ΕΛ.Ι.Δ.Ε.Κ.
Ελληνικό Ίδρυμα Έρευνας & Καινοτομίας

Περιγραφή Χρηματοδοτούμενου Ερευνητικού Έργου

1η Προκήρυξη Ερευνητικών Έργων ΕΛ.ΙΔ.Ε.Κ. για την ενίσχυση των Μελών ΔΕΠ και Ερευνητών/τριών και την προμήθεια ερευνητικού εξοπλισμού μεγάλης αξίας

Τίτλος Ερευνητικού Έργου: “Data Science for All”

Επιστημονικός Υπεύθυνος: Ιωάννης Τσαμαρδίνος

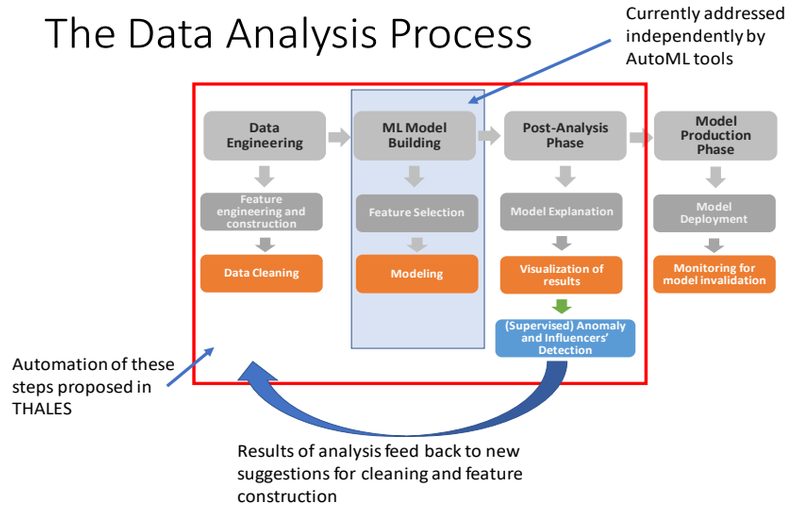
Φιλικός προς τον αναγνώστη τίτλος: “THALES”

Επιστημονική Περιοχή: Μαθηματικά και Επιστήμες της Πληροφορίας

Φορέας Προέλευσης και Χώρα: Ελλάδα

Φορέας Υποδοχής: Ινστιτούτο Εφαρμοσμένων Μαθηματικών, Ινστιτούτο Τεχνολογίας και Έρευνας Κρήτης (ΙΥΜ-ΙΤΕ)

Συνεργαζόμενος Φορέας: Τμήμα Επιστήμης Υπλογιστών Πανεπιστημίου Κρήτης (ΤΕΥ-ΠΚ)



Ποσό Χρηματοδότησης: 169.515,50 Ευρώ

Διάρκεια Χρηματοδότησης: 36 μήνες

ΕΛ.Ι.Δ.Ε.Κ.
Ελληνικό Ίδρυμα Έρευνας & Καινοτομίας

Σύνοψη Ερευνητικού Έργου

Η Επιστήμη Δεδομένων (Data Science) υπόσχεται να φέρει επανάσταση στη σύγχρονη επιστήμη, το επιχειρείν και την κοινωνία γενικότερα. Ωστόσο, εξακολουθεί να είναι σε μεγάλο βαθμό μια χειρωνακτική και χρονοβόρα διαδικασία στην οποία εύκολα υπεισέρχονται σφάλματα και απαιτεί σημαντική τεχνογνωσία. Ως εκ τούτου, σχεδόν κάθε εταιρία λογισμικού ανάλυσης δεδομένων, μια νέα γενιά τεχνοβλαστών σε Ευρώπη και Αμερική, καθώς και εξέχουσες ακαδημαϊκές ερευνητικές ομάδες, επικεντρώνονται πλέον στην αυτοματοποίηση των διαδικασιών ανάλυσης δεδομένων. Δυστυχώς όμως, είμαστε μακριά από αυτόν τον στόχο. Ιδιαίτερα, η Μηχανική Δεδομένων (Data Engineering, καθαρισμός δεδομένων, προεπεξεργασία και κατασκευή χαρακτηριστικών) - που θεωρείται ως το πιο χρονοβόρο βήμα της διαδικασίας ανάλυσης - δεν έχει αυτοματοποιηθεί. Επιπλέον, αντιμετωπίζεται ανεξάρτητα από το επόμενο βήμα, αυτό της Αναλυτικής Δεδομένων. Στο έργο THALES προτείνουμε ένα ερευνητικό πρόγραμμα για να γεφυρώσουμε τη Μηχανική Δεδομένων με την Αναλυτική Δεδομένων (Data Analytics) σε ένα ολοκληρωμένο αλγοριθμικό πλαίσιο, να αυτοματοποιήσουμε πλήρως τα βήματα της ανάλυσης και κάνουμε την Επιστήμη Δεδομένων προσβάσιμη σε όλους. Συγκεκριμένα, εστιάζουμε στη χρήση προβλεπτικών μοντέλων για την στοχοποίηση των δεδόμενων προς «καθαρισμό», την αυτοματοποιημένη κατασκευή χαρακτηριστικών (features) από σχεσιακές βάσεις δεδομένων και το σχεδιασμό μαζικά παράλληλων, κλιμακούμενων, αλγορίθμων πολλαπλής επιλογής χαρακτηριστικών (multiple feature selection). Ακόμα περισσότερο, τονίζουμε την ενοποίηση των παραπάνω σε ένα ενιαίο αλγοριθμικό πλαίσιο και την υλοποίησή τους ως Machine-Learning-as-a-Service. Τα αποτελέσματα είναι επιστημονικού ενδιαφέροντος αλλά και εξαιρετικά εμπορικά εκμεταλλεύσιμα. Ο ΕΥ έχει ήδη δημιουργήσει μια εταιρία τεχνοβλαστό του Πανεπιστημίου που θα χρησιμεύσει στο μέλλον ως μέσο εμπορικής εκμετάλλευσης.

