ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΕΡΓΑΣΤΗΡΙΟ ΓΛΩΣΣΩΝ ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΥ & ΤΕΧΝΟΛΟΓΙΑΣ ΛΟΓΙΣΜΙΚΟΥ

Διπλωματικές 2012-2013

στην περιοχή της Μηχανικής Μάθησης

1. Μέθοδοι Παλινδρόμησης, Εφαρμογή στην Πρόβλεψη Τιμών

Οι μέθοδοι παλινδρόμησης χρησιμοποιούνται σε πλήθος εφαρμογών όπου στόχος είναι η πρόβλεψη της τιμής μίας συνεχούς μεταβλητής. Παραδείγματα προβλημάτων όπου εφαρμόζονται με επιτυχία μέθοδοι παλινδρόμησης αποτελούν η πρόβλεψη της πιστοληπτικής ικανότητας πελατών σε τράπεζες, η πρόβλεψη των τιμών των μετοχών στο χρηματιστήριο, ακόμη και η μέτρηση μικροσκοπικών αλλαγών στην ελλειπτικότητα των γαλαξιών με σκοπό τον εντοπισμό της σκοτεινής ύλης από τη NASA. Σε κάποια από αυτά τα προβλήματα, υπάρχουν περισσότερες από μία μεταβλητές τις οποίες θέλουμε να προβλέψουμε ταυτόχρονα και οι οποίες ενδέχεται να παρουσιάζουν συσχετίσεις μεταξύ τους (π.χ. τιμές τραπεζικών μετοχών, αεροπορικών εισιτηρίων, κ.λ.π.). Σκοπός της διπλωματικής είναι η μελέτη της σχετικής βιβλιογραφίας και η πειραματική αξιολόγηση υπαρχουσών αλλά και καινοτόμων μεθόδων παλινδρόμησης οι οποίες εκμεταλλεύονται τις συσχετίσεις μεταξύ των πολλαπλών μεταβλητών στόχων σε διάφορα σύνολα δεδομένων.

Απαραίτητα προσόντα: Καλή γνώση Java και Αγγλικών.

Επικοινωνία:

Ελευθέριος Σπυρομήτρος, e-mail: espyromi@csd.auth.gr, www: http://users.auth.gr/espyromi

2. Ταξινόμηση Ροών Δεδομένων, Εφαρμογή στην Ταξινόμηση Ηλεκτρονικών Εγγράφων

Η ανάπτυξη αλγορίθμων κατηγοριοποίησης δεδομένων τα οποία μπορεί να ανήκουν ταυτόχρονα σε παραπάνω από μία κατηγορίες έχει γνωρίσει εξαιρετική άνθηση τα τελευταία χρόνια εξαιτίας της πληθώρας των εφαρμογών που έχουν να κάνουν με την αυτόματη επισήμανση δεδομένων πολλαπλών ετικετών όπως εικόνες, ειδησεογραφικά άρθρα, μουσικά κομμάτια κ.α. Στην παρούσα διπλωματική θα εστιάσουμε σε αλγορίθμους μάθησης πολλαπλών ετικετών οι οποίοι θα έχουν τη δυνατότητα να διαχειριστούν πολύ μεγάλο όγκο δεδομένων, να παράγουν προβλέψεις σε πραγματικό χρόνο και να προσαρμόζονται σε τυχόν αλλαγές της κατανομής των δεδομένων. Ιδιαίτερο βάρος θα δοθεί στην υλοποίηση ή/και επέκταση αλγορίθμων από τη διεθνή βιβλιογραφία και την πειραματική τους αξιολόγηση.

Απαραίτητα προσόντα: Καλή γνώση Java και Αγγλικών.

Επιθυμητά προσόντα: Γνώση του weka.

