Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ

ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΚΙΝΟΥΜΕΝΩΝ ΑΝΤΙΚΕΙΜΕΝΩΝ ΒΑΣΕΙ ΣΗΜΑΣΙΟΛΟΓΙΚΑ ΕΠΑΥΞΗΜΕΝΩΝ ΣΥΝΟΨΕΩΝ

Μαρκάκης Γεώργιος

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς

Οκτώβριος 2021

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. ……. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών.

Τα μέλη της Επιτροπής ήσαν:

- Αναπλ Καθηγητής Ν. Πελέκης (Επιβλέπων)

- Επίκουρος Καθηγητής Χ. Ευαγγελάρας

- Αναπλ. Καθηγητής Ε. Κοφίδης

Η έγκριση της Διπλωματικής Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN

APPLIED STATISTICS

CLASSIFICATION OF MOVING OBJECTS BASED ON SEMANTICALLY ENRICHED SYNOPSES

By

Markakis George

Thesis

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece

October 2021

Στους γονείς μου

Εμμανουήλ και Ζαχαρένια

Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή της διπλωματικής εργασίας μου, κ. Πελέκη Νικόλαο, για την γνώση, την έμπνευση αλλά και τις πολύτιμες συμβουλές που μου παρείχε σε όλη την διάρκεια των σπουδών μου στο Μεταπτυχιακό Πρόγραμμα, καθώς και την περίοδο της συγγραφής της παρούσας διπλωματικής εργασίας.

Θα ήθελα να ευχαριστήσω και τα μέλη της τριμελούς κ. Χ. Ευαγγελάρα και κ. Ε. Κοφίδη για τις χρήσιμες παρατηρήσεις τους πάνω στην εργασία καθώς και στην παρουσίαση της.

Επιπλέον θα ήθελα επίσης να ευχαριστήσω τους καθηγητές του προγράμματος μεταπτυχιακών σπουδών για τις γνώσεις που μου μετέδωσαν ο καθένας ξεχωριστά. Φυσικά δεν θα μπορούσα να παραλείψω τους συμφοιτητές μου, οι οποίοι μου προσέφεραν ένα περιβάλλον στο οποίο μπόρεσα να αναπτυχθώ ακόμα περισσότερο .

Τέλος, ευχαριστώ τους γονείς μου για όλη την υποστήριξη τους και την ενθάρρυνση τόσο στις προπτυχιακές μου όσο και στις μεταπτυχιακές μου σπουδές.

Περίληψη

Η ανάπτυξη στον τομέα της τεχνολογίας της πληροφορίας και της επικοινωνίας, ιδίως στην ανίχνευση κινητής τηλεφωνίας και στην ασύρματη επικοινωνία, μας πλημμυρίζει δεδομένα που περιέχουν γεωγραφικές θέσεις που ποικίλλουν χρονικά. Αν και αυτό το είδος δεδομένων σχετίζεται επίσης με προκλήσεις όπως η εξάντληση της χωρητικότητας αποθήκευσης και το εύρος ζώνης μετάδοσης δεδομένων, οι ερευνητές έχουν δείξει ότι αυτά τα σύνολα δεδομένων αποτελούν πολύτιμο πόρο. Η ανάλυσή τους μπορεί να οδηγήσει σε λύσεις για σημαντικά ερευνητικά προβλήματα σε διάφορους τομείς, όπως πολεοδομικός σχεδιασμός, μεταφορά, οικολογική συμπεριφορά, ανάλυση αθλητικών σκηνών, παρακολούθηση και ασφάλεια.

Στην εργασία επιχειρείται η προσπάθεια να αναλυθούν και να κατηγοριοποιηθούν δεδομένα κίνησης με την συνδρομή αλγορίθμων Μηχανικής Μάθησης. Παρουσιάζονται τεχνικές που ασχολούνται με το εν λόγω ζήτημα, καθώς και ορισμοί και τρόποι χειρισμού δεδομένων που είναι σημαντικοί για τον αναγνώστη ώστε να μπορέσει να του δώσει πληροφορίες σχετικά με το εν λόγω ζήτημα. Επιπλέον, παρουσιάζεται η μέθοδος MasterMovelets (Ferrero et al., 2020) η οποία χρησιμοποιήθηκε και για τις ανάγκες του πειραματικού μέρους της εργασίας.

Abstract

Development in the field of information and communication technology, especially in mobile telephony detection and wireless communication, floods us with data containing geographical locations that vary over time. Although this type of data is also associated with challenges such as depletion of storage capacity and data bandwidth, researchers have shown that these data sets are a valuable resource. Their analysis can lead to solutions to important research problems in various fields, such as urban planning, transportation, ecological behavior, sports scene analysis, monitoring and security.

This thesis attempts to analyze and classify kinetic data with the help of Machine Learning algorithms. Bibliographic techniques dealing with this issue are presented, as well as definitions and handling of data that is important to the facilitator so that he can provide information on the issue. In addition, the MasterMovelets (Ferrero et al., 2020) method is presented which was used for the needs of the experimental part of the work.

Περιεχόμενα

Πίνακας Σχημάτων

Κατάλογος Πίνακων

1. Εισαγωγή

Στην παρούσα ενότητα γίνεται παρουσίαση του σκοπού και του αντικειμένου της εργασίας, προκειμένου ο αναγνώστης να μπορεί να κατανοήσει το θέμα αυτής και τις απαιτήσεις του. Επιπλέον δίνεται η διάρθρωση του κειμένου, ώστε να μπορέσει να γίνει άμεση εύρεση των στοιχείων που επιθυμεί ο αναγνώστης να διαβάσει.

