Πετράκης Κωνσταντίνος *ΑΜ:1041589* *ΥΔΑ*

Detecting Subdimensional Motifs: An Efficient Algorithm for Generalized Myltivariate Pattern Discovery

Στην παρούσα εργασία παρουσιάζεται και αξιολογείται εμπειρικά ένας γραμμικού χρόνου αλγόριθμος για την αποδοτική και ακριβής, μη επιβλεπόμενη ανίχνευση μοτίβων (patterns) σε δεδομένα χρονοσειρών πολλών μεταβλητών, τα οποία μοτίβα μπορεί να εκτείνονται σε ένα υποσύνολο του συνόλου των διαστάσεων της χρονοσειράς. Τέτοιου είδους subdimensional μοτίβα προκύπτουν σε κατανεμημένα συστήματα αισθητήρων, εξόρυξη πολυμέσων, ανάλυση αισθητήρων σώματος και από δεδομένα ανίχνευσης κίνησης. Ο αλγόριθμος που αναπτύσσεται εδώ αναζητά ζευγάρια όμοιων, σταθερού μεγέθους, υποακολουθιών και χρησιμοποιεί αυτά τα μοτίβα 'σπόρους' για να καθορίσει επιπλέον εμφανίσεις του ίδιου μοτίβου. Για την ανίχνευση των όμοιων ζευγαριών οι συγγραφείς επιστρατεύουν τη μέθοδο τοπικής διακριτοποίησης SAX, η οποία μετατρέπει τις υποακολουθίες σε words (strings) χαρακτήρων, και έπειτα χρησιμοποιούν πολλές επαναλήψεις τυχαίων προβολών επ' αυτών των words για τον εντοπισμό προσεγγιστικά ίσων υποακολουθιών σε γραμμικό χρόνο. Μόλις βρεθεί ένα πιθανό μοτίβο 'σπόρος' ο αλγόριθμος καθορίζει τη συνάφεια κάθε διάστασης για αυτό το μοτίβο. Εν συνεχεία εκτιμά το μέγεθος της γειτονιάς του μοτίβου και αναζητά επιπλέον εμφανίσεις του τελευταίου.

Οι ισοδύναμες προβολές καταμετρούνται σε ένα πίνακα συγκρούσεων (collision matrix). Τα στοιχεία του collision matrix αντιστοιχούν στο πλήθος των επαναλήψεων στο οποίο κάθε ζευγάρι υποακολουθιών ήταν ισοδύναμα έπειτα από την τυχαία προβολή. Ο πίνακας ενημερώνεται μετά από κάθε επανάληψη κατακερματίζοντας τις λέξεις που προέκυψαν και αυξάνοντας το στοιχεία που αντιστοιχούν σε ισοδύναμο ζευγάρι. Τελικά τα στοιχεία του collision matrix θα αντιπροσωπεύουν το σχετικό βαθμό ομοιότητας μεταξύ υποακολουθιών, παρέχοντας με αυτόν τον τρόπο στους ερευνητές τη δυνατότητα να εστιάζουν στην ανάλυση μόνο των ζευγαριών με σχετικά υψηλές τιμές ομοιότητας, γλυτώνοντας σε υπολογιστικό χρόνο.

Για πολυμετάβλητα δεδομένα όμως, κάθε διάσταση οδηγεί και σε ένα ξεχωριστό SAX word. Προς αυτή την κατεύθυνση, οι ερευνητές αυξάνουν τα στοιχεία του coliision matrix για κάθε διάσταση με την οποία ταιριάζει το word, αντί να συνενώνουν τις προβολές των λέξεων από κάθε διάσταση και να κατακερματίζουν το string που προκύπτει. Με αυτόν τον τρόπο αυξάνοντας τη σχετική καταχώριση του πίνακα για κάθε διάσταση που ταιριάζει λαμβάνεται υπ' όψιν από τον αλγόριθμο η πρόσθετη υποστήριξη (support) που παρέχουν πολλές όμοιες διαστάσεις.