Πρωτοτυπία του Ερευνητικού Έργου

Οι βασικές θεωρητικές, αλγοριθμικές και συστημικές προκλήσεις που αντιμετωπίζει το πρόγραμμα THALES είναι:

•Η προεπεξεργασία και η μοντελοποίηση των δεδομένων έχουν αντιμετωπιστεί μέχρι σήμερα από τις κοινότητες των ΒΔ και ΜΜ κατα βάση ως ανεξάρτητα προβλήματα σε διαφορετικά βήματα μιας ροής ανάλυσης δεδομένων. Σε αυτό το πλαίσιο, δυνητικά τεράστιοι όγκοι δεδομένων καθαρίζονται άσκοπα ακόμη και όταν δεν συμμετέχουν στην εκπαίδευση του τελικού μοντέλου. Παρομοίως, μια πληθώρα χαρακτηριστικών συνήθως κατασκευάζονται χωρίς αναγκαστικά να έχουν κάποια προβλεπτική αξία για τα μοντέλα που τελικά θέλουμε να κατασκευάσουμε.

•Ενα μεγάλο πλήθος αναλυτικών διεργασιών πρέπει να εκτελεστεί επαναληπτικά ώστε να αξιολογηθούν μέσω δοκιμής και σφάλματος με κάθε νέα λειτουργία καθαρισμού ή νέα ιδέα για κατασκευή χαρακτηριστικών. Με αυτό τον τρόπο τεράστιοι όγκοι δεδομένων εξάγονται δυνητικά ξανά και ξανά από μια βάση δεδομένων σε μορφή επίπεδων πινάκων οι τροφοδοτούν τους αλγορίθμους επιλογής μεταβλητών και εκμάθησης μοντέλων.

•Οι αναλυτικές διεργασίες συνήθως επεξεργάζονται το σύνολο των διαθέσιμων δεδομένων. Κατα συνέπεια δύσκολα κλιμακώνουν σε δεδομένα μεγάλης κλίμακας. Στην ιδανική περίπτωση, θα θέλαμε να επεξεργαστούμε (αυτόματα) μόνο τα υποσύνολα δεδομένων που αρκούν για την λήψη ισχυρών αποφάσεων απο διαφορετικούς αλγόριθμους εκμάθησης.

Long-Term Vision	Targeted scientific breakthrough
Bridge Data Engineering (Cleaning and Feature Construction) with Machine Learning Perspectives into an integrated algorithmic framework	Create theoretical and algorithmic underpinnings for interweaving solutions of Data Engineering and Model Building tasks to create Automated Data Cleaning and Feature Construction pipelines
Automate the entire pipeline of predictive analytics	Create algorithms for simultaneous Cleaning, Feature Construction, and Feature Selection that are fully automated, efficient, and scalable to Big Data
Make Data Science accessible to All types of users, increasing productivity by at least 10-fold	Encapsulate the algorithmic advances to industrial-strength, distributed, robust and versatile implementations that are offered as MLaaS

Αναμενόμενα αποτελέσματα & Αντίκτυπος του Ερευνητικού Έργου

Στόχος 1: Αυτοματοποίηση καθαρισμού δεδομένων σε κλειστό βρόχο. Η βασική ιδέα είναι να παραλείψουμε τον μη στοχευμένο καθαρισμό όλων των δεδομένων και των τιμών χαρακτηριστικών τους που δεν επηρεάζουν το τελικό μοντέλο μηχανικής εκμάθησης, που παράγεται απο επόμενα βήματα ανάλυσης. Αντ 'αυτού, ο στόχος είναι να σχεδιαστούν αλγόριθμοι καθαρισμού δεδομένων που καθοδηγούνται από τις αστοχίες και την αβεβαιότητα του παραγόμενου προβλεπτικού μοντέλου. Αυτοί οι αλγόριθμοι θα εστιάζουν στα δείγματα εκμάθησης και τα χαρακτηριστικά πρόβλεψης που είναι πιθανά «βρώμικα» και απαιτούν κάποια δράση καθαρισμού.