1.1 Αντικείμενο – Σκοπός εργασίας

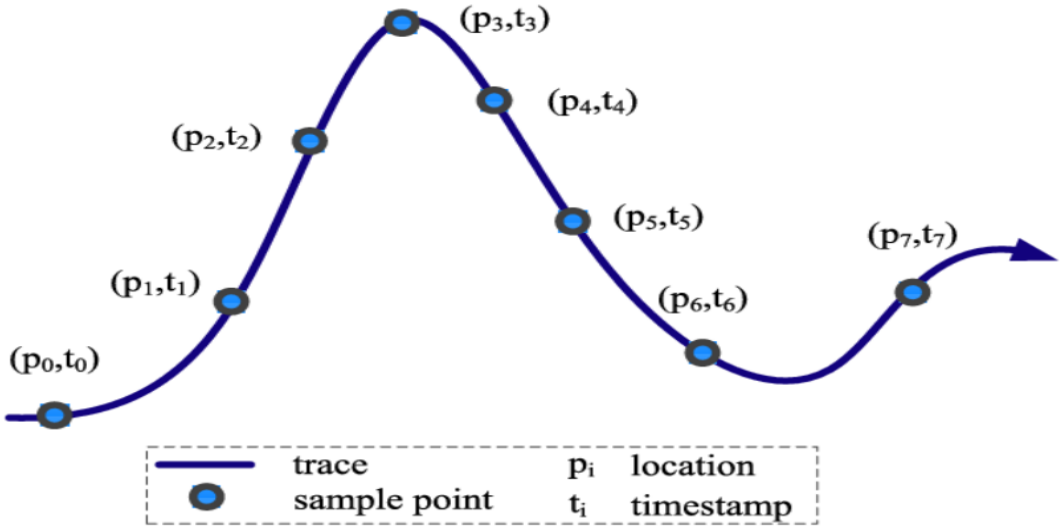
Στην εργασία επιχειρείται η προσπάθεια να αναλυθούν και να κατηγοριοποιηθούν δεδομένα κίνησης με την συνδρομή αλγορίθμων Μηχανικής Μάθησης. Βέβαια, οι μέθοδοι αντικατοπτρίζονται στην κατηγοριοποίηση των παρατηρήσεων με βάση τις σημάνσεις τους, δηλαδή χαρακτηριστικά που δεν αφορούν απλώς τις κινήσεις τους, αλλά αποτελούν μια επιπλέον πληροφορία που μπορεί να συνδράμει στην ανάδειξη ενός χαρακτηριστικού τους.

Είναι γεγονός ότι νέα προϊόντα ολοένα και δημιουργούνται στην αγορά. Όμως σε συνδυασμό με την ανάπτυξη της τεχνητής νοημοσύνης, η οποία έχει εξελίξει τις συσκευές παρακολούθησης και μεγάλο αριθμό κινούμενων αντικειμένων, απαιτείται πιο διεξοδική συλλογή δεδομένων ώστε να παραχθούν σημαντικές πληροφορίες.

Η κινούμενη κατηγοριοποίηση είναι ένας αποτελεσματικός τρόπος για την ανάλυση δεδομένων που βρίσκονται εν κινήσει και έχει εφαρμοστεί στην αναγνώριση προτύπων, ανάλυση δεδομένων, Μηχανική Μάθηση. Επιπλέον, η κινούμενη ομαδοποίηση (fication) εφαρμόζεται συχνά στη λήψη πληροφοριών συντεταγμένων σε κινητά δεδομένα. Πιο συγκεκριμένα, μπορεί να επικαλεστεί ως μέθοδος πρόβλεψης κινούμενων αντικειμένων (Chen et al., 2010), ελέγχου της κυκλοφορίας στους δρόμους

(Atev et al., 2006; Gurung et al., 2014; Gonzalez et al., 2007), νοητικής δραστηριότητας (Bashir et al., 2007; Oneata et al., 2016; Feng et al., 2017), τρισδιάστατης πρόβλεψης (Dai et al., 2017), πρόβλεψη καιρού (Carlos et al., 2013), αλλά ακόμη και στο πεδίο της γεωγραφίας (Di et al., 2015).

Υπάρχουν πολλοί τρόποι καταγραφής των δεδομένων των μεταβλητών αντικειμένων. Το γεγονός αυτό εξαρτάται κυρίως από τη μορφή των συσκευών, την κίνηση αντικειμένου που πραγματοποιείται και άλλους λόγους που αφορούν τους σκοπούς της άντλησης των εκάστοτε δεδομένων. Κλασσική εφαρμογή εντοπίζεται στις συσκευές παρακολούθησης του παγκόσμιου συστήματος εντοπισμού θέσης (GPS) όπου δημιουργείται ένας ανιχνευτής ο οποίος εξετάζει τις μεταβολές της κίνησης αντικειμένου ως $Trajectory = (Tr_1, Tr_2, ..., Tr_N)$, το οποίο είναι μια ακολουθία πορείας N σημείων στο γεωγραφικό χώρο και το Tr_i υποδηλώνει γεωγραφικές συντεταγμένες και χρονική στιγμή μέσω του διανύσματος $Tr_i = (x_i, y_i, t_i)$.



Σχήμα 1 : Τροχιά από δειγματοληψία ιχνών.

Πηγή : Feng & Zhu et al., 2016.

Στο μοντέλο αυτό μπορούν να προστεθούν και άλλοι παράμετροι όπως η ταχύτητα, κατεύθυνση, επιτάχυνση (Ying et al., 2010; Ying et al., 2011). Η μεγάλη διαφορά από κλασσικές συσκευές GPS, είναι ότι οι τελευταίες μπορούν να συμπεριλάβουν μονάχα δεδομένα για τη θέση του κινούμενου σώματος. Εναλλακτικά, υπάρχει η δυνατότητα εύρεσης δεδομένων από εικόνες ακόμη και videos. Συχνή είναι η χρήση τέτοιων μεθόδων και στις περιπτώσεις, όπου τα μοντέλα είναι οι ίδιες οι εικόνες κάτι που συναντάται συχνά στη βιβλιογραφία (Cai et al., 2016; Oneata et al., 2016). Δημιουργείται αρχικώς ακολουθία από στοιχεία εικόνων σε πολλά καρέ ώστε να πραγματοποιηθεί μια ανάλυση ευαισθησίας και λεπτομέρειας. Έτσι, δημιουργείται μια κίνηση (ή τροχιά), η οποία συμπίπτει σε μεγάλο βαθμό με οπτικές ροές.

Στην περίπτωση των videos, συλλέγονται δεδομένα που έχουν καταγραφεί από κινήσεις αντικειμένων (Dai et al., 2012; Ishikawa et al., 2008). Οπότε μέσω των εφαρμογών της κατηγοριοποίησης κινήσεων από εικόνες ή βίντεο, ανακτώνται χωρο-χρονικές συνιστώσες και πληροφορίες εικόνας, όπως λεπτομέρειες εικόνων ή χρήση κάποιας κλίμακας μέτρησης. Βεβαίως, τα σημασιολογικά δεδομένα κινήσεων είναι το μέρος που η πιο πρόσφατη βιβλιογραφία διερευνά περισσότερο (Choi et al., 2017; Sadiq et al., 2015). Αυτό που τα θέτει σε τέτοιο επίπεδο ώστε να υπερτερούν με τα υπόλοιπα δεδομένα, είναι το περιεχόμενό τους, το οποίο είναι πλούσιο από ενδείξεις με σκοπό τη βελτιστοποίηση της ακρίβειας της κατηγοριοποίησης. Επιπροσθέτως, η εφαρμογή τους είναι άμεση, για αυτό το λόγο μειώνεται σε αισθητό βαθμό ο υπολογιστικός χρόνος που απαιτείται στις κλασσικές περιπτώσεις.

