μάθησης, ενώ μελετήθηκε ξεχωριστά η τεχνική της Μεταφοράς Μάθησης. Σε ένα πρώτο βήμα, διερευνήθηκε η αποτελεσματικότητα της Μπεϋζιανής βελτιστοποίησης στην αυτόματη επιλογή του βέλτιστου αλγορίθμου μάθησης και στην ρύθμιση των υπερ-παραμέτρων του. Για να τονιστεί εξίσου η σημασία της παραγωγής ερμηνεύσιμων και επεξηγήσιμων μοντέλων μηχανικής μάθησης, η αναζήτηση λύσεων περιορίστηκε σε αυτούς τους αλγορίθμους που βασίζονται σε δέντρα ή κανόνες. Αποτέλεσμα της σύμπραξης αυτής ήταν η δημιουργία μοντέλων με κατανοητή από τον άνθρωπο αναπαράσταση, τα οποία ταυτόχρονα χαρακτηρίζονται από υψηλή ακρίβεια.

Επιπλέον προτείνεται η αξιοποίηση στρατηγικών Ενεργούς Μάθησης και Μεταφοράς Μάθησης, τεχνικές που επικεντρώνονται στην αντιμετώπιση του προβλήματος της έλλειψης επαρκών δεδομένων για την εκπαίδευση ενός μοντέλου μάθησης. Η Ενεργή Μάθηση αφορά κυρίως προβλήματα όπου, ενώ είναι διαθέσιμα αρκετά δεδομένα, μόνο για λίγα από αυτά είναι γνωστή η ετικέτα τους. Η μεταφορά μάθησης αφορά κυρίως προβλήματα όπου δεν είναι διαθέσιμα αρκετά δεδομένα εκπαίδευσης για ένα πρόβλημα, αλλά είναι διαθέσιμα για ένα άλλο, σχετικό πρόβλημα. Αναπτύχθηκαν δύο και καινοτόμες μεθοδολογίες: μια υβριδική μέθοδος Ενεργούς Μάθησης βασισμένη σε ασαφείς ταξινομητές που ενσωματώνει μηχανισμούς αυτόματης βελτιστοποίησης και μία μέθοδος μεταφοράς γνώσης μεταξύ προγνωστικών μοντέλων μάθησης. Οι νέοι προτεινόμενες μέθοδοι εφαρμόζονται σε προβλήματα ταξινόμησης και παλινδρόμησης με τη χρήση εκπαιδευτικών δεδομένων που περιγράφουν τη δραστηριότητα των φοιτητών κυρίως στη διαδικτυακή πλατφόρμα

μάθησης Moodle. Η πειραματική διερεύνηση τεκμηριώνει την αποτελεσματικότητα των μεθόδων αυτών για το πρόβλημα της πρόβλεψης των μαθησιακών αποτελεσμάτων.

Η συστηματική έρευνα που πραγματοποιήθηκε ελπίζουμε να αποτελεί μια χρήσιμη συμβολή και να ανοίγει νέους ορίζοντες στο πεδίο της Εξόρυξης Γνώσης από εκπαιδευτικά δεδομένα και της Μαθησιακής Αναλυτικής, οδηγώντας στην αξιοποίηση και την ανάπτυξη εργαλείων που υποβοηθούν το έργο του εκπαιδευτή και της μάθησης μέσω απτών δεδομένων.

Περίληψη στα Αγγλικά

Nowadays, large amount of educational data is becoming more and more available. Most of these data emanate from institutional student information systems, virtual learning environments, attendance monitoring systems and library systems. These systems can record any student activity that is supported, such as reading, writing, exams taking, tasks performed and peer communications, information on staff, content and the institution and so forth. The analysis of these datasets attracted much attention among researchers for its potential to identify patterns, trends and predictions that can be used to optimize the learning process and its outcomes. In the field of Education, Educational Data Mining and Learning Analytics have expanded rapidly, emerging as dynamic interdisciplinary fields of study. Their ultimate common goal is to understand and improve the learning process and environment.

Predicting students' progression and learning outcomes is one of the most important task in the field of Educational Data Mining and Learning Analytics. Regardless of how the problem is being approached, its solution is of high practical value, as it provides information on which students need additional support or are at risk of dropping out. With a view to encourage and motivate those students, remedial actions could be organized, such as early alerts and advising interventions.

In real-world problems, it is more of a rule than an exception that the available data may contain inaccurate data, are incomplete and/or are appeal to subjectivity. It is also often necessary the exported learning models to have an easy-to-understand representation and transparency of how they work, in order to forge a synergy between humans and machines. At the same time, the rapid development and the wide variety of advanced Machine Learning solutions require specialized knowledge and experience in the relevant fields from their users, thus preventing the participation of inexperienced ones.

This dissertation deals with the general problem of creating predictive classification and regression models in the field of Education. The main objective is to develop methodologies that efficiently construct reliable learning models for predicting the performance of students in the final examinations of semester courses of higher education.

To this end, six new research studies are described which provide new methods of extracting knowledge from educational data. Regardless of their individual differences, the studies were based upon two basic principles: the handling of incomplete information that generally exists in digital data (educational and not) and the employment of techniques that produce high-performance models with intuitively interpretable structure, without excluding non-specialists in the field due to the complexity of the available options that usually surround them.

In this context, Fuzzy Learning Algorithms, Automated Machine Learning techniques and Active Learning strategies were employed and progressively combined, while the Transfer Learning technique was separately studied. In a first step, the effectiveness of Bayesian optimization in the automatic selection of the optimal learning algorithm and in the adjustment of their hyper-parameters was investigated. To equally highlight the importance of producing interpretable and explainable Machine Learning models, the search space was restricted to tree-based and rule-based algorithms. The result of this consolidation was the formation of models with human-understandable representation, which were also characterized by high accuracy.

In addition, the use of Active Learning and Transfer Learning strategies is proposed. These techniques are focused on addressing the difficulty of collecting sufficient data that are required to fit a learning model. Active Learning is well-motivated in problems, where unlabeled data may be abundant or easy to come by, but training labels are difficult, or expensive to obtain. Transfer Learning is well-motivated in problems, where there is a lack of data regarding a problem, but there is plenty of data about another related one. Two distinct and innovative methodologies have been developed: a hybrid Active Learning method based on Fuzzy classifiers that incorporates automated optimization mechanisms and a Transfer Learning method between predictive learning models.The new proposed methods are applied to classification and regression problems using mainly educational data describing the student activity on the Moodle online learning platform. The experimental results

document the effectiveness of these methods for the problem of predicting students' learning outcomes.

Hopefully, the systematic research that has been conducted could be considered as a useful contribution in the fields of Educational Data Mining and Learning Analytics. At the same time, it could lead to the development of tools that assist instructors and learning through data-driven decisions.