Επικοινωνία:

Ελευθέριος Σπυρομήτρος, e-mail: espyromi@csd.auth.gr, www: http://users.auth.gr/espyromi

3. Ανάπτυξη Δυναμικών Μεθόδων Παραγωγής Συστάσεων

Η παραγωγή συστάσεων παίζει σημαντικό ρόλο σαν εργαλείο προώθησης προϊόντων σε ηλεκτρονικά καταστήματα καθώς επιτρέπει στους χρήστες να εντοπίσουν ευκολότερα και

να αποφασίσουν τι θα αγοράσουν. Η διπλωματική αυτή θα στηριχθεί και θα επεκτείνει μέθοδο παραγωγής συστάσεων η οποία αναπτύχθηκε από μέλη της ομάδας Μηχανικής Μάθησης και Ανακάλυψης Γνώσης και η οποία έχει διακριθεί στο διεθνή διαγωνισμό Data Mining Cup 2011 (http://www.data-mining-cup.de/en/). Η παραγωγή συστάσεων θα βασίζεται σε δεδομένα clickstream και ιδιαίτερο βάρος θα δοθεί στην κλιμάκωση σε δεδομένα μεγάλου όγκου και την προσαρμογή σε τυχόν αλλαγές των προτιμήσεων των χρηστών.

Απαραίτητα προσόντα: Καλή γνώση Java και Αγγλικών.

Επιθυμητά προσόντα: Γνώση θεωρίας γράφων.

Επικοινωνία:

Ελευθέριος Σπυρομήτρος, e-mail: espyromi@csd.auth.gr, www: http://users.auth.gr/espyromi

4. Ανακάλυψη Γνώσης από Αλληλουχίες MicroRNA

Η πρόσφατη ανακάλυψη της λειτουργίας μια μικρής αλληλουχίας RNA (microRNA ή miRNA) έχει προβάλει νέα μεγάλα ερωτήματα στην επιστημονική κοινότητα. Ένα μόριο miRNA μπορεί να παρέμβει στη διαδικασία της πρωτεϊνοσύνθεσης και να καταστείλει την έκφραση κάποιων γονιδίων ενός οργανισμού. Αυτή η δράση του miRNA ενδέχεται να σχετίζεται με την εκδήλωση διάφορων ασθενειών, γεγονός που έχει προκαλέσει μεγάλο ερευνητικό ενδιαφέρον. Η χρήση τεχνικών μηχανικής μάθησης μπορεί να βοηθήσει στην ανάλυση τέτοιων δεδομένων και να οδηγήσει στην ανακάλυψη νέας γνώσης. Σκοπός της διπλωματικής εργασίας είναι η ανάπτυξη υπολογιστικών μεθόδων και εργαλείων για την ανάλυση δεδομένων με στόχο την κατανόηση της λειτουργίας των miRNA.

Απαραίτητα προσόντα: Καλή γνώση Java και Αγγλικών και ενδιαφέρον για βιολογικά θέματα.

Επικοινωνία: Ιωάννης Καβακιώτης, e-mail: ikavak@csd.auth.gr

5. Η Μηχανική Μάθηση στη Διαδικτυακή Διαφήμιση

Η Διαδικτυακή Διαφήμιση αποτελεί σήμερα την ταχύτερα αναπτυσσόμενη μέθοδο διαφήμισης παγκοσμίως. Νέες εξελιγμένες τεχνολογίες έχουν ώς σκοπό τους την ευφυή-στοχευμενη προβολή διαφήμισης στο κοινό. Αυτές οι τεχνολογίες βασίζονται στο μεγάλο όγκο δεδομένων που συλλέγονται, καθιστώντας το αντικείμενο κατάλληλο για την εφαρμογή τεχνικών Μηχανικής Μάθησης. Σκοπός της διπλωματικής εργασίας είναι καταρχήν, η μελέτη της χρήσης τεχνικών Μηχανικής Μάθησης στα συστήματα παροχής διαδικτυακής διαφήμισης σήμερα. Επίσης, θα γίνει ανάπτυξη περιβάλλοντος δοκιμών και εξομοίωσης για τεχνικές Διαδικτυακής Διαφήμισης, σε πραγματικά δεδομένα, καθώς και η υλοποίηση στη συνέχεια κατάλληλου αλγορίθμου Μηχανικής Μάθησης με σκοπό την βέλτιστη και ευφυή παροχή διαφήμισης.

