**ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ**
**ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ**

**ΕΡΓΑΣΤΗΡΙΟ ΓΛΩΣΣΩΝ ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΥ & ΤΕΧΝΟΛΟΓΙΑΣ ΛΟΓΙΣΜΙΚΟΥ**

# Διπλωματικές 2013-2014

***στην περιοχή της Μηχανικής Μάθησης***

## 1. Μέθοδοι παλινδρόμησης πολλαπλών στόχων, εφαρμογή σε δεδομένα χρονοσειρών

Οι μέθοδοι παλινδρόμησης (regression) είναι μια κατηγορία στατιστικών μεθόδων που χρησιμοποιούνται ευρέως στην πρόβλεψη τιμών. Παραδείγματα προβλημάτων όπου βρίσκουν εφαρμογή οι μέθοδοι παλινδρόμησης αποτελούν η πρόβλεψη της ωριαίας παραγωγής ηλεκτρικής ενέργειας από αιολικά πάρκα [1] ή του ενεργειακού φόρτου σε σταθμούς παραγωγής ηλεκτρικής ενέργειας [2]. Σε τέτοιου είδους προβλήματα τα δεδομένα έχουν τη μορφή χρονοσειράς (time series) και απαιτείται κατάλληλος μετασχηματισμός των δεδομένων για την εφαρμογή των παραδοσιακών μεθόδων παλινδόμησης. Επιπλέον, σε πολλά προβλήματα αυτού του είδους απαιτείται η ταυτόχρονη πρόβλεψη των τιμών πολλών και συσχετισμένων μεταξύ τους μεταβλητών (π.χ. τιμές τραπεζικών μετοχών). Σκοπός της διπλωματικής, είναι η μελέτη της σχετικής βιβλιογραφίας στην πρόβλεψη χρονοσειρών και η εφαρμογή καινοτόμων μεθόδων παλινδρόμησης για πολλαπλές μεταβλητές που έχουν αναπτυχθεί από την ομάδα Μηχανικής Μάθησης και Ανακάλυψης Γνώσης σε δεδομένα με χαρακτηριστικά χρονοσειράς (π.χ [1],[2]).

**Σύνδεσμοι:**
[1] http://www.kaggle.com/c/GEF2012-wind-forecasting
[2] http://www.kaggle.com/c/global-energy-forecasting-competition-2012-load-forecasting).

**Απαραίτητα προσόντα**: Καλή γνώση προγραμματισμού (Java) και Αγγλικών.

**Επικοινωνία:** Λευτέρης Σπυρομήτρος, espyromi@csd.auth.gr, http://users.auth.gr/espyromi

## 2. Μέθοδοι ταξινόμησης πολλαπλών ετικετών, εφαρμογή στην ταξινόμηση ροών 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 προγραμματισμού (Java) και Αγγλικών.

**Επικοινωνία:** Λευτέρης Σπυρομήτρος, espyromi@csd.auth.gr, 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 προγραμματισμού (Java) και Αγγλικών.

**Επικοινωνία:** Λευτέρης Σπυρομήτρος, espyromi@csd.auth.gr, http://users.auth.gr/espyromi

4. **Χρήση Μηχανικής Μάθησης για ευφυή παροχή Διαδικτυακής Διαφήμισης**

Η Διαδικτυακή Διαφήμιση αποτελεί σήμερα την ταχύτερα αναπτυσσόμενη μέθοδο διαφήμισης παγκοσμίως. Νέες εξελιγμένες τεχνολογίες έχουν ώς σκοπό τους την ευφυή - στοχευμένη 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 τη βέλτιστη και ευφυή παροχή διαφήμισης.

