Πετράκης Κωνσταντίνος *ΑΜ: 1041589* *ΥΔΑ*

Algorithms for Characterization and Trend Detection in Spatial Databases

Στην παρούσα εργασία παρουσιάζονται κάποιοι νέοι αλγόριθμοι για 2 προβλήματα. Για τον χωρικό χαρακτηρισμό (spatial characterization) των αντικειμένων μιας χωρικής Βάσης Δεδομένων (SDBS), τον καθορισμό δηλαδή της κλάσης του αντικειμένου όχι μόνο από τα μη-χωρικά χαρακτηριστικά του αλλά και από τα χαρακτηριστικά γειτονικών του αντικειμένων και την χωρική ανάλυση τάσεων (spatial trend analysis) η οποία συνίσταται στον καθορισμό κάποιων μοτίβων αλλαγής σε κάποια μη-χωρικά χαρακτηριστικά στη γειτονιά κάποιου αντικειμένου της Βάσης.

Δοθέντος του neighborhood graph, στο οποίο περιλαμβάνονται ακμές μόνο μεταξύ χωρικά γειτονικών αντικειμένων, το σύνολο των αντικειμένων στόχων targets, το μέγιστο πλήθος γειτόνων max-neighbors και το επίπεδο εμπιστοσύνης confidence, ο χωρικός χαρακτηρισμός συνίσταται στη ανίχνευση κάθε πιθανής άγνωστης ιδιότητας prop για τα αντικείμενα εκείνα που απέχουν λιγότερο από max-neighbors από το σύνολο αντικειμένων targets και ο παράγοντας συχνότητας (frequency factor) τους είναι τουλάχιστον ίσος με το confidence που δίνεται. Τελικά ο χαρακτηρισμός δίνεται ως μια περιγραφή των ιδιοτήτων (χωρικών η μη) οι οποίες συναντώνται στα αντικείμενα του συνόλου targets αλλά όχι σε όλη τη Βάση.

Για την ανίχνευση χωρικών τάσεων (trends) εκκινώντας από ένα αντικείμενο O, οι συγγραφείς εκτελούν παλινδρόμηση (regression) στις σχετικές τιμές των χαρακτηριστικών των αντικειμένων στα μονοπάτια του neighborhood graph για να περιγράψουν την 'κανονικότητα' της αλλαγής δηλ. την τάση (το trend). Στη παλινδρόμηση ανεξάρτητη μεταβλητή αποτελεί η απόσταση από το αντικείμενο εκκίνησης O ενώ οι διαφορές στις τιμές των χαρακτηριστικών έχουν τον ρόλο των εξαρτημένων μεταβλητών. Με αυτό τον τρόπο η συσχέτιση μεταξύ των παρατηρούμενων τιμών των χαρακτηριστικών και των τιμών που προέβλεψε η παλινδρόμηση, αποτελεί ένα μέτρο του confidence για το ανακαλυφθέν trend. Σε αυτό το μήκος κύματος οι συγγραφείς παρουσιάζουν 2 αλγορίθμους, τον global-trend και τον local-trends, με τον μεν πρώτο να επιστρέφει το σημαντικότερο χωρικό trend, αυτό με το μέγιστο μήκος (με τη μεγαλύτερη 'έκταση') και τον δεύτερο να επιστρέφει 2 σύνολα μονοπατιών από το neighborhood graph, ένα με τα θετικά trends και ένα με τα αρνητικά.

Οι παραπάνω αλγόριθμοι δοκιμάστηκαν πειραματικά στο σύνολο δεδομένων ATKIS 500 το οποίο αποτελείται από γεωγραφικά δεδομένα για τη Βαυαρία σχετικά με την οικονομική κατάσταση ανά περιοχή (περιλαμβάνει 1 χωρικό χαρακτηριστικό και 52 μη χωρικά). Ο αλγόριθμος χαρακτηρισμού καταφέρνει να παράγει έναν κανόνα ο οποίος να περιγράφει τις επιλεγμένες (target) περιοχές λαμβάνοντας υπ' όψιν μη-χωρικά χαρακτηριστικά αλλά και τη γειτονιά των επιλεγμένων περιοχών. Ακόμη οι αλγόριθμοι global-trend και local-trends για spatial trend detection δοκιμάστηκαν αποτελεσματικά για την εύρεση trends σε περιοχές όσον αφορά το μέσο ενοίκιο, για επίπεδα ελάχιστης συσχέτισης (confidence) από 0.6 μέχρι 0.9. Και οι 2 αλγόριθμοι παρήγαγαν συγκρίσιμα αποτελέσματα όσον αφορά το confidence με τον αλγόριθμο global-trend να εκτελεί περισσότερες πράξεις μεταξύ γειτόνων σε σχέση με τον local-trends, όπως ήταν αναμενόμενο.