Απαιτούμενα Προσόντα: Γνώσεις προγραμματισμού σε Java ή C/C++

Επικοινωνία: Ανέστης Φαχαντίδης, e-mail: afa@csd.auth.gr, www: http://users.auth.gr/afa

6. Χρήση Προχωρημένων Τεχνικών Ενισχυτικής Μάθησης σε Πολυ-Πρακτορικά Συστήματα

Η Ενισχυτική Μάθηση σε πολυ-πρακτορικά συστήματα δεν αποτελεί απλή αναγωγή της κλασσικής μονο-πρακτορικής ενισχυτικής μάθησης. Οι θεωρητικές προκλήσεις που τίθενται αλλά και οι πιθανές εφαρμογές της νέας συτής γνωστικής περιοχής είναι πολλές και σημαντικές. Σκοπός της διπλωματικής είναι κατ' αρχήν, η μελέτη προχωρημένων τεχνικών ενισχυτικής μάθησης όπως η Ιεραρχική Ενισχυτική Μάθηση και η Μεταφορά Μάθησης, στο πλαίσιο των πολυ-πρακτορικών συστημάτων. Επίσης, σκοπός της διπλωματικής είναι και η πρόταση ενός νέου αλγορίθμου πολυ-πρακτορικής Ενισχυτικής Μάθησης που μπορεί να προέλθει από την εξέλιξη και προσαρμογή υπαρχόντων αλγορίθμων από την μονο-πρακτορική Ενισχυτική Μάθηση. Ως πεδίο εφαρμογής του

νέου αλγορίθμου θα χρησιμοποιηθεί το Robocup Soccer Simulation ή/και υπο-προβλήματα αυτού, με σκοπό την επίδειξη βελτιωμένων αποτελεσμάτων.
Απαιτούμενα Προσόντα: Γνώσεις προγραμματισμού σε Java ή C/C++
Επικοινωνία: Ανέστης Φαχαντίδης, e-mail: afa@csd.auth.gr, www: http://users.auth.gr/afa

7. Μάθηση από Δεδομένα Πολλαπλών Ετικετών

Δεδομένα πολλαπλών ετικετών (multi-label data), ονομάζουμε τα δεδομένα εκείνα τα οποία έχουν σημανθεί με μία ή παραπάνω ετικέτες (labels ή tags) από ένα πεπερασμένο σύνολο ετικετών. Τα δεδομένα αυτά διαφέρουν από τα δεδομένα που χρησιμοποιούνται στο κλασικό πρόβλημα της ταξινόμησης (classification), όπου κάθε δεδομένο ανήκει σε μία και μόνο κλάση από ένα πεπερασμένο σύνολο κλάσεων. Για παράδειγμα, πολλά τραγούδια των Scorpions θα μπορούσαν να σημανθούν τόσο με την ετικέτα «ροκ», όσο και με την ετικέτα «μπαλάντα», αλλά και με πολλές άλλες ετικέτες που θα μπορούσαν να αφορούν το συναίσθημα του κομματιού, τη γλώσσα των στίχων, τα μουσικά όργανα που χρησιμοποιήθηκαν στη σύνθεση, κ.α. Τα τελευταία χρόνια, η μάθηση από δεδομένα πολλαπλών ετικετών (multi-label learning) είναι ένα πρόβλημα που παρουσιάζει πολύ μεγάλο ενδιαφέρον, επειδή ανακύπτει σε πολλές ενδιαφέρουσες εφαρμογές όπως στην ανάλυση δεδομένων κειμένου (ιστοσελίδες, άρθρα σε blogs, κ.α.), βιολογικών δεδομένων (λειτουργία πρωτεϊνών), εικόνων (σημασιολογική σήμανση) και μουσικής (ταξινόμηση τραγουδιών σε συναισθήματα). Η αναγκαιότητα χρήσης τεχνικών μάθησης οφείλεται κυρίως στο ότι οι συλλογές δεδομένων (εικόνων, μουσικής, κειμένων, βιολογικών δεδομένων κτλ.) έχουν στις μέρες μας πολύ μεγάλο μέγεθος, και η χειροκίνητη σήμανση τους με ετικέτες έχει μεγάλο χρονικό και οικονομικό κόστος. Απαιτούνται λοιπόν τεχνικές που από ένα μικρό σύνολο δεδομένων, για το οποίο οι ετικέτες είναι γνωστές, μαθαίνουν ένα μοντέλο, το οποίο μπορεί στη συνέχεια να παράγει αυτόματα τις ετικέτες για τα υπόλοιπα δεδομένα, τα οποία δεν έχουν σημανθεί.