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

**Σχετικό υλικό:**

http://research.microsoft.com/en-us/um/beijing/events/mload-2010/mload2010.pdf

http://pages.stern.nyu.edu/~fprovost/Papers/MLOAD.pdf

http://chercheurs.lille.inria.fr/~gabillon/publications/internship/GabillonInternshipReport.pdf

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

5. **Μέθοδοι Ενισχυτικής Μάθησης για εφαρμογή στον αυτόματο έλεγχο παρκαρίσματος ενός οχήματος**

Η Ενισχυτική Μάθηση αποτελεί μια οικογένεια τεχνικών Μηχανικής Μάθησης και έχει ως σκοπό την μάθηση μίας βέλτιστης πολιτικής δράσεων ενός πράκτορα που κινείται σε περιβάλλοντα περιορισμένης πληροφορίας. Σκοπός της διπλωματικής εργασίας είναι κατ' αρχήν, η μελέτη των προχωρημένων τεχνικών ενισχυτικής μάθησης με χρήση μοντέλου, για μέγιστη αποδοτικότητα δεδομένων. Στη συνέχεια, θα αναπτυχθεί μια εφαρμογή εξομοίωσης ενός οχήματος που επιχειρεί να παρκάρει αυτόνομα μεταξύ δύο σταθμευμένων οχημάτων. Δεν απαιτείται υλοποίηση σύνθετου ή ρεαλιστικού γραφικού περιβάλλοντος (απλή δι-διαστατη αναπαράσταση). Το εικονικό όχημα θα διαθέτει ελεγκτή αυτόνομου παρκαρίσματος που έχει εκπαιδευτεί με τεχν. Ενισχυτικής Μάθησης. Η γνώση που διαθέτει σε κάθε στιγμή το όχημα είναι αποκλειστικά τα δεδομένα από τους 5 αισθητήρες κίνησης που υπάρχουν σε κάθε προφυλακτήρα. Ο ελεγκτής που θα εκπαιδευτεί θα έχει τη δυνατότητα να προσαρμοστεί σε διάφορα μεγέθη του οχήματος που ελέγχει καθώς και σε διάφορες διαστάσεις του ελεύθερου χώρου parking.

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

**Σχετικό υλικό:** http://j.mp/rlparking

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

## 6. Ανάπτυξη μεθόδου Ενισχυτικής Μάθησης για Συστήματα συστάσεων

Τα συστήματα συστάσεων (recommender systems) δίνουν την δυνατότητα σε διαδικτυακές υπηρεσίες πωλήσεων να παρέχουν προσωποποιημένες προτάσεις αγοράς στους πελάτες τους. Σκοπός ενός τέτοιου συστήματος είναι η μάθηση μοντέλων επιθυμίας των πελατών είτε βάση της ομαδοποίησης των προσωπικών τους χαρακτηριστικών είτε βάσει ομαδοποίησης των προϊόντων. Σκοπός της διπλωματικής εργασίας είναι κατ' αρχήν, η μελέτη των υπαρχουσών μεθόδων συστημάτων συστάσεων και η πρόταση καινοτόμου επέκτασης τους με χρήση Ενισχυτικής Μάθησης. Η μέθοδος που θα προκύψει θα δοκίμαστει στην πλατφόρμα ηλ. καταστήματος - ανοιχτού κώδικα Magento, με σκοπό την ανάπτυξη αντιστοίχου plug-in και την δοκιμή με εικονικά δεδομένα. Πέρα από τα σημαντικά γνωστικά εφόδια που θα αποκτήσει ο φοιτητής σε τεχνολογίες αιχμής της Μηχανικής Μάθησης, η συγκεκριμένη διπλωματική θα του παρέχει και την εμπειρία ανάπτυξης για την πιο διαδεδομένη πλατφόρμα ηλεκτρονικού εμπορίου, Magento.