Mining Recent Temporal Patterns for Event Detection in Multivariate Time Series Data

Σε αυτήν την εργασία οι συγγραφείς παρουσιάζουν ένα framework εξόρυξης πρόσφατων χρονικών μοτίβων (Recent Temporal Patterns) από χρονοσειρές με σκοπό την εύρεση γενικότερων μοτίβων πρόβλεψης σε προβλήματα παρακολούθησης και ανίχνευσης γεγονότων σε περίπλοκες χρονοσειρές πολλών μεταβλητών. Σε γενικές γραμμές το framework αρχικά μετατρέπει τις χρονοσειρές σε ακολουθίες χρονικών διαστημάτων πετυχαίνοντας έτσι μια μορφή χρονικής αφαίρεσης και στη συνέχεια κατασκευάζει πιο περίπλοκα χρονικά μοτίβα, εκκινώντας από τις πιο πρόσφατες παρατηρήσεις και πηγαίνοντας προς τα πίσω στο χρόνο με τη χρήση ειδικών χρονικών τελεστών.

Ειδικότερα, πρώτο βήμα της μέθοδος αποτελεί η εφαρμογή ενός δυναμικού μετασχηματισμού στα αρχικά δεδομένα χρονοσειρών, μετατρέποντας τις αριθμητικές μεταβλητές των χρονοσειρών σε ακολουθίες διαστημάτων, χρησιμοποιώντας χρονική αφαίρεση και ένα πεπερασμένο αλφάβητο Σ επιτρεπτών αφαιρέσεων. Με αυτόν τον τρόπο κάθε δείγμα αναπαρίσταται σαν μια πολυμετάβλητη ακολουθία καταστάσεων (Multivariate State Sequence). Οι συγγραφείς συνδυάζουν βασικές καταστάσεις αυτών των MSS με τη χρήση μόνο των χρονικών σχέσεων πριν (before) και συνύπαρξης (co-occurs) για να σχηματίσουν χρονικά μοτίβα διαστημάτων (time interval patterns). Στην συνέχεια ορίζονται οι συνθήκες που πρέπει να ικανοποιεί ένα τέτοιο μοτίβο για να θεωρηθεί πρόσφατο και οι σχέσεις προθέματος και υπέρ-μοτίβο πίσω επέκτασης (backward-extension superpattern) με βάση τις οποίες καθορίζεται εάν ένα πρόσφατο χρονικό μοτίβου (RTP) είναι συχνό (frequent).

Ακολούθως πρέπει να εκτελεστεί η εξόρυξη των πρόσφατων χρονικών μοτίβων (RTP). Για την ανίχνευση συχνών RTP χρησιμοποιούνται τα συχνά RTP από κάθε κλάση ξεχωριστά, με βάση ελάχιστα τοπικά supports, που καθορίζονται για κάθε κλάση, ώστε να αποφευχθεί η απώλεια μοτίβων σε μη-ισορροπημένα (unbalanced) δεδομένα αλλά και επειδή η εξόρυξη μοτίβων που είναι συχνά σε κάποια κλάση είναι πιο αποδοτική από την εξόρυξη συχνών μοτίβων στο σύνολο των δεδομένων. Ο αλγόριθμος εξόρυξης που παρουσιάζεται εδώ λαμβάνει τα MSS από μία κλάση, μια παράμετρο gap και μια παράμετρο σ για το ελάχιστο support για την κλάση και παράγει όλα τα χρονικά μοτίβα που ικανοποιούν το ελάχιστο support. Ο αλγόριθμος λειτουργεί σε 2 φάσεις, στη φάση της παραγωγής υποψηφίων όπου παράγονται υποψήφια (k+1)-μοτίβα επεκτείνοντας συχνά k-RTP προς τα πίσω στο χρόνο και τη φάση μέτρησης όπου απορρίπτονται τα πρόσφατα μοτίβα με support λιγότερο από σ. Οι συγγραφείς επίσης προτείνουν και 2 τρόπους με τους οποίους μπορούν να υλοποιηθούν πιο αποδοτικά αυτές οι 2 φάσεις.