Η ομάδα Μηχανικής Μάθησης και Ανακάλυψης Γνώσης (http://mlkd.csd.auth.gr) έχει αναπτύξει σε γλώσσα Java το λογισμικό Mulan (http://mulan.sourceforge.net), ένα ανοιχτού κώδικα λογισμικό για μάθηση από δεδομένα πολλαπλών ετικετών, το οποίο χρησιμοποιείται διεθνώς από ερευνητές αλλά και από απλούς χρήστες εργαλείων ανάλυσης δεδομένων.

Στο παραπάνω πλαίσιο, προτείνονται οι εξής πτυχιακές εργασίες:

(α) *Scalable multi-label learning*. Ανάπτυξη τεχνικών μάθησης από τεράστιου όγκου δεδομένα πολλαπλών ετικετών χρησιμοποιώντας είτε παράλληλο προγραμματισμό σε νέας τεχνολογίας επεξεργαστές και κάρτες γραφικών (π.χ. τεχνολογία CUDA), είτε τη φιλοσοφία Map/Reduce πάνω από την υποδομή του Apache Hadoop όπως γίνεται στη βιβλιοθήκη Apache Mahout (http://mahout.apache.org/).

(β) Multi-label learning and background knowledge. Ανάπτυξη τεχνικών μηχανικής μάθησης, οι οποίες μαθαίνουν αυτόματα και στη συνέχεια αξιοποιούν περιορισμούς μεταξύ των ετικετών (π.χ. σχέσεις γονέα-παιδιού, σχέσεις αποκλειστικής διάζευξης, σχέσεις άρνησης) με στόχο η μάθηση να γίνεται πιο αποτελεσματικά και πιο αποδοτικά.

(γ) Stratification in multi-label learning. Ανάπτυξη τεχνικών δειγματοληψίας από δεδομένα πολλαπλών ετικετών με στόχο την καλύτερη αντιπροσώπευση ετικετών και συνδυασμών ετικετών

στο δείγμα. Η διπλωματική αυτή θα στηριχθεί και θα επεκτείνει τη δουλειά που περιγράφεται στη δημοσιευμένη εργασία: http://mlkd.csd.auth.gr/publication_details.asp?publicationID=346

Επιθυμητά προσόντα: Καλή γνώση προγραμματισμού σε Java (ή/και C για την εργασία α).
Επικοινωνία: Γρηγόρης Τσουμάκας, e-mail: greg@csd.auth.gr, www: http://users.auth.gr/greg

Οι ενδιαφερόμενοι μπορούν να επικοινωνήσουν με το διδάσκοντα του μαθήματος της Μηχανικής Μάθησης για να εκδηλώσουν το ενδιαφέρον τους για κάποιο-α από τα θέματα.

Περισσότερες λεπτομέρειες για τα ερευνητικά ενδιαφέροντα της ομάδας μηχανικής μάθησης και ανακάλυψης γνώσης (MLKD) θα βρείτε στη διεύθυνση http://mlkd.csd.auth.gr.