Οι ερευνητές αξιολόγησαν τον παραπάνω αλγόριθμο χρησιμοποιώντας συνθετικά δεδομένα στα οποία είχαν τοποθετήσει ένα ή περισσότερα τεχνητά μοτίβα καθώς και πραγματικά δεδομένα αισθητήρων σώματος από την εκτέλεση μια άσκησης γυμναστικής στα οποία πρώτα δοκίμασαν να προσθέσουν αυξημένα επίπεδα θορύβου σε μία μόνο διάσταση 'θορύβου' και δεύτερον να προσθέσουν επιπλέον άσχετες διαστάσεις κάθε μια με μέτρια επίπεδα θορύβου. Από τα πειράματα με τα συνθετικά δεδομένα παρατηρείται πως ο χρόνος ανίχνευσης του μοτίβου αυξάνεται γραμμικά με το χρόνο όταν χρησιμοποιείται σαν μετρική η L1 απόσταση και τετραγωνικά με το χρόνο όταν χρησιμοποιείται DTW. Από τα πειράματα με τα πραγματικά δεδομένα στη πρώτη περίπτωση φάνηκε πως η ακρίβεια του αλγορίθμου παραμένει σχετικά καλή καθώς αυξάνεται η ποσότητα θορύβου στην επιπλέον διάσταση σε σχέση με αλγορίθμους που αναζητούν μοτίβα σε όλες τις διαστάσεις. Για την δεύτερη περίπτωση η απόδοση του αλγορίθμου μειώνεται καθώς προστίθενται διαστάσεις με θόρυβο αλλά πιο αργά σε σχέση με αλγορίθμους που αναζητούν μοτίβα σε όλες τις διαστάσεις.

On Data mining, compression and Kolmogorov complexity

Στην παρούσα εργασία οι συγγραφείς δείχνουν πως η εξόρυξη δεδομένων είναι στενά συνδεδεμένη με την συμπίεση και την πολυπλοκότητα Kolmogorov, και δεδομένου πως η τελευταία είναι μη υπολογίσιμη συμπεραίνουν πως δεν μπορεί να υπάρξει μια θεωρία ανάλογη της σχεσιακής άλγεβρας για τις βάσεις δεδομένων η οποία θα αυτοματοποιήσει πλήρως την εξόρυξη δεδομένων.

Οι συγγραφείς αρχικά παρουσιάζουν τις έννοιες της Kolmogorov πολυπλοκότητας μέσω παραδειγμάτων ανίχνευσης outliers σε γραμμικές και log-log κλίμακες. Επιχειρηματολογούν πως σχεδόν όλες οι λειτουργίες της εξόρυξης δεδομένων είναι στενά συνδεμένες με την συμπίεση, την πολυπλοκότητα Kolmogorov και την υπό συνθήκη πολυπλοκότητα Kolmogorov. Πιο συγκεκριμένα η ομαδοποίηση στη μη-επιβλεπόμενη μάθηση και η ανίχνευση ακραίων τιμών (outliers) σχετίζονται με την συμπίεση χωρίς απώλειες, ενώ η κατηγοριοποίηση στην επιβλεπόμενη μάθηση και ο ορισμός συνάρτησης απόστασης σχετίζονται με την υπό συνθήκη συμπίεση.