Τέλος το κάθε δείγμα μετατρέπεται σε ένα δεικνύον (indicator) διάνυσμα χρησιμοποιώντας αυτά τα μοτίβα που επέστρεψε ο αλγόριθμος εξόρυξης και η τελική κατηγοριοποίηση εκτελείται με την εφαρμογή οποιουδήποτε αλγορίθμου μηχανικής μάθησης σε αυτά τα διανύσματα.

Το framework εφαρμόστηκε σε υγειονομικά EHR δεδομένα διαβητικών ασθενών με στόχο την ορθή κατηγοριοποίηση των διαταραχών που παρουσιάζει ο κάθε ασθενής. Τα αποτελέσματα των πειραμάτων δείχνουν πως τα χαρακτηριστικά που βασίζονται σε χρονικά μοτίβα έχουν ευεργετικά αποτελέσματα στην κατηγοριοποίηση καθώς υπερτερούν των χαρακτηριστικών που βασίζονται στις πιο πρόσφατες τιμές. Όσον αφορά το πλήθος των μοτίβων, ο αριθμός αυτών που εξάγονται με χρήση του RTP είναι τουλάχιστον μία τάξη μεγέθους μικρότερος από τον αριθμό αυτών που εξάγονται με TP. Επίσης τα μοτίβα που εξήγαγε ο αλγόριθμος αυτόματα χωρίς την ενσωμάτωση κάποιας πρότερης κλινικής γνώσης είναι σε πλήρη συμφωνία με τις οδηγίες ιατρικής διάγνωσης. Τέλος από την μέτρηση των χρόνων εκτέλεσης των αλγορίθμων που δοκιμάστηκαν φαίνεται πως η μέθοδος που προτείνεται είναι η πιο αποδοτική σε σχέση με άλλες.

Locally Adaptive Dimensionality Reduction for Indexing Large Time Series Databases

Στην παρούσα εργασία παρουσιάζεται μια νέα τεχνική μείωσης της διαστατικότητας για δεδομένα χρονοσειρών, η APCA, με στόχο την αποδοτική αναζήτηση με βάση την ομοιότητα σε Βάσεις χρονοσειρών. Σε αντίθεση με άλλες παρόμοιες τεχνικές η APCA προσεγγίζει κάθε χρονοσειρά με ένα σύνολο τμημάτων σταθερής τιμής και μεταβλητού μήκους έτσι ώστε να ελαχιστοποιείται το σφάλμα ανακατασκευής για κάθε χρονοσειρά. Οι συγγραφείς παρουσιάζουν πως μπορεί να ευρετηριαστεί η APCA, προτείνουν 2 μετρικές απόστασης που επιτρέπουν γρήγορη ακριβής αναζήτηση και ακόμη πιο γρήγορη προσεγγιστική αναζήτηση και συγκρίνουν την τεχνική τους με όλες τις υπόλοιπες.