Οι ημι-επιβλεπόμενες (semi-supervised) και επιβλεπόμενες (supervised) κατηγοριοποιήσεις κινήσεων τροχιάς συναντιούνται συχνά και αξίζει να αναφερθούν. Σε αυτές τις περιπτώσεις χρησιμοποιείται στην εισαγωγή των δεδομένων μία διάκριση σε δεδομένα δοκιμής από ένα συγκεκριμένο σύνολο αρίθμησης των δεδομένων που

υπάρχουν (test_set), και ένα σύνολο προς εκτίμηση με ταυτόχρονη σύγκριση με τις πραγματικές τιμές των δεδομένων του συγκεκριμένου συνόλου (train_set). Με αυτό τον τρόπο, οι αλγόριθμοι εκπαιδεύονται πάνω στα δεδομένα του train_set, στα οποία υπάρχουν ετικέτες (labels) ώστε να αντιλαμβάνεται ο κώδικας την διάκριση τους με τα υπό εξέταση δεδομένα.

Τα σημασιολογικά δεδομένα που μπορούν να συλλεχθούν από προηγμένης τεχνολογίας συσκευές, ενημερώνουν για σκοπούς κατηγοριοποίησης και αυτό συμβάλλει στην ελάττωση του πλήθους των παρατηρήσεων στο σύνολο της εκπαίδευσης test_data και στη βελτιστοποίηση της ακρίβειας των αποτελεσμάτων που θα εξαχθούν από τον αλγόριθμο. Βέβαια, ο καθορισμός των δεδομένων για τους σκοπούς της εκπαίδευσης του κώδικα, θα πρέπει να πραγματοποιείται από εξειδικευμένους επιστήμονες με ειδικές δοκιμές.

Ουκ ολίγες είναι οι περιπτώσεις που η αρίθμηση των δεδομένων είναι σχεδόν ακατόρθωτη. Κλασσικό παράδειγμα αποτελεί η προσπάθεια κατηγοριοποίησης ή διάκρισης των κινήσεων των πεζών συγκεντρωμένοι σε μεγάλα πλήθη. Σε αυτή την περίπτωση, η αναγνώριση του κάθε ατόμου είναι πολύ δύσκολη, καθώς τέτοιου είδους φωτογραφίες υστερούν σε ποιότητα ώστε να ληφθεί η πληροφόρηση που επιζητείται.

Στις παραπάνω περιπτώσεις υπάρχει μια άλλη μέθοδος Μηχανικής Μάθησης που συνιστάται. Επομένως, γίνεται αντιληπτό ότι σε αυτά τα κινητά δεδομένα χρειάζεται να μην δοθούν ετικέτες («labels»). Συνεπώς, οι τρόποι με τους οποίους σε αυτή την περίπτωση, θα συνδράμουν στην κατηγοριοποίηση των κινήσεών τους θα είναι οι λεγόμενοι μη εποπτευόμενες μέθοδοι κατηγοριοποίησης τροχιάς. Με βάση αυτές, δύναται να ομαδοποιηθούν τα σημασιολογικά δεδομένα.

Στη βιβλιογραφία εντοπίζονται κάποια διακεκριμένα papers για την σύνοψη του ζητήματος της κατηγοριοποίησης των κινούμενων αντικειμένων (Kong et al., 2018; Mazimpaka & Timpf et al 2016; Morris &Trivedi et al., 2008;). Στο ένα από αυτά τα άρθρα δίνεται έμφαση στην ιστορική αναδρομή των κινητών δεδομένων και στον τρόπο με τον οποίο χρησιμοποιούνται σε διάφορα παραδείγματα, όπως στην εφαρμογή κινούμενων ατόμων (Kong et al., 2018). Επιπλέον, δίνεται έμφαση σε μεθόδους εξόρυξης δεδομένων εστιάζοντας σε εκείνα που εμπεριέχουν σημασιολογικό χαρακτήρα (Mazimpaka & Timpf et al., 2016; Li et al., 2017).

1.2 Διάρθρωση κειμένου

Κατά την ανάλυση της εργασίας πραγματοποιείται ομοιόμορφη κατανομή στα κεφάλαια που ακολουθούν σε αυτό το κείμενο. Ο χωρισμός του κειμένου πραγματοποιείται σε πέντε κεφάλαια, σε κάθε ένα από τα οποία γίνεται ανάλυση και πραγμάτευση ενός ζητήματος, πράγμα απαραίτητο για την ολοκλήρωση της εργασίας. Σκοπός του κάθε κεφαλαίου είναι να κάνει αντιληπτή στον αναγνώστη την συλλογιστική πορεία που ακολουθήθηκε, και να κάνει κατανοητά τα μέσα που τέθηκαν προς χρήση, προκειμένου να πραγματοποιηθεί η εργασία και να γίνει παρουσίαση των αποτελεσμάτων της.

Στο κεφάλαιο αυτό γίνεται η αρχική παρουσίαση του αντικειμένου της εργασίας, του θέματος το οποίο πραγματεύεται, αλλά πραγματοποιείται και αναφορά στους στόχους της. Σκοπός του κεφαλαίου είναι να πραγματοποιηθεί η παρουσίαση των πραγματευόμενων από την εργασία αντικειμένων, ο καθορισμός των στόχων και του αντικειμένου της, και τέλος, να γίνει φανερός ο τρόπος μελέτης, ως προς τον αναγνώστη.

Στη συνέχεια της εργασίας, ακολουθεί βιβλιογραφική επισκόπηση συναφών θεμάτων. Βεβαία, ο σκοπός της μελέτης είναι η κατηγοριοποίηση κινούμενων αντικειμένων με βάση τη σημασία των τροχιών, οπότε βρίσκεται στο επίκεντρο των αναφορών που θα πραγματοποιηθούν.

Το τρίτο κεφάλαιο παρουσιάζει τα τεχνικά χαρακτηριστικά της μεθόδου MasterMovelets. Η μέθοδος αυτή θα χρησιμοποιηθεί για το πειραματικό μέρος της παρούσας εργασίας.

Το τέταρτο κεφάλαιο παρουσιάζει τα δεδομένα που χρησιμοποιούνται και το πειραματικό μέρος της εργασίας όπως αυτή έχει υλοποιηθεί, μέσα από την παράθεση εικόνων και επεξηγήσεων καθώς και των μεθοδολογιών που ακολουθήθηκαν. Τέλος, μετά την παρουσίαση της εφαρμογής σημειώνονται τα συμπεράσματα τα οποία προέκυψαν από την ανάπτυξη της παρούσας εργασίας. Στο πέμπτο κεφάλαιο παρουσιάζονται οι μελλοντικές επεκτάσεις που μπορεί να προκύψουν από την εργασία αυτή και τέλος στο έκτο κεφάλαιο οι βιβλιογραφικές αναφορές που χρησιμοποιήθηκαν.