Η πολυπλοκότητα Kolmogorov λέει πως η πολυπλοκότητα ενός bit string είναι το συντομότερο πρόγραμμα μιας γενικής μηχανής Turing (UTM) που μπορεί να παράγει το επιθυμητό string. Ένα string με Kolmogorov πολυπλοκότητα μεγαλύτερη από το μήκος του string καλείται ασυμπίεστο. Με βάση τα παραπάνω εάν ένα σύνολο δεδομένων, μοντελοποιημένο σαν bit string, περιέχει κάποια μοτίβα και συσχετίσεις, η πολυπλοκότητα Kolmogorov του θα είναι χαμηλή. Το πρόβλημα έγκειται στο ότι δεν μπορούν να ανακαλυφθούν τα μοτίβα που εκμεταλλεύτηκε η UTM χαμηλότερης πολυπλοκότητας διότι είναι αδύνατο να εφαρμοστή αντίστροφη μηχανική (reverse engineer) σε αυτή. Αυτό προκύπτει από το γεγονός πως η πολυπλοκότητα Kolmogorov ενός αυθαίρετου string είναι μη-υπολογίσημη (undecidable), μπορεί ωστόσο να προσεγγιστεί μέσω της Lemper-Ziv κωδικοποίησης. Εν συνεχεία οι συγγραφείς δίνουν παραδείγματα αλγορίθμων Μηχανικής Μάθησης οι οποίοι χρησιμοποιούν όρους υπό συνθήκη Kolmogorov πολυπλοκότητας χωρίς να την καθορίζουν ρητά. Στη κατηγοριοποίηση με δέντρα απόφασης για κάθε κόμβο χρειάζεται ένα μέτρο ομοιογένειας που θα καθορίσει εάν θα χωριστεί ο κόμβος, το οποίο μέτρο συνήθως είναι η εντροπία, η οποία είναι στενά συνδεδεμένη με την πολυπλοκότητα Kolmogorov. Στην ομαδοποίηση ο καθορισμός του αριθμού των ομάδων k μπορεί να γίνει με τη χρήση μη-παραμετρικών τεχνικών (π.χ. MDL), οι οποίες σχετίζονται με την πολυπλοκότητα Kolmogorov. Όλες οι συναρτήσεις απόστασης στην ουσία προσπαθούν να μετρήσουν το κόστος μετατροπής ενός αντικειμένου σε ένα άλλο, γεγονός που μπορεί να μοντελοποιηθεί με την υπό συνθήκη πολυπλοκότητα Kolmogorov K(x|y), με τη string editing απόσταση, τη Time Warping απόσταση, την Eschera-Fu απόσταση στα ARG graphs, τη tf-idf στάθμιση και την ομοιότητα συνημίτονου να αποτελούν μερικά παραδείγματα τέτοιων συναρτήσεων.

Οι συγγραφείς εξετάζοντας την εξόρυξη δεδομένων από την σκοπιά της Kolmogorov πολυπλοκότητας καταλήγουν από τη μία πως αυτή η θεώρηση θα βοηθήσει στο σχεδιασμό μη-παραμετρικών αλγορίθμων για το πρόβλημα και από την άλλη πως δεν θα μπορέσουμε ποτέ να ξέρουμε αν ο αλγόριθμος που εφαρμόζουμε είναι ο καλύτερος, μόνο πόσο καλός είναι σε σύγκριση με άλλους. Ακόμη οι συγγραφείς με αφορμή την εφαρμογή ενός power law για την μοντελοποίηση τεχνητών και πραγματικών δεδομένων σε κλίμακες log-log παρατηρούν πως ο στόχος είναι να βρεθούν εκείνα τα μοντέλα που πετυχαίνουν την καλύτερη συμπίεση. Τέλος για να δείξουν πως η αναζήτηση καλών μοντέλων μπορεί να είναι δύσκολη, χρησιμοποιούν το παράδειγμα ενός fractal, του τριγώνου Sierpinksi, εικάζοντας πως ένα σύνολο δεδομένων με υψηλή fractal διαστατικότητα είναι πιο δύσκολο να συμπιεστεί και θα απαιτούσε ένα πιο πολύπλοκο μοντέλο σε σχέση με ένα χαμηλότερης fractal διαστατικότητας σύνολο δεδομένων, συμπεραίνοντας έτσι πως η fractal διαστατικότητα ενός συνόλου δεδομένων συνδέεται με την Kolmogorov πολυπλοκότητα.

