ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΑΡΟΥΣΙΑΣΗ / ΕΞΕΤΑΣΗ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Ξένου Ρουμπίνη

Μεταπτυχιακή Φοιτήτρια

Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

Επόπτης Μεταπτ. Εργασίας: Αναπλ. Καθηγητής, Ι. Τσαμαρδίνος

Δευτέρα, 23/1/2017, 16:00

Αίθουσα Τηλεδιάσκεψης Κ206, Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

" Ένας καθοδηγητής ανάλυσης δεδομένων βασισμένος σε κανόνες"

ΠΕΡΙΛΗΨΗ

Η ανάλυση των δεδομένων είναι μια αναδυόμενη και σε κάθε περίπτωση χρήσιμη επιστήμη. Οι διάφοροι αλγόριθμοι μηχανικής μάθησης παρέχουν τη δυνατότητα να εκπαιδεύονται σε ένα σύνολο δεδομένων, αποτελούμενο από κάθε είδους, πραγματικές ή προσομοιωμένες, παρατηρήσεις, και να δημιουργούν ένα μοντέλο που τα περιγράφει.

Με το πέρασμα του χρόνου, ο αριθμός των περιπτώσεων, σε επαγγελματικό ή καθημερινό επίπεδο, που διευκολύνεται από την ανάπτυξη μιας μεθόδου εξόρυξης δεδομένων γίνεται όλο και μεγαλύτερος.

Προκειμένου να καλυφθούν οι πολυποίκιλες ανάγκες που προκύπτουν, ο χώρος των διαθέσιμων αλγορίθμων και μεθοδολογιών εξόρυξης δεδομένων ολοένα αυξάνεται, καθιστώντας την εξερεύνηση τους ως μια επίπονη και χρονοβόρο διαδικασία, ακόμη και για τους πιο πεπειραμένους αναλυτές δεδομένων.

Μία ακόμη δυσκολία, είναι η απαίτηση ξεχωριστής κατάλληλης μεθοδολογίας και ερμηνείας για κάθε διαφορετικό τύπο δεδομένων.

Μια μεγάλη προσπάθεια έχει δοθεί στην ανάπτυξη καθοδηγητών εξόρυξης δεδομένων, με σκοπό να βοηθήσουν το χρήστη να ξεπεράσει τα παραπάνω εμπόδια. Μέχρι στιγμής, οι καθοδηγητές ταξινομούνται ως αυτοματοποιημένοι και συνεργατικοί.

Στην εργασία αυτή, σχεδιάστηκε και αναπτύχθηκε ένα αυτοματοποιημένο έξυπνο σύστημα, το RB-DMA , το οποίο, βασισμένο στην OntoDM οντολογία, σε συνδυασμό με ένα σύνολο κανόνων εκφρασμένων με την βοήθεια του συστήματος drools, προτείνει τις πιο κατάλληλες ροές εργασιών εξόρυξης δεδομένων , διατεταγμένες με βάση την αποτελεσματικότητά τους για μια δεδομένη ανάλυση.

Η προσέγγισή μας παρέχει, σε χρήστες οποιουδήποτε επιπέδου γνώση, όλες τις αποφάσεις που χρειάζονται προκειμένου να προβούν σε μία ανάλυση με αξιόπιστα αποτελέσματα. Για έναν χρήστη με πλήρη άγνοια, η ανάγκη "να γίνει έμπειρος" εξαλείφεται. Από την άλλη πλευρά, το σύστημα θα λειτουργεί περισσότερο ως ένας μηχανισμός υπενθύμισης των διαθέσιμων βέλτιστων πρακτικών για τον έμπειρο χρήστη. Ακόμη, το σύστημά μας βοηθάει στην μείωση του απαιτούμενου χρόνου για να πραγματοποιηθεί μία ανάλυση, εγγυώμενο, στις περισσότερες περιπτώσεις, σχεδόν βέλτιστα αποτελέσματα εκτελώντας μόνο τις πρώτες K καλύτερες ροές.

Τελευταίο, αλλά όχι λιγότερο σημαντικό, το σύστημα καλύπτει έως 200 σενάρια ανάλυσης δεδομένων (ταξινόμηση δύο κλάσεων, και παλινδρόμηση με διαφορετικούς τύπους και μεγέθη δεδομένων), βοηθώντας των αναλυτή να ξεπεράσει το προαναφερθέν πρόβλημα της ξεχωριστής διαχείρισης διαφορετικών δεδομένων.

Ksenou Roubini

M.Sc. Thesis

Computer Science Department

University of Crete

Master's Thesis Supervisor: Associate Professor I. Tsamardinos

Monday, 23/1/2017, 16:00
Room K206, Computer Science dept., University of Crete

"A rule based data mining assistant"

ABSTRACT

Data analysis is an emerging and by any means useful science. The various Machine Learning algorithms provide the ability of being trained on a dataset, consisted of any kind, real world or simulated, observations, and generate a model describing them.

With the passage of time the amount of professional, or everyday life cases, facilitated by the deployment of a data mining method is getting larger.

In order to cover the manifold needs, the available data mining algorithm and methodologies space is growing, rendering their exploration into a taught and time consuming task, even for the most senior data analysts.

One more difficulty, is the demanding of separate interpretation and suitable data mining tasks for each different dataset type.

A lot of effort has been given towards the development of Data Mining Assistants, trying to help the user overcome the above obstacles. So far, the IDAs are classified as automated and cooperative.

In this Thesis we designed and developed an automated intelligent system, the RB-DMA (Rule Based Data Mining Assistant), which, based on an extension of the OntoDM data mining ontology \cite{panov2010representing}, combined with a set of rules written in Drools \cite{proctor2011drools}, proposes the most appropriate data mining workflows, ordered based on their efficiency for a given analysis.

Our approach provides, to any level of prior knowledge, users all the decisions they need in order to conduct an analysis with trustful results. For the ignorant user , the need to "become an expert" is eliminated. On the other side the system will operate more as a reminder of the available best practices for the senior one. Even more, the time consumption is reduced due to our system guaranteeing close to best results only by executing the first k best workflows, in most cases.

Last, but not least, the system covers up to 200 data analysis scenarios (binary classification and regression with different data types and sizes), enabling the data analyst to deal with the separate prior interpretation problem.