2. Βιβλιογραφική Επισκόπηση

Οι τροχιές και η εφαρμογή τους σε εργασίες κατηγοριοποίησης είναι οι βασικές έννοιες που είναι απαραίτητες για την κατανόηση της παρούσας εργασίας. Για το σκοπό αυτό, σε αυτήν την ενότητα περιγράφονται οι κύριες έννοιες, ξεκινώντας από την εξήγηση των δεδομένων τροχιάς, το πρόβλημα κατηγοριοποίησης τροχιάς και τις τεχνικές Μηχανικής Μάθησης που υπάρχουν.

Η χωρική τροχιά είναι ένα ίχνος που δημιουργείται από ένα κινούμενο αντικείμενο σε γεωγραφικούς χώρους, που συνήθως αντιπροσωπεύεται από μια σειρά χρονικά ταξινομημένων σημείων, για παράδειγμα, $p_1 \rightarrow p_2 \rightarrow \cdots \rightarrow p_n$, όπου κάθε σημείο αποτελείται από ένα σύνολο γεω-χωρικών συντεταγμένων και μια χρονική σήμανση όπως $p = (x, y, t)$.

Η πρόοδος στις τεχνολογίες απόκτησης τοποθεσίας έχει δημιουργήσει μια πληθώρα χωρικών τροχιών που αντιπροσωπεύουν την κινητικότητα διαφόρων κινούμενων αντικειμένων, όπως άνθρωποι, οχήματα και ζώα. Τέτοιες τροχιές προσφέρουν πρωτοφανείς πληροφορίες για την κατανόηση κινούμενων αντικειμένων και τοποθεσιών, προωθώντας ένα ευρύ φάσμα εφαρμογών σε κοινωνικά δίκτυα βάσει τοποθεσίας (Zheng 2011), ευφυή συστήματα μεταφοράς και αστικούς υπολογιστές (Zheng et al., 2014). Ο χρήση αυτών των εφαρμογών με τη σειρά του απαιτεί συστηματική έρευνα σε νέες τεχνολογίες υπολογιστών για την ανακάλυψη γνώσεων από δεδομένα τροχιάς. Υπό αυτές τις συνθήκες, η εξόρυξη δεδομένων τροχιάς έχει γίνει όλο και περισσότερο σημαντικό θέμα έρευνας.

Στην παρούσα ενότητα θα παρουσιαστούν τα βασικότερα στοιχεία για την εξόρυξη δεδομένων τροχιάς, τις τεχνικές που έχουν αναπτυχθεί για την ανάλυσή τους καθώς

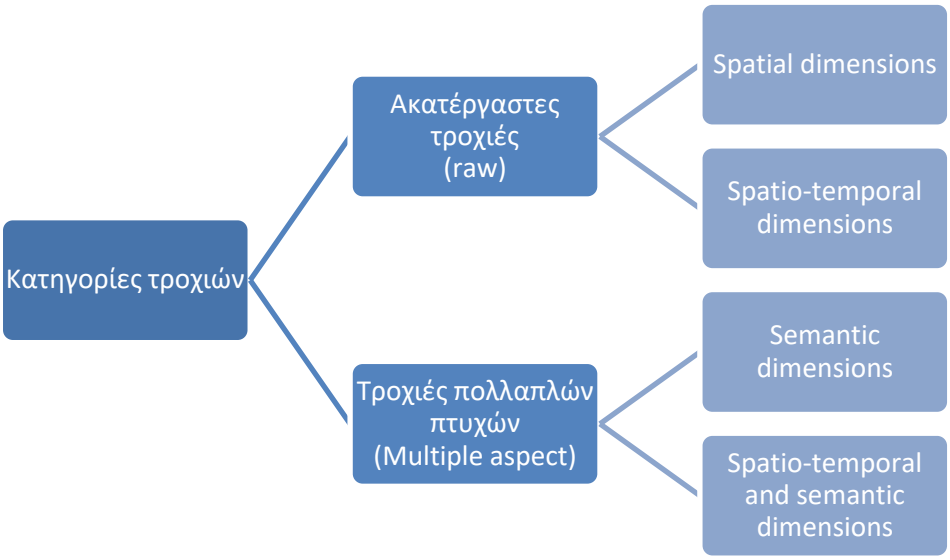
και οι βασικές μέθοδοι Μηχανικής Μάθησης που χρησιμοποιούνται κατά κόρον για την επεξεργασία τέτοιων δεδομένων.

2.1 Trajectory data and trajectory data mining

Η ιδέα μιας τροχιάς πηγάζει απ' την έννοια της κίνησης σωμάτων, παρέχοντας μια εικόνα της πορείας τους. Αυτό το θέμα, όμως, εφαρμόζεται σε πολλούς τομείς, στους οποίους είναι απαραίτητο είτε να εξαχθούν άμεσα αποτελέσματα κινήσεων (π.χ. GPS, άτομα, οχήματα, ζώα) ή έμμεσα, βρίσκοντας ιδιαίτερα χαρακτηριστικά τους, όπως η ταχύτητα τους, η τοποθεσία τους κάποια χρονική στιγμή. Με δεδομένες τις ανωτέρω ανάγκες και σε συνδυασμό με τη ραγδαία εξέλιξη των τεχνολογικών μέσων, οι τροχιές αποτελούν βασικό άξονα για ένα ευρύ φάσμα εφαρμογών με πιο συνήθη τα μέσα κοινωνικής δικτύωσης, τις ηλεκτρονικές εφαρμογές αλλά και τις «έξυπνες» μεταφορές. Τουτέστιν, η άντληση πληροφοριών από κινητά δεδομένα έχει ευρεία χρήση στις μέρες μας και αναμένεται να εξελιχθεί ακόμη περισσότερο στο μέλλον ικανοποιώντας τις ολοένα και αυξανόμενες ανάγκες και με τη συνεισφορά της σφαιρικότητας των γνώσεων που προέρχεται από ένα εκτενές φάσμα επιστημονικών.

2.1.1 Ορισμοί

Πριν παρουσιαστούν οι εργασίες που υπάρχουν βιβλιογραφικά για το ζήτημα της εξόρυξης και επεξεργασίας τροχιών, πρέπει πρώτα να δοθούν οι ορισμοί της τροχιάς και οι συνδεδεμένοι ορισμοί με αυτούς, ώστε να γίνουν κατανοητές οι μέθοδοι που παρουσιάζονται στα ακόλουθα κεφάλαια. Στην ακόλουθη εικόνα παρουσιάζεται ο διαχωρισμός των τροχιών, των οποίων οι ορισμοί θα παρατεθούν στην παρούσα ενότητα.



Σχήμα 2 : Κατηγορίες τροχιών.