Representative Clustering of Uncertain Data

Στην παρούσα εργασία οι συγγραφείς παρουσιάζουν και αξιολογούν πειραματικά ένα πλαίσιο (framework) με σκοπό την εφαρμογή οποιουδήποτε αλγόριθμου ομαδοποίησης για την εξαγωγή μιας ουσιαστικής ομαδοποίησης ενός αβέβαιου συνόλου δεδομένων, βασισμένο στη σημασιολογία πιθανών κόσμων (possible worlds), το οποίο υπολογίζει ένα σύνολο αντιπροσωπευτικών ομαδοποιήσεων οι οποίες δεν απέχουν περισσότερο από ένα κατώφλι από την πραγματική ομαδοποίηση των πραγματικών αλλά άγνωστων δεδομένων. Προτείνουν μια λύση με βάση τη δειγματοληψία της Βάσης Δεδομένων (DB), παρουσιάζουν μια μεθοδολογία μέσω της οποίας μπορούν να αξιολογούν τη ποιότητα μιας ομαδοποίησης ενός possible world σε σχέση με την πραγματική ομαδοποίηση και δείχνουν πως η εμπιστοσύνη στα αποτελέσματα της ομαδοποίησης σε πιθανούς κόσμους μπορεί να βελτιωθεί υπολογίζοντας ένα σύνολο πολλαπλών αντιπροσωπευτικών ομαδοποιήσεων, η κάθε μια με σημαντική πιθανότητα να μοιάζει με την αληθινή άγνωστη ομαδοποίηση.

Μία αβέβαιη DB ορίζεται από ένα σύνολο πιθανών καταστάσεων που ονομάζονται πιθανοί κόσμοι και κάθε πιθανός κόσμος συνδέεται με την αντίστοιχη πιθανότητα να είναι η πραγματική κατάσταση της Βάσης. Η προσέγγιση τους έγκειται στη δειγματοληψία της αβέβαιης DB και στη παράγωγη |X| πιθανών κόσμων, σαν ένα σύνολο δειγματικών βέβαιων βάσεων δεδομένων. Εφαρμόζουν τον επιλεγμένο αλγόριθμο ομαδοποίησης σε κάθε έναν από τους πιθανούς κόσμους λαμβάνοντας το σύνολο των πιθανών ομαδοποιήσεων PC. Για οποιαδήποτε ομαδοποίηση C στο σύνολο PC η υποστήριξη (support) της ομαδοποίησης C.supp είναι ίση με το πλήθος των εμφανίσεων της ομαδοποίησης C στο multiset PC. Αποδεικνύεται πως η ποσότητα C.supp/|X| (υποστήριξη προς το μέγεθος του δείγματος) είναι ένας αμερόληπτος εκτιμητής της πιθανότητας η ομαδοποίηση C να είναι η πραγματική ομαδοποίηση του δείγματος.

Μια ομαδοποίηση τώρα είναι τ-φ αντιπροσωπευτική εάν η πιθανότητα αυτή η ομαδοποίηση να απέχει από την πραγματική ομαδοποίηση της Βάσης λιγότερο από τ, είναι τουλάχιστον φ. Οι συγγραφείς προσεγγίζουν αυτήν την πιθανότητα σαν το πλήθος των πιθανών ομαδοποιήσεων στο σύνολο PC που απέχει λιγότερο από τ από την εκάστοτε ομαδοποίηση διαιρεμένο με το σύνολο των δειγμάτων |X|. Η πιθανότητα αυτή αποτελεί αμερόληπτο εκτιμητή για τη πιθανότητα μια ομαδοποίηση να είναι τ-φ αντιπροσωπευτική. Οι ερευνητές χρησιμοποιούν ένα κάτω φράγμα γι αυτή τη πιθανότητα του αμερόληπτου εκτιμητή υπό ένα επίπεδο σημαντικότητας α το οποίο καθορίζεται από το χρήστη.