Ακριβής αναζήτηση σημαίνει να μην υπάρχουν false dismissals, δηλαδή αντικείμενα που είναι κοντά στην πραγματικότητα να εμφανίζονται μακριά αφού γίνει η δεικτοδότηση, ενώ η προσεγγιστική αναζήτηση δεν το εγγυάται αυτό. Για να μην υπάρχουν false dismissals αρκεί να ικανοποιείται το lower bounding lemma δηλαδή η απόσταση 2 αντικειμένων στο ευρετήριο πρέπει να είναι μικρότερη από την πραγματική τους απόσταση στον αρχικό χώρο. Για την αναπαράσταση κάθε τμήματος σταθερής τιμής στην APCA χρειάζονται 2 αριθμοί, ένας για τη μέση τιμή των δεδομένων στο τμήμα και ένας για το μήκος του τμήματος. Οι 2 μετρικές που παρουσιάζονται για τη μέτρηση της απόστασης μεταξύ μιας query χρονοσειράς και μιας χρονοσειράς σε APCA αναπαράσταση είναι οι εξής. Η πρώτη μετρική D_{AE} αντιστοιχεί στην ευκλείδεια απόσταση μεταξύ query και της APCA αναπαράστασης της χρονοσειράς με την οποία συγκρίνω το query, αλλά δεν ικανοποιεί την τριγωνική ανισότητα (ούτε το lower bounding lemma) γι' αυτό και δεν μπορεί να χρησιμοποιηθεί για ακριβές ταίριασμα, είναι όμως χρήσιμη για προσεγγιστικό ταίριασμα. Η δεύτερη μετρική D_{LB} υπολογίζει την απόσταση αφού πρώτα μετατραπεί και το query σε APCA αναπαράσταση, και αποτελεί lower bound για την ευκλείδεια απόσταση. Για τη αποδοτική μετατροπή της χρονοσειράς σε APCA χρησιμοποιείται ένας αλγόριθμος που πρώτα αντιστοιχίζει το πρόβλημα σε ένα πρόβλημα wavelet συμπίεσης, για το οποίο υπάρχουν λύσεις, και στη συνέχεια μετατρέπεται αυτή η λύση στην αναπαράσταση APCA με μικρές τροποποιήσεις. Στόχος είναι πάντα η ποιότητα της προσέγγισης δηλ. το όσο γίνεται μικρότερο σφάλμα ανακατασκευής.

Η μετατροπή μιας χρονοσειράς σε APCA αναπαράσταση ισοδυναμεί με την αντιστοίχιση της σε έναν N-διάστατο χώρο, όπου N το πλήθος των τμημάτων σταθερής τιμής που χρησιμοποιείται. Η δεικτοδότηση γίνεται με τη χρήση μια πολυδιάστατης δομής δεικτοδότησης π.χ. R-tree στον APCA N-διάστατο χώρο και είναι ακριβώς αυτή η δομή ευρετηρίου που χρησιμοποιείται για την αποδοτική απάντηση σε K-NN queries και range queries. Οι συγγραφείς παρουσιάζουν ένα νέο τρόπο κατασκευής των MBR με βάση τις μέγιστες και ελάχιστες τιμές που λαμβάνει μια χρονοσειρά στα τμήματα σταθερής τιμής της APCA αναπαράστασης της. Επιπλέον αφού ορίσουν την ελάχιστη απόσταση μεταξύ ενός query και ενός MBR, αποδεικνύουν πως αυτή η ελάχιστη απόσταση του MBR από το query αποτελεί lower bound για την απόσταση οποιασδήποτε χρονοσειράς C κάτω από οποιονδήποτε κόμβο φύλλο του R-tree, από το query.

Οι συγγραφείς σύγκριναν πειραματικά σε δυο σύνολα δεδομένων την μέθοδο APCA με όλες τις άλλες τεχνικές ως προς την ταχύτητα απόκρισης σε ένα query για μια σειρά επιλεγμένων μειωμένων διαστάσεων και μήκη query. Έχουν διασφαλίσει πως οι υλοποιήσεις έχουν γίνει με τέτοιο τρόπο ώστε να μην μεροληπτούν υπέρ κάποιας μεθόδου, και το πέτυχαν αυτό συγκρίνοντας το pruning power των διάφορων προσεγγίσεων. Τα αποτελέσματα που λαμβάνονται δείχνουν πως η APCA υπερτερεί σημαντικά των DFT και DWT, για μια τάξη μεγέθους. Με βάση αυτό οι συγγραφείς συμπεραίνουν πως η τεχνική APCA παρουσιάζει λιγότερα false alarms επομένως και χαμηλότερο κόστος αναζήτησης. Ακόμη σύγκριναν τις διάφορες τεχνικές, συμπεριλαμβανομένης και της γραμμικής αναζήτησης, σε ένα ήδη υλοποιημένο σύστημα (το Sun Ultra Enterprise 450 machine) μετρώντας τα κόστη I/O και CPU. Η τεχνική APCA υπερισχύει σημαντικά σε σχέση με τις άλλες τεχνικές και στα κόστη I/O και στα CPU κόστη.