Πηγή : Ferrero et al., 2020.

Η εικόνα μιας τροχιάς παρουσιάζεται στο Σχήμα 3. Η τροχιά ενός κινούμενου αντικειμένου είναι μια συνεχής συνάρτηση τ (t) του χρόνου t, έτσι ώστε δεδομένου ενός στιγμιαίου t, επιστρέφει τη θέση του κινούμενου αντικειμένου. Στην πραγματικότητα, η τροχιά του κινούμενου αντικειμένου καταγράφεται από ένα πεπερασμένο σύνολο παρατηρήσεων σε διακριτές χρονικές σημάνσεις t_1, t_2, ..., t_n. Ουσιαστικά μια τροχιά τ ενός κινούμενου αντικειμένου ορίζεται ως μια πολυγωνική γραμμή με n κορυφές. Προκειμένου να ορισθεί η γραμμή αυτή υποτίθεται ότι η ταχύτητα κίνησης του αντικειμένου κατά μήκος ενός τμήματος γραμμής είναι σταθερή. Πρέπει ωστόσο να σημειωθεί ότι για δύο τροχιές τ_1, τ_2, οι χρόνοι παρατήρησης θα μπορούσαν να είναι διαφορετικοί.

Σχήμα 3 : Παράδειγμα ακατέργαστης τροχιάς.

Πηγή : Ferrero et al., 2020.

Η διεξοδική επεξήγηση τροχιών επιδέχεται αρκετή μαθηματική ανάλυση στο υπόβαθρο της. Γι' αυτό και συνήθως χρησιμοποιείται η έννοια των διανυσμάτων όπου σε κάθε διάσταση τους παρουσιάζεται και ένα υποκείμενο χαρακτηριστικό όπως η θέση αντικειμένου (γεωγραφικές συντεταγμένες), η ταχύτητα του, η κατεύθυνσή του και οτιδήποτε άλλο ενδιαφέρει τον ερευνητή. Μια γνωστή μέθοδος που συναντάται συχνά στην Στατιστική Ανάλυση δεδομένων είναι η Principal Component Analysis (PCA), στην οποία είναι εφικτό να μειωθεί η διάσταση του προβλήματος και να βρεθούν οι κύριες συνιστώσες με μεθόδους «ορθογωνιοποίησης» των μεταβλητών (στην περίσταση μας τα βασικά χαρακτηριστικά μιας τροχιάς που αναλύεται με τα εκάστοτε δεδομένα) (Bashir et al., 2007). Επιπλέον, μια ακόμη αρκετά αποτελεσματική τεχνική είναι η Discrete Fourier Transformation (DFT), στην οποία μια τροχιά δύναται να ερμηνευθεί ως συνδυασμός συντελεστών Fourier με μικρό σφάλμα (Hu, 2013; Naftel & Khalid, 2006).

Σχήμα 4 : Παράδειγμα τροχιάς πολλαπλών πτυχών.

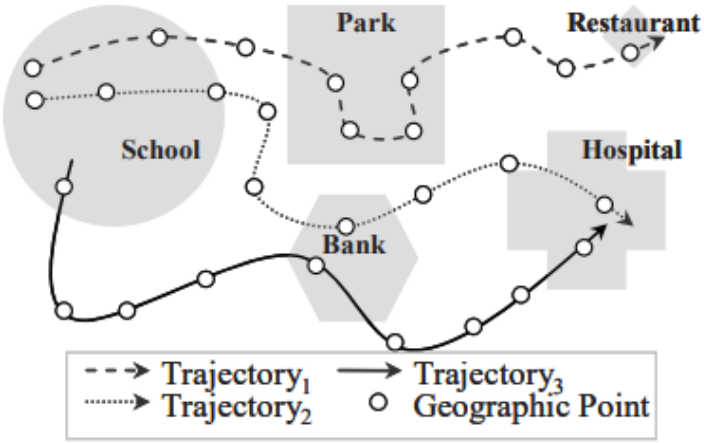
Πηγή : Ferrero et al., 2020.

Με δεδομένη μια τροχιά $T = \{p_1, p_2, ..., p_{ni}\}$ του μεγέθους n, μια υπο-τροχιά $s_T = \{p_a, p_b ..., \eta\ p_{mi}\}$ του T είναι μια συνεχόμενη ακολουθία όπου $a >= 1$, και $m <= n$. Γενικά, μία υπο-τροχιά αποτελεί ένα μέρος της συνολικής τροχιάς. Λαμβάνοντας υπόψη την έννοια της υπο-τροχιάς, οι τελευταίες έννοιες που απαιτούνται για την κατανόηση της κατηγοριοποίησης της τροχιάς είναι τα χαρακτηριστικά τους, τα οποία χωρίζονται σε καθολικά (global) και τοπικά. Ένα καθολικό χαρακτηριστικό είναι ένα

χαρακτηριστικό ή μοτίβο του οποίου το νόημα σχετίζεται με την τροχιά έναρξης. Λαμβάνοντας υπόψη μια τροχιά *T*, τα καθολικά χαρακτηριστικά είναι πληροφορίες που εξάγονται από ολόκληρη την τροχιά *T*. Τα χαρακτηριστικά αυτά μπορούν, για παράδειγμα, να είναι η μέση ταχύτητα μιας τροχιάς κατά τη διάρκεια ολόκληρης της κίνησης, μεταξύ άλλων. Από την άλλη πλευρά, τα τοπικά χαρακτηριστικά είναι τα χαρακτηριστικά ή τα μοτίβα που εξάγονται από τις υπο-τροχιές και η σημασία τους σχετίζεται μόνο με ένα τμήμα της τροχιάς.

Με δεδομένη την τροχιά *T*, τα τοπικά χαρακτηριστικά είναι μοτίβα που εξάγονται από υπο-τροχιά του *T*. Τα έργα των (Ferrero et al., 2018) και (Ferrero et al., 2020) πρότειναν τη χρήση σχετικών υπο-τροχιών ως τα χαρακτηριστικά για την κατηγοριοποίηση τροχιάς, όπου οι σχετικές υπο-τροχιές ονομάζονται κινητά (movelets), και η συνάφειά τους μετράται από την ικανότητά τους να διακρίνουν τις κλάσεις.

Τα τελευταία χρόνια, προέκυψε μια νέα φυλή μεθόδων πρόβλεψης τοποθεσίας, που ονομάζεται πρόβλεψη βάσει γενικού προτύπου. Τέτοιες μέθοδοι πρόβλεψης χρησιμοποιούν συνήθως τις συχνές κοινές συμπεριφορές των χρηστών που εξάγονται από συλλογές τροχιών GPS χρηστών για κινητές συσκευές, για να προβλέψουν την επόμενη κίνηση ενός χρήστη. Το Σχήμα 5 δείχνει μερικά παραδείγματα της τροχιάς GPS, η οποία συνήθως αποτελείται από μια ακολουθία χωρο-χρονικών σημείων (σε μορφή γεωγραφικού πλάτους, μήκους και χρόνου).



Σχήμα 5 : Παράδειγμα σημασιολογικής τροχιάς.

Πηγή : Ying et al., 2011.

Μεταξύ των μεθόδων πρόβλεψης που βασίζονται σε γενικά μοτίβα, χρησιμοποιήθηκαν ευρέως τεχνικές εξόρυξης διαδοχικών μοτίβων (Lu et al, 2009; Monreale et al, 2009) για την ανάλυση μοτίβων σε σύνολα δεδομένων κίνησης χρηστών κινητών. Ωστόσο, τείνουν να προβλέπουν δημοφιλείς τοποθεσίες όπου επισκέφτηκαν οι περισσότεροι, οδηγώντας στο πρόβλημα των «ανισόρροπων» δεδομένων. Επιπλέον, αυτές οι μέθοδοι πρόβλεψης βάσει προτύπων κάνουν συνήθως μια πρόβλεψη μόνο εάν μια αναμενόμενη κίνηση έχει πλήρη αντιστοιχία με το πρόθεμα ενός μοτίβου, οδηγώντας σε απώλεια ανάκλησης στις προβλέψεις.

Αν και τα ζητήματα της ανακάλυψης των συχνών μοτίβων των χρηστών κινητών στις τροχιές τους έχουν συζητηθεί στη βιβλιογραφία, οι υπάρχουσες μελέτες εξετάζουν κυρίως μόνο τα γεωγραφικά χαρακτηριστικά των τροχιών των χρηστών (Lu et al, 2009; Monreale et al, 2009). Μία γεωγραφική τροχιά αποτελείται συνήθως από μια ακολουθία γεωγραφικών σημείων (αντιπροσωπεύεται ως «γεωγραφικό πλάτος, μήκος») με ετικέτες χρονικών σημείων. Ως αποτέλεσμα, το συχνό μοτίβο της συμπεριφοράς κίνησης του χρήστη με βάση τη γεωγραφική τροχιά περιορίζεται από τις γεωγραφικές ιδιότητες των δεδομένων τροχιάς. Για παράδειγμα, όπως δείχνει το

Σχήμα 5, η γεωγραφική απόσταση και το σχήμα μεταξύ $Τροχιάς_1$ και $Τροχιάς_2$ είναι πιο κοντά και παρόμοια από εκείνη μεταξύ $Τροχιάς_1$ και $Τροχιάς_3$. Έτσι, ορισμένες τεχνικές πρόβλεψης θέσης θα προέβλεπαν τον προορισμό της $Τροχιάς_1$ με βάση τη γεωγραφική του ομοιότητα με την $Τροχιά_2$. Επιπλέον, τέτοιες στρατηγικές πρόβλεψης λαμβάνουν υπόψη μόνο τις τοποθεσίες που έχουν επισκεφθεί προηγουμένως και συνεπώς δεν λειτουργούν καλά όταν λαμβάνονται υπόψη τοποθεσίες που δεν έχουν επισκεφθεί προηγουμένως. Υποστηρίζουμε ότι η απλή χρήση γεωγραφικών πληροφοριών για την πρόβλεψη του προορισμού μιας τροχιάς ή την επόμενη τοποθεσία ενός χρήστη δεν επαρκεί.

Η έννοια της σημασιολογικής τροχιάς έχει προταθεί από τους Alvares et al. (2007). Μια σημασιολογική τροχιά (semantic trajectory) αποτελείται από μια ακολουθία τοποθεσιών που επισημαίνονται με σημασιολογικές ετικέτες (που ονομάζονται σημασιολογικές τοποθεσίες) για να καταγράψουν συγκεκριμένα ορόσημα. Αυτές οι σημασιολογικές ετικέτες τοποθεσιών υποδηλώνουν τις δραστηριότητες που διεξάγονται στην τροχιά.

Οι τροχιές που συλλαμβάνουν την ίδια συμπεριφορά κίνησης ενός χρήστη είναι πιθανό να έχουν κάποιες «ενδείξεις» που αναφέρονται σε αυτά τα χωρικά (spatial) και προσωρινά σημεία δεδομένων μεταξύ των τροχιών. Δεδομένου ότι οι τροχιές μπορεί να έχουν αθόρυβη διάρκεια, αυτά τα χωρικά και χρονικά (spatio-temporal points) συντεταγμένα σημεία δεδομένων πρέπει να προσδιορίζονται προσεκτικά και να χρησιμοποιούνται για να συναχθούν διαδρομές τροχιάς.

Στην πράξη, οι τροχιές λαμβάνονται με συσκευές ή σχήματα απόκτησης τροχιάς, τα οποία δυστυχώς ενδέχεται να εισάγουν χωρική και χρονική προκατάληψη στα σημεία δεδομένων των τροχιών. Για παράδειγμα, η ακρίβεια θέσης του GPS (Global-

Position System) έχει εγγενή χωρική προκατάληψη. Επιπλέον, οι χρόνοι εμφάνισης των σημείων δεδομένων που λαμβάνονται από την ίδια ακριβώς συμπεριφορά κίνησης δεν είναι πάντα οι ίδιοι. Για παράδειγμα, ένας εργαζόμενος πηγαίνει στο γραφείο του από το σπίτι του στις 8:00 π.μ. καθημερινά. Τα σημεία δεδομένων των τροχιών που καταγράφουν αυτήν τη συμπεριφορά κίνησης συνήθως δεν έχουν τον ίδιο χρόνο εμφάνισης. Ένας λόγος είναι ότι η συσκευή τοποθέτησης χρειάζεται λίγο χρόνο για να προσδιορίσει τη θέση. Έτσι, ακόμη και αν αυτός ο χρήστης φεύγει από το σπίτι του στις 8:00 π.μ. καθημερινά, τα σημεία δεδομένων των τροχιών ενδέχεται να έχουν κάποια χρονική προκατάληψη.

Μετά την παρουσίαση των βασικών ορισμών για τις κατηγορίες και τα χαρακτηριστικά των τροχιών, στην επόμενη υπο-ενότητα παρουσιάζεται η κατηγοριοποίηση των τροχιών, οι μέθοδοι που χρησιμοποιούνται βιβλιογραφικά και οι ορισμοί αυτών και της ομαδοποίησης των τροχιών.

2.1.2 Εξόρυξη δεδομένων τροχιάς

Στο σημείο αυτό αναφέρεται το θέμα που αφορά την εξόρυξη δεδομένων και κυρίως από τροχιές. Στόχος της διεργασίας αυτής είναι η εύρεση προτύπων μέσω των εκάστοτε δεδομένων. Μέσω αυτής επιτυγχάνονται δύο καταστάσεις που αφορούν την περιγραφή και την πρόβλεψη των δεδομένων, αντιστοίχως. Η περιγραφή αφορά κυρίως την ανακάλυψη δεδομένων που μπορούν με κάποιον τρόπο να ερμηνευθούν και να έχουν μια υπόσταση όταν μελετηθούν. Η πρόβλεψη συναντάται όταν θεωρείται πως κάποιες μεταβλητές μπορούν μέσω κάποιας συνάρτησης να προβλέψουν τις τιμές μερικών υπό εξέταση άγνωστων μεταβλητών.

Η εξόρυξη προτύπων τροχιάς στοχεύει στην ανακάλυψη και την περιγραφή των μοτίβων κίνησης που κρύβονται στις τροχιές. Παρέχει πληροφορίες σχετικά με το πότε και πού εμφανίζεται το μοτίβο, καθώς και για τις οντότητες που εμπλέκονται σε αυτό. Στη βιβλιογραφία έχουν αναφερθεί πολλοί τύπων κινήσεων. Μια ολοκληρωμένη εικόνα αυτών μπορεί να βρεθεί στην έρευνα που πραγματοποίησαν οι Dodge et al. (2008). Ομοίως, έχουν αναπτυχθεί διάφορες μέθοδοι για την εξόρυξη αυτών των προτύπων. Οι μέθοδοι μπορούν να χωριστούν σε τρεις κατηγορίες: 1) επαναλαμβανόμενη εξόρυξη προτύπων, 2) συχνή εξόρυξη προτύπων και 3) ομαδική εξόρυξη προτύπων.

Οι μέθοδοι εξόρυξης δεδομένων δύναται να διακριθούν σε δύο κύριες κατηγορίες. Στη μία συμπεριλαμβάνονται οι πρωτογενείς τεχνικές, οι οποίες αποσκοπούν στην κατηγοριοποίηση δεδομένων με κριτήριο τις ιδιότητες των τροχιών τους (ίσως λαμβάνοντας υπόψιν κάποια μετρική). Στην άλλη πτυχή ενυπάρχουν οι δευτερογενείς μέθοδοι που έχουν μια συνδυαστική έννοια καθώς εφαρμόζουν εν μέρει τεχνικές εξόρυξης ή τις συσχετίζουν μαζί με κλασσικές στατιστικές αναλύσεις. Ακόμη, δύο από αυτές είναι οι πιο σημαντικές, ο λόγος για τις «clustering» και «classification» (Ferrero et al., 2020).

Η επαναλαμβανόμενη εξόρυξη μοτίβων εφαρμόζεται βασικά σε τροχιές ενός κινούμενου αντικειμένου, ενώ η συχνή εξόρυξη μοτίβων εφαρμόζεται σε τροχιές πολλαπλών κινούμενων αντικειμένων αλλά σε σχετικό χρόνο. Δηλαδή, τα αντικείμενα ενδέχεται να μην κινούνται ταυτόχρονα και η μόνη προϋπόθεση είναι ότι επισκέπτονται περίπου τα ίδια μέρη στην ίδια σειρά. Παρόμοια με τη συχνή εξόρυξη προτύπων, η ομαδική εξόρυξη μοτίβων ισχύει για τροχιές πολλαπλών αντικειμένων, αλλά η κίνηση θεωρείται σε απόλυτο χρόνο, δηλαδή, τα αντικείμενα κινούνται μαζί. Η εξόρυξη επαναλαμβανόμενων προτύπων αφορά τα κανονικά μοτίβα κίνησης, όπως η κίνηση

ενός μετακινούμενου, η οποία επαναλαμβάνεται κάθε μέρα, ή η κίνηση ενός πειραματικού πουλιού, το οποίο επαναλαμβάνεται κάθε εποχή. Ένα επαναλαμβανόμενο μοτίβο ονομάζεται επίσης περιοδικό μοτίβο επειδή το αντικείμενο ακολουθεί περίπου την ίδια διαδρομή μετά από μια σχεδόν σταθερή χρονική περίοδο.

Η ανακάλυψη περιοδικών προτύπων περιπλέκεται από την κατά προσέγγιση φύση των προτύπων από άποψη χώρου και χρόνου. Δηλαδή, το αντικείμενο δεν επισκέπτεται ακριβώς την ίδια τοποθεσία στις αντίστοιχες χρονικές στιγμές της περιόδου και η περίοδος δεν έχει ακριβώς την ίδια τιμή σε διαφορετικούς κύκλους.

Βασικές εργασίες που σχετίζονται με την ανάλυση των τροχιών είναι η εξόρυξη μοτίβων τροχιάς, η αβεβαιότητα τροχιάς, η ανίχνευση ακραίων τιμών και η κατηγοριοποίηση.

- **Αβεβαιότητα τροχιάς**: Τα αντικείμενα κινούνται συνεχώς, ενώ οι θέσεις τους μπορούν μόνο να ενημερωθούν σε διακριτές ώρες, αφήνοντας αβέβαια τη θέση ενός κινούμενου αντικειμένου μεταξύ δύο ενημερώσεων. Για να βελτιωθεί η χρησιμότητα των τροχιών, μια σειρά από έρευνες προσπάθησαν να «μοντελοποιήσουν» και να μειώσουν την αβεβαιότητα των τροχιών. Αντίθετα, ένας κλάδος έρευνας στοχεύει στην προστασία του απορρήτου ενός χρήστη όταν ο χρήστης αποκαλύπτει τις τροχιές της.

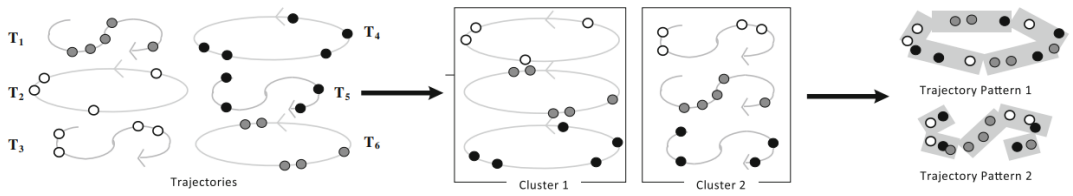
- **Εξόρυξη προτύπων τροχιάς**: Ο τεράστιος όγκος χωρικών τροχιών επιτρέπει ευκαιρίες για ανάλυση των μοτίβων κινητικότητας των κινούμενων αντικειμένων, τα οποία μπορούν να αναπαρασταθούν από μια μεμονωμένη τροχιά που περιέχει ένα συγκεκριμένο μοτίβο ή μια ομάδα τροχιών που μοιράζονται παρόμοια μοτίβα.

- **Κατηγοριοποίηση τροχιάς**: Χρησιμοποιώντας εποπτευόμενες μαθησιακές προσεγγίσεις, μπορούμε να ταξινομήσουμε τροχιές ή τμήματα μιας τροχιάς σε ορισμένες κατηγορίες, οι οποίες μπορεί να είναι δραστηριότητες (όπως πεζοπορία και φαγητό) ή διαφορετικοί τρόποι μεταφοράς, όπως περπάτημα και οδήγηση.

- **Ανίχνευση ακραίων τροχιών**: Διαφορετικά από τα μοτίβα τροχιάς που συμβαίνουν συχνά σε δεδομένα τροχιάς, οι ακραίες τροχιές (γνωστές ανωμαλίες) μπορεί να είναι αντικείμενα (τροχιά ή τμήμα τροχιάς) που διαφέρουν σημαντικά από άλλα στοιχεία όσον αφορά κάποια μέτρηση ομοιότητας. Μπορεί επίσης να είναι γεγονότα ή παρατηρήσεις (που αντιπροσωπεύονται από μια συλλογή τροχιών) που δεν συμμορφώνονται με το αναμενόμενο μοτίβο (π.χ. κυκλοφοριακή συμφόρηση που προκαλείται από τροχαίο ατύχημα).

2.1.3 Κατηγοριοποίηση τροχιών

Στην ενότητα αυτή, αναπτύσσονται στοιχεία για το ζήτημα της κατηγοριοποίησης τροχιάς (Trajectory Classification), το οποίο ανήκει στις πρωτογενείς μεθόδους εξόρυξης που αναφέρθηκαν στην προηγούμενη ενότητα. Είναι αντιληπτό ότι για να οριστεί με ορθό τρόπο το πρόβλημα που διατυπώνεται, απαιτείται μια προετοιμασία όσον αφορά κάποιες ορολογίες οι οποίες επικαλούνται συχνά στο μεγαλύτερο μέρος της εργασίας.



Σχήμα 6 : Παράδειγμα κατηγοριοποίησης τροχιάς.

Πηγή : Hung et al., 2015.

Αρχικά, ως Trajectory T θεωρείται μια ακολουθία στοιχείων από το σύνολο $<e_1,...,e_m>$, όπου κάθε στοιχείο e_i είναι ένα l-διάστατο διάνυσμα, της μορφής $e_i = (d_{1,i}, ..., d_{l,i})$ με κάθε $d_{k,i}$ να αντιπροσωπεύει την τιμή μιας παραμέτρου k του αντικειμένου για τη χρονική στιγμή i (Furtado et al., 2015).

Τώρα, ως Trajectory Classification, ορίζουμε το πρόβλημα της κατηγοριοποίησης κινούμενων τροχιών T_j ως την εκπαίδευση μιας εκτιμήτρια συνάρτησης f η οποία θα δέχεται σαν όρισμα T_j που ανήκουν στο $\boldsymbol{T} = \{(T_1, classT_1), ..., (T_n, classT_n)\}$ και θα το αντιστοιχεί στην πιο ακριβή επισήμανση από τις προαναφερθείσες $classT_i$ (Tan et al., 2005; Lee et al., 2008).

Στο σημείο αυτό αναφέρεται η βασική διαφορά που έχει το trajectory classification σε σχέση με την απλή κατηγοριοποίηση δεδομένων (Data Classification). Η κατηγοριοποίηση δεδομένων μπορεί να επιτευχθεί με βάση κατάλληλο κατηγοριοποιητή, ο οποίος μέσα από μια διαδικασία εκπαίδευσης προσφέρει το βέλτιστο αποτέλεσμα. Στη κατηγοριοποίηση τροχιών είναι πολύ πιθανόν να επιτευχθεί μέσω βέλτιστων χαρακτηριστικών τα οποία δεν απαιτούν ολόκληρη τη τροχιά αλλά μέρος της. Έτσι, ως λογικό επακόλουθο πρέπει να καθοριστεί η έννοια του υποσυνόλου μιας τροχιάς (Ferrero et al., 2020).

Η κατηγοριοποίηση τροχιάς έχει ως καθήκον τον εντοπισμό του κινούμενου αντικειμένου που εκτέλεσε μια δεδομένη τροχιά (Lee et al., 2008). Στη βιβλιογραφία

υπάρχουν πολλά έργα που εκτελούν κατηγοριοποίηση τροχιάς για διαφορετικούς σκοπούς: εύρεση ενός επιπέδου τυφώνα, καθορισμός ενός είδους ζώου, πρόβλεψη ενός τρόπου μεταφοράς, αναγνώρισης χρήστη και ούτω καθεξής.

Πριν παρουσιαστεί η κατηγοριοποίηση της τροχιάς, είναι απαραίτητο να κατανοηθεί η έννοια της κατηγοριοποίησης. Στην εξόρυξη δεδομένων, η κατηγοριοποίηση είναι ένα σημαντικό και πολύ διερευνημένο θέμα που στοχεύει στη διάκριση των κλάσεων σε ένα σύνολο δεδομένων (Nikam et al., 2015). Ένα σύνολο δεδομένων αποτελείται από πολλά στοιχεία, όπου κάθε στοιχείο είναι ένα σύνολο χαρακτηριστικών ή χαρακτηριστικών και κάθε στοιχείο έχει μια ετικέτα κλάσης. Ο στόχος των αλγορίθμων κατηγοριοποίησης είναι να εκπαιδεύσει ένα μοντέλο ικανό να εκχωρήσει τη σωστή ετικέτα των στοιχείων χωρίς ετικέτα σε ένα σύνολο δεδομένων, με το μικρότερο σφάλμα κατηγοριοποίησης. Τα μοντέλα κατηγοριοποίησης εκπαιδεύονται χρησιμοποιώντας τα επισημασμένα στοιχεία ενός συνόλου δεδομένων, πράγμα που σημαίνει ότι οι τεχνικές κατηγοριοποίησης χρειάζονται δύο σύνολα δεδομένων: ένα εκπαιδευτικό (train_set) και ένα σύνολο δοκιμών (test_set).

Το σετ κατάρτισης (train_set) αποτελείται από τα στοιχεία ενός συνόλου δεδομένων των οποίων οι κλάσεις είναι γνωστές από τον αλγόριθμο κατηγοριοποίησης, ο οποίος χρησιμοποιεί το σετ προπόνησης για την πρόκληση ενός μοντέλου κατηγοριοποίησης. Το σύνολο δοκιμών (test_set), αντ 'αυτού, αποτελείται από στοιχεία των οποίων οι κλάσεις είναι άγνωστες από τους αλγόριθμους κατηγοριοποίησης και χρησιμοποιούνται για την επικύρωση του εάν ένