ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΑΡΟΥΣΙΑΣΗ / ΕΞΕΤΑΣΗ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Μεταπτυχιακός Φοιτητής
Γιαννούλης Μιχαήλ

Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης
Επόπτης Μεταπτ. Εργασίας: Καθηγητής, Β. Χριστοφίδης

Πέμπτη, 05/03/2020, 10:00
Αίθουσα Τηλεδιάσκεψης Κ206, Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

" Πειραματική Αξιολόγηση Ανιχνευτών Ανωμαλιών σε Ροές Δεδομένων "

ΠΕΡΙΛΗΨΗ

Η πειραματική αξιολόγηση των αλγόριθμων ανίχνευσης ανωμαλιών χωρίς επίβλεψη αποτελεί μια σταθερή πρόκληση σε διάφορους τομείς έρευνας και εφαρμογών. Ωστόσο, λίγα είναι γνωστά όσον αφορά τα πλεονεκτήματα και τις αδυναμίες των μεθόδων ανίχνευσης ανωμαλιών εντός σύνδεσης και το αντίκτυπο των παραμέτρων τους. Η παρούσα ερευνητική δημοσίευση αποσκοπεί στο σχεδιασμό και την ανάπτυξη ενός πλαισίου συγκριτικής αξιολόγησης για την εξαγωγή μιας εκτενούς πειραματικής μελέτης πάνω στις δεντρικές και πλησιέστερου γείτονα μεθόδους των κορυφαίων μη επιτηρούμενων ανιχνευτών ανωμαλιών εντός σύνδεσης (συμπεριλαμβανομένου και των εκτός σύνδεσης) σε συνεχή ροή, μέσα από μια μεγάλη ποικιλία από πολυδιάστατα δεδομένα τα οποία έχουν μολυνθεί είτε με ανωμαλίες υποχώρου είτε πλήρους

διανυσματικού χώρου. Αρχικά, παρουσιάζουμε τη σημασιολογία και τις λειτουργίες των ανιχνευτών μέσω ενός περιεκτικού παραδείγματος. Στη συνέχεια, εισάγουμε το περιβάλλον πειραματικής μελέτης το οποίο παρέχει περιγραφική (μέτα) ανάλυση των δεδομένων και τις επιλογές υλοποίησης των ανιχνευτών, θέτοντας επίσης το σύνολο των υπερπαραμέτρων και των υποψήφιων τιμών τους. Η δίκαιη αξιολόγηση των ανιχνευτών εξασφαλίζεται μέσω επαρκούς ανάλυσης κρίσιμων μεθοδολογικών ερωτημάτων όπως η προσομοίωση και διαμοιρασμός ροής δεδομένων, τα πρωτόκολλα και οι μετρικές αξιολόγησης, η βελτιστοποίηση των ανιχνευτών και η κατάταξή τους. Μέσω αυτής της μελέτης, διαπιστώνουμε ότι οι ανιχνευτές εντός σύνδεσης όχι μόνο προσεγγίζουν πολύ καλά τους ανιχνευτές εκτός σύνδεσης (0.777 έναντι 0.778 αντίστοιχα, τιμές κατάταξης) αλλά τους ξεπερνούν υπό συνθήκες. Επίσης διαπιστώνουμε με έκπληξη την αντίσταση του δυναμικού μοντέλου των ανιχνευτών εντός σύνδεσης στην κλιμάκωση της διάστασης των δεδομένων και των υποχώρων τους. Πάραυτα, παρουσιάζουν μειωμένη απόδοση καθώς κλιμακώνονται οι τιμές των υπερπαραμέτρων των παραθύρων τους. Εξετάζουμε επίσης τα θεμελιώδη στοιχεία ενός δυναμικού μοντέλου, υπογραμμίζοντας την ανάγκη μηχανισμού λήθης. Από όσο γνωρίζουμε, αυτή είναι η πιο ολοκληρωμένη προσπάθεια πειραματικής αξιολόγησης ανιχνευτών ανωμαλιών σε σύνδεση πάνω σε πολυδιάστατα δεδομένα.

Giannoulis Michail
M.Sc. Thesis

Computer Science Department
University of Crete
Master's Thesis Supervisor: Professor, V. Chrisophides

Thursday, 05/03/2020, 10:00
Room K206, Computer Science Dept., University of Crete

“Benchmarking Anomaly Detectors on Streaming Data”

ABSTRACT

The experimental evaluation of unsupervised anomaly detection algorithms is a constant challenge within diverse research areas and applications domains. However, little is known regarding the strengths and weaknesses of online anomaly detection methods and the impact of their parameters. This paper elaborates on the design and development of a benchmark framework to perform an extensive experiment study on tree and nearest-neighbor based methods of top-notch unsupervised online outlier detectors (including their offline) in streaming manner, across a wide variety of multivariate datasets contaminated by sub and full space outliers. Initially, we present the semantics and functionalities of the detectors through a comprehensive example. Then, we introduce the benchmark environment providing a descriptive (meta) analysis of the datasets and implementation choices of detectors, posing also the set of their hyper-parameters and candidate values. The fair evaluation of detectors is guaranteed through an adequately analysis of critical methodological questions such as stream simulation and partitioning, evaluation protocols and metrics, detectors optimization and ranking. Through this study, we ascertain that online detectors not only approximate very well offline detectors (0.777 vs 0.778, respectively; Ranking value) but also outperform them under certain conditions. In addition, we surprisingly establish the resistance of online detectors' dynamic model on scaling data and subspace dimensionality. Nevertheless, they shown a decreasing performance while scaling window parameters. We also examine the fundamentals of a dynamic model highlighting the need for a forgetting mechanism. To the best of our knowledge, this is the most complete online anomaly detection benchmark attempt on multivariate data.