Οι συγγραφείς δίνουν επίσης έναν τρόπο ώστε να επιλεγούν εκείνοι οι αντιπροσωπευτικοί κόσμοι με υψηλή εμπιστοσύνη και οι οποίοι ταυτόχρονα απέχουν λιγότερο από την πραγματική ομαδοποίηση. Το πετυχαίνουν αυτό εφαρμόζοντας έναν αλγόριθμο ομαδοποίησης C', με βάση την απόσταση, στο σύνολο PC κατασκευάζοντας έτσι το σύνολο των αντιπροσωπευτικών ομαδοποιήσεων (RPC). Αυτή η ομαδοποίηση αντιπροσωπευτικών κόσμων επιστρέφει έναν α-σημαντικό αντιπρόσωπο, με φραγμένη από κάτω πιθανότητα φ, για κάθε μετά-ομάδα που προέκυψε από την εφαρμογή του αλγορίθμου C' στο σύνολο PC. Για την επιλογή εκπροσώπου κάθε μετά-ομάδας είτε απαιτείται όλες οι πιθανές ομαδοποιήσεις στη μετά-ομάδα να εκπροσωπούνται είτε εκπροσωπούνται μόνο εκείνες οι ομαδοποιήσεις που ικανοποιούν το κατώφλι απόφασης τ_{max} που δίνει ο χρήστης.

Στα πειράματα χρησιμοποιήθηκε σαν αλγόριθμος ομαδοποίησης ο DBSCAN και σαν C' ο PAM. Για την αξιολόγηση της μεθόδου αρχικά χρησιμοποιήθηκαν διάφορα σύνολα πραγματικών δεδομένων. Από την εφαρμογή της μεθόδου σε ένα μικρό παράδειγμα με συνθετικά δεδομένα προέκυψε πως η ομαδοποίηση που ήταν πιο κοντά στην πραγματική έλαβε την μεγαλύτερη πιθανότητα. Τα πειράματα έδειξαν πως η ποιότητα των αποτελεσμάτων βελτιώνεται όταν επιστρέφονται μέχρι και 10 αντιπρόσωποι. Επίσης γίνεται εμφανές πως ένα σύνολο 4 αντιπροσωπευτικών cluster έχει ικανοποιητικά αποτελέσματα για το μέσο χρήστη. Τέλος αυξάνοντας τον αριθμός των δειγμάτων |X| αυξάνεται και η εμπιστοσύνη (confidence).

Dependency Clustering Across Measurement Scales

Σε αυτήν την εργασία παρουσιάζεται και αξιολογείται πειραματικά ο αλγόριθμος Scenic για την ομαδοποίηση ετερογενών τύπων δεδομένων, συνεχή αριθμητικών και διακριτών κατηγορικών, οργανώνοντας τα αντικείμενα και τα χαρακτηριστικά τους σε ένα χώρο χαμηλότερης διάστασης ακολουθώντας την αρχή του ελάχιστου μήκους περιγραφής (MDL) για την αποδοτική συμπίεση των δεδομένων.

Η βασική ιδέα είναι η θεώρηση μιας ομάδας σαν ένα σύνολο αντικειμένων τα οποία χαρακτηρίζονται από ένα μοναδικό μοτίβο συσχέτισης των χαρακτηριστικών. Αυτά τα μοτίβα ανιχνεύονται προβάλλοντας τα αντικείμενα σε έναν χαμηλότερης διάστασης διανυσματικό χώρο ο οποίος καλείται χώρος αντικειμένων-χαρακτηριστικών (AO space) και συμβάλει στην ερμηνεία των αποτελεσμάτων. Με αυτόν τον τρόπο η ομαδοποίηση σχετίζεται άμεσα με την μη επιβλεπόμενη κατηγοριοποίηση και τη συμπίεση. Σε αυτό το AO χώρο γίνεται η αναζήτηση συστάδων, οι οποίες εξηγούν ή προβλέπουν τις αρχικές μικτού τύπου (mixed-type) τιμές χαρακτηριστικών των δεδομένων με υψηλή ακρίβεια. Όσον αφορά τα αριθμητικά χαρακτηριστικά που προβάλλονται στον AO χώρο, απαραίτητη προϋπόθεση είναι να διατηρείται η διάταξη και τα διαστήματα μεταξύ τους από τον αρχικό χώρο.

Η ιδέα κλειδί της εργασίας είναι η θεώρηση της ομαδοποίησης σαν ένα πρόβλημα μη-επιβλεπόμενης κατηγοριοποίησης. Δοθείσης της αναπαράστασης των αντικειμένων και των χαρακτηριστικών στον χώρο AO η συγγραφέας θέλει να προβλέψει τις αρχικές τιμές των χαρακτηριστικών, δηλαδή τα μικτού τύπου δεδομένα εισόδου, με υψηλή ακρίβεια. Θεωρώντας πως μία ομάδα αντιστοιχεί σε μία άγνωστη κλάση των δεδομένων, ο στόχος της ομαδοποίησης είναι να βρεθούν συμπλέγματα αντικειμένων τα οποία μεγιστοποιούν τη ακρίβεια πρόβλεψης. Για την αποφυγή υπέρ-προσαρμογής (overfitting) η ιδέα της μη-επιβλεπόμενης κατηγοριοποίησης συνδυάζεται με την αρχή συμπίεσης MDL. Απώτερος στόχος της ομαδοποίησης είναι η να βρεθεί ένας διαχωρισμός των αντικειμένων σε ομάδες ώστε να ελαχιστοποιείται το συνολικό Μήκος Περιγραφής (Description Length), το οποίο για κάθε ομάδα δίνεται από το άθροισμα του κόστους ανακατασκευής (RE) και της πολυπλοκότητα του μοντέλου (MC) της ομάδας.

Για την εύρεση ενός καλού AO χώρου χρησιμοποιείται ο αλγόριθμος Princals ο οποίος αντιστοιχίζει τα αντικείμενα στο νέο χώρο χαμηλότερης διάστασης με τέτοιο τρόπο ώστε αυτά τα αντικείμενα να είναι όσο πιο κοντά γίνεται στις πραγματικές τιμές των χαρακτηριστικών τους και όσο πιο μακριά γίνεται από τις υπόλοιπες τιμές. Όσον αφορά την εύρεση των ομάδων ο αλγόριθμος αρχικά αναθέτει κάθε αντικείμενο στην ομάδα στην οποία έχει τα ελάχιστα κόστη κωδικοποίησης και στη συνέχεια ενημερώνει το μοντέλου της ομάδας ενημερώνοντας τον AO χώρο. Αρχικά όλα τα αντικείμενα ανήκουν σε μια ομάδα και διαχωρίζονται σε υποομάδες όσο παρατηρείται βελτίωση στα κόστη κωδικοποίησης. Κατά τη φάση διαχωρισμού η ομαδοποίηση εκτελείται στον AO χώρο. Τελικά η καλύτερη AO διαστατικότητα κάθε μεμονωμένης ομάδας καθορίζεται με την αρχή MDL.

Όσον αφορά την πειραματική αξιολόγηση ο αλγόριθμος Scenic συγκρίνεται με τους αλγορίθμους INCONCO, K-Mean Mixed, K-means και K-modes σε σύνολα συνθετικών και πραγματικών δεδομένων. Σαν μετρική χρησιμοποιείται η κανονικοποιημένη αμοιβαία πληροφορία (NMI). Για τα συνθετικά δεδομένα ο αλγόριθμος Scenic ήταν ο μόνος που ανίχνευσε την αριθμητική συσχέτιση στην πρώτη ομάδα και τις μικτού τύπου εξαρτήσεις στην δεύτερη ομάδα, ομαδοποιώντας τέλεια το σύνολο δεδομένων. Για το $1^{ο}$ σύνολο πραγματικών δεδομένων, Abalone, επαληθεύεται πως ο διαχωρισμός των δεδομένων σε 4 ομάδες από τον αλγόριθμο Scenic είναι λογικός και σύμφωνος με την πρότερη γνώση που υπάρχει για τα δεδομένα αλλά και πως αυτές οι 4 ομάδες παρουσιάζουν σημαντικές διαφορές μεταξύ τους, όπως είναι επιθυμητό. Για το $2^{ο}$ πραγματικό σύνολο δεδομένων, Acute Inflamations, ο αλγόριθμος Scenic είναι ο μοναδικός αλγόριθμος με καλή απόδοση και στα δυο χαρακτηριστικά αξιολόγησης, δημιουργώντας ομάδες στον AO χώρο οι οποίες επιτρέπουν τον ευδιάκριτο διαχωρισμό των ασθενών μεταξύ των δυο παθήσεων.