**Απαιτούμενα Προσόντα:** Γνώσεις προγραμματισμού σε PHP, προαιρετικά γνώση Python.

**Σχετικό υλικό:**

http://j.mp/recommendersystems

https://www.magentocommerce.com

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

## 7. Πρόβλεψη Τοξικότητας

Η περιβαλλοντική ρύπανση και η έκθεση σε τοξικές και επικίνδυνες χημικές ουσίες, φυσικούς παράγοντες (π.χ. ακτινοβολία) και παθογόνους οργανισμούς είναι γνωστό ότι μπορούν να προκαλέσουν φθορές, νοσηρότητα και θνησιμότητα στα διάφορα βιολογικά συστήματα. Οι επιστημονικοί τομείς που αναπτύχθηκαν τις τελευταίες δεκαετίες για την έρευνα των ποικίλων τοξικών επιδράσεων, κυρίως των χημικών ουσιών, καλύπτονται κάτω από τον όρο της επιστήμης της Τοξικολογίας, με ιδιαίτερη έμφαση στα τοξικολογικά προβλήματα του ανθρώπου. Ο στόχος της Τοξικολογίας Πρόβλεψης (Predictive Toxicology) είναι να προβλέψει με ακρίβεια τις αρνητικές επιπτώσεις των χημικών ουσιών κατά την απουσία πειραματικών δεδομένων. Οι In Silico τεχνικές πρόβλεψης τοξικότητας είναι πολύ γρήγορες και πολύ φτηνές σε σχέση με τις In Vivo και In Vitro τεχνικές καθώς δεν απαιτούν ούτε χημικά υλικά, ούτε φυσικές ενώσεις. Για τον λόγο αυτό είναι ιδανικές και ως επί το πλείστον εφαρμόσιμες σε όλες τις περιπτώσεις όπου πρέπει να γίνει εκτίμηση τοξικότητας γρήγορα και/ή με περιορισμένους πόρους, όπως για παράδειγμα μια πρώτη εκτίμηση τοξικότητας σε πιθανές ουσίες φαρμάκων. Στόχος της παρούσας εργασίας είναι η εφαρμογή μεθόδων μηχανικής μάθησης σε πραγματικά σύνολα δεδομένων με στόχο την πρόβλεψη τοξικότητας.

**Απαραίτητα Προσόντα:** Γνώσεις Προγραμματισμού σε Java και ενδιαφέρον για βιολογικά θέματα

**Σχετικό Υλικό:**

http://www.predictive-toxicology.org/

https://openaccess.leidenuniv.nl/bitstream/handle/1887/12954/Chapter%205.pdf?sequence=8

**Επικοινωνία:** Ιωάννης Καβακιώτης, e-mail: ikavak@csd.auth.gr,
www: http:// sites.google.com/site/ikavakiotis/

## 8. Εξόρυξη Γνώσης από Επιστημονικές Δημοσιεύσεις

Καθημερινά προστίθεται μεγάλος όγκος επιστημονικών δημοσιεύσεων στις ψηφιακές βιβλιοθήκες εκδοτικών οίκων (π.χ. SpringerLink, ScienceDirect), επιστημονικών οργανώσεων (π.χ. ACM digital library, IEEE explore), εμπορικών μηχανών αναζήτησης (π.χ. Google Scholar) και αποθετηρίων ανοιχτής πρόσβασης (π.χ. arXiv.org, CiteSeerX). Η διπλωματική αυτή έχει ως στόχο την ανάπτυξη τεχνικών εξόρυξης χρήσιμης γνώσης από τέτοιες ψηφιακές βιβλιοθήκες επιστημονικών δημοσιεύσεων με σκοπό την απάντηση ερωτημάτων όπως: α) Ποιες είναι οι επιστημονικές περιοχές που αναπτύσσονται και ποιες εκείνες που φθίνουν; β) ποιοι είναι οι κορυφαίοι ερευνητές σε μια επιστημονική περιοχή; γ) Ποια δημοσίευση οδήγησε στην άνθιση μιας νέας περιοχής; δ) Ποιες δημοσιεύσεις σχετίζονται σημασιολογικά με κάποια συγκεκριμένη ερευνητική περιοχή ή ερευνητικό αντικείμενο (ακόμα και αν δεν περιέχουν συγκεκριμένο keyword στο κείμενο τους); Κυρίως μας ενδιαφέρει η σημασιολογική κατηγοριοποίηση των δημοσιεύσεων σε πολλαπλές κατηγορίες (ερώτημα δ), αλλά η διπλωματική θα μπορούσε να εστιάσει και σε κάποιο από τα υπόλοιπα ερωτήματα, ή ακόμα και σε κάποιο άλλο ερώτημα που δεν αναφέρεται. Επιπλέον, κυρίως μας ενδιαφέρει η αξιοποίηση της πληροφορίας του **ετερογενούς γράφου** που σχηματίζεται με τις δημοσιεύσεις ως κόμβους και τις εξής ετερογενείς σχέσεις: α) κοινοί συγγραφείς μεταξύ δημοσιεύσεων, β) αναφορά προς μια δημοσίευση εντός μιας άλλης δημοσίευσης, γ) κοινές λέξεις μεταξύ δημοσιεύσεων.

**Σχετικό υλικό:**

http://link.springer.com/article/10.1007/s13278-011-0034-8
http://www.cs.bris.ac.uk/Publications/Papers/2001525.pdf
http://www.readcube.com/articles/10.1186/1471-2105-9-525

**Επικοινωνία:** Γρηγόρης Τσουμάκας, e-mail: greg@csd.auth.gr, www: http://users.auth.gr/greg

## 9. Μάθηση από Δεδομένα Πολλαπλών Ετικετών που έχουν Αποκτηθεί με Crowdsourcing

Το crowdsourcing είναι μια φθηνή λύση για τον σημασιολογικό χαρακτηρισμό πολυμέσων (π.χ. εικόνες, βίντεο, μουσική). Ωστόσο, η ποιότητα του χαρακτηρισμού είναι συνήθως μειωμένη, λόγω του ότι οι άνθρωποι που εμπλέκονται στη διαδικασία του χαρακτηρισμού δεν είναι ειδικοί. Έχουν προταθεί διάφορες τεχνικές για τη μάθηση από τέτοιου είδους δεδομένα (π.χ. μοντελοποίηση της ποιότητας των ανθρώπων που εκτελούν την εργασία, συγκερασμός των χαρακτηρισμών πολλών ανθρώπων για το ίδιο αντικείμενο). Στόχος της διπλωματικής είναι η μελέτη, υλοποίηση και σύγκριση τεχνικών μηχανικής μάθησης από τέτοιου είδους δεδομένα πολλαπλών ετικετών.

**Σχετικό υλικό:**

http://dl.acm.org/citation.cfm?id=1743478
http://jmlr.csail.mit.edu/papers/volume13/raykar12a/raykar12a.pdf
http://ir.ischool.utexas.edu/csdm2011/proceedings/csdm2011_kumar.pdf

**Επικοινωνία:** Γρηγόρης Τσουμάκας, e-mail: greg@csd.auth.gr, www: http://users.auth.gr/greg

### 10. Πρόβλεψη Πωλήσεων στη Βιομηχανία Τροφίμων

Η πρόβλεψη λιανικών πωλήσεων μπορεί να βοηθήσει μια επιχείρηση πώλησης τροφίμων στην ορθότερη σύνταξη των ημερήσιων παραγγελιών των καταστημάτων και επομένως στην ελλάτωση των απωλειών κέρδους είτε λόγω έλλειψης ειδών είτε λόγω καταστροφής ειδών των οποίων έχει παρέλθει η ημερομηνία λήξης. Επιπλέον, μπορεί να βοηθήσει στον αποδοτικότερο σχεδιασμό της παραγωγής των τροφίμων στο εργοστάσιο. Στόχος της διπλωματικής αυτής είναι η εφαρμογή τεχνικών παλινδρόμησης πολλαπλών στόχων σε πραγματικά δεδομένα πωλήσεων ζαχαροπλαστείων της Θεσσαλονίκης στο πλαίσιο ερευνητικού έργου. Η διπλωματική περιλαμβάνει τη μελέτη της σχετικής βιβλιογραφία, την προ-επεξεργασία των δεδομένων, την εφαρμογή των αλγορίθμων μηχανικής μάθησης, την αξιολόγηση των αποτελεσμάτων και τη συγγραφή επιστημονικής δημοσίευσης.

**Σχετικό υλικό:** http://www.win.tue.nl/~mpechen/projects/sligro/

**Επικοινωνία:** Γρηγόρης Τσουμάκας, e-mail: greg@csd.auth.gr, www: http://users.auth.gr/greg

*Οι ενδιαφερόμενοι μπορούν να επικοινωνήσουν με τα μέλη της ομάδας Μηχανικής Μάθησης και Ανακάλυψη Γνώσης (MLKD) για περισσότερες πληροφορίες και για να εκδηλώσουν το ενδιαφέρον τους για κάποιο/α από τα θέματα. Περισσότερες λεπτομέρειες για τα ερευνητικά ενδιαφέροντα της ομάδας MLKD θα βρείτε στη διεύθυνση http://mlkd.csd.auth.gr.*