ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΑΡΟΥΣΙΑΣΗ / ΕΞΕΤΑΣΗ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Καλουρής Δημήτριος

Μεταπτυχιακός Φοιτητής

Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

Επόπτης Μεταπτυχιακής Εργασίας: Καθηγητής, Ι. Τσαμαρδίνος

Δευτέρα, 21 Φεβρουαρίου 2022, ώρα 09:30 π.μ.

Join Zoom Meeting

https://zoom.us/j/98725868189

" Αυτοματοποιημένη Κατασκευή Χαρακτηριστικών σε Σχεσιακά Δεδομένα"

Περίληψη

Στην Μηχανική Μάθηση συχνά μαθαίνουμε μοντέλα από ένα μοναδικό πίνακα. Όμως, στην εποχή των μεγάλων δεδομένων στην οποία ζούμε πολύ συχνά τα δεδομένα είναι μοιρασμένα σε πολλούς διαφορετικούς πίνακες μέσα σε μια βάση για καλύτερη αποδοτικότητα. Για να δουλέψουν με τα σχεσιακά αυτά δεδομένα δεν είναι σπάνιο οι επιστήμονες να δημιουργούν ένα-ένα τα χαρακτηριστικά σε μια διαδικασία που είναι κυρίως ενστικτώδης και απαιτεί γνώσεις πεδίου. Στόχος είναι η δημιουργία ενός Αυτοματοποιημένου Αλγορίθμου Κατασκευής Χαρακτηριστικών, τον οποίο οποιασδήποτε χωρίς ειδικές γνώσεις μπορεί να χρησιμοποιήσει. Πολλοί αλγόριθμοι έχουν προταθεί που μετατρέπουν μια βάση από πολλούς πίνακες σε έναν, αλλά κανείς από αυτούς δεν λαμβάνει υπόψη τα πιο περίπλοκα σχήματα βάσεων. Επιπλέον αυτοί οι αλγόριθμοι, κατά την παραγωγή των χαρακτηριστικών, συσσωρεύουν ένα μεγάλο πλήθος από χαρακτηριστικά πριν εκτελέσουν κάποιον αλγόριθμο επιλογής χαρακτηριστικών ενώ ακόμη οι αλγόριθμοι επιλογής που χρησιμοποιούν δεν είναι βελτιστοποιημένοι. Για αυτό και εμείς δημιουργήσαμε έναν επιγραμμικό αλγόριθμο κατασκευής χαρακτηριστικών ο οποίος εκτελεί ενώσεις και αθροιστικές συναρτήσεις στους πίνακες για να παράξει χαρακτηριστικά, ενώ κρατάει μόνο τα πιο χρήσιμα από αυτά μέσα από ένα μοντέλο υπόλοιπα που υπολογίζει υπόλοιπα. Τέλος προτείνουμε έναν αλγόριθμο επιλογής χαρακτηριστικών που κλιμακώνει σε μεγάλου όγκου δεδομένα, ο οποίος μπορεί να βλέπει έναν πίνακα σε κομμάτια και μετά να αθροίσει την πληροφορία χωρίς μεγάλο κόστος στην απόδοση.

University of Crete

Computer Science Department

M.Sc. Thesis

Kalouris Dimitrios

Master's Thesis Supervisor: Professor, I. Tsamardinos

Monday, 21 February 2022, 09:30 a.m.

Join Zoom Meeting

https://zoom.us/j/98725868189

"Automated Feature Engineering on Relational Data"

Abstract

Machine learning typically learns from a single table. However, in the age of big data it is often the case that data are distributed across many different table in a relational database for efficiency. To work with relational data, it is not rare for scientists perform feature engineering manually and intuitively. Additionally, many algorithms that produce a single table from a relational database have been proposed for this problem but none of them takes into account complex relational data schemas or they are limited in the paths they follow and the combinations of joins and aggregations they perform during feature generation. Moreover, these algorithms, during feature generation, accumulate large number of features before performing feature selection and the feature selection algorithms are not optimized. To this end we created SRFGA a novel online feature engineering algorithm that performs joins and aggregations on the tables to create features and keeps only the most useful features, using the residuals calculated by a model to guide the feature selection. This algorithm can be used without any knowledge expertise, and it also unifies all the previous works in terms of visited paths and actions performed.