Στόχος 2: Αυτοματοποίηση κατασκευής προβλεπτικών χαρακτηριστικών από σχεσιακές βάσεις δεδομένων. Ο στόχος είναι να σχεδιαστούν αλγόριθμοι που παράγουν αυτόματα χαρακτηριστικά με αυξημένη προβλεπτική ικανότητα, εκτελώντας ερωτήματα SQL (συνδέσεις) απευθείας στη βάση δεδομένων, πρωτού χρησιμοποιηθουν απο κάποιο αλγόριθμο μηχανικής μάθησης. Οι αλγόριθμοι θα βασίζονται σε στατιστικές ευρετικές μεθόδους που αποφεύγουν την συστηματική κατασκευή και δοκιμή ολων των πιθανών σχεσιακών χαρακτηριστικών από εμπειρογνώμονες αναλυτές.

Στόχος 3: Κλιμάκωσιμοι αλγόριθμοι κατασκευής και επιλογής μεταβλητών για δεδομένα μεγάλης κλίμακας. Ο στόχος είναι να σχεδιαστούν αλγόριθμοι που επιτρέπουν τον μαζικό-παράλληλο υπολογισμό επιλογής χαρακτηριστικών ελαχιστοποιόντας το κόστος επικοινωνίας δικτύου και επιστρέφοντας αποτελέσματα που προσεγγίζουν ικανοποιητικά τους υπολογισμούς που εκτελούν οι αντίστοιχοι κεντρικοποιημένοι αλγόριθμοι.
Επιπλέον, αυτοί οι αλγόριθμοι θα βασίζονται μόνο σε παραδείγματα εκπαίδευσης, αποφεύγοντας την γραμμική επεξεργασία δεδομένων μεγάλης κλίμακας.

Στόχος 4: Ολοκλήρωση του καθαρισμού δεδομένων, της κατασκευής και της επιλογής χαρακτηριστικών για δεδομένα μεγάλης κλίμακας, σε κλειστό βρόχο. Ο στόχος είναι να ενσωματωθούν τα παραπάνω βήματα σε ένα ενιαίο πλαίσιο που αυτοματοποιεί πλήρως ένα μεγαλύτερο μέρος της διαδικασίας ανάλυσης δεδομένων από αυτό που είναι δυνατόν σήμερα, τουλάχιστον για δεδομένα που είναι αποθηκευμένα σε μια σχεσιακή βάση δεδομένων.

Η σημασία της χρηματοδότησης

Παρά τις σημαντικές προσπάθειες έρευνας και ανάπτυξης για την ενίσχυση του χαρακτήρα αυτοεξυπηρέτησης ορισμένων αναλυτικών διεργασιών, όπως η αυτοματοποίηση της προεπεξεργασίας δεδομένων ή της ρύθμισης μοντέλων Μηχανικής Μάθησης, η ανάπτυξη υψηλής ποιότητας ροών αναλυτικών διεργασιών εξακολουθεί να είναι μια εγγενής επαναληπτική διαδικασία που απαιτεί εμπειρογνώμονες που συνδυάζουν εξειδίκευση στη στατιστική, τη βελτιστοποίηση και τη Μηχανική Μάθηση με δεξιότητες βάσεων δεδομένων και προγραμματισμού. Το προτεινόμενο έργο είναι φιλόδοξο από πολλές απόψεις. Πρώτον, θα συνδυάσει δύο, επί του παρόντος ξεχωριστές και διακριτές, ερευνητικές προβληματικές και μεθόδους από τις κοινότητες των βάσεων δεδομένων και της Μηχανικής Μάθησης. Δεύτερον, θα διερευνήσει τους κοινούς στόχους των δύο κοινοτήτων σχετικά με την ευρεία και μεγάλης κλίμακας αναλυτική δεδομένων για να δημιουργήσει μια ερευνητική ατζέντα νέων προβλημάτων γύρω από τη μάθηση μοντέλων και τη μηχανική δεδομένων. Τρίτον, θα αξιοποιήσει τη βιομηχανική δυναμική σχετικά με την αυτοματοποίηση επιμέρους καθηκόντων ανάλυσης δεδομένων που βασίζονται σε πραγματικές μελέτες περιπτώσεων, συλλογών δεδομένων κ.λπ. Από την άποψη αυτή, αναμένεται να παραχθούν νέοι αλγόριθμοι, θεωρητικά αποτελέσματα, συστήματα και εργαλεία που θα δημοσιευτούν σε κορυφαία διεθνή περιοδικά και συνέδρια, αλλά και να διοργανωθούν εξειδικευμένα επιστημονικά fora και άλλες ερευνητικές και επιστημονικές δραστηριότητες.

ΕΛ.Ι.Δ.Ε.Κ.
Ελληνικό Ίδρυμα Έρευνας & Καινοτομίας

ΕΠΙΚΟΙΝΩΝΙΑ

Λ. Συγγρού 185 & Σάρδεων 2
ΤΚ. 17121, Νέα Σμύρνη, Ελλάδα
210 64 12 410, 420
communication@elidek.gr
www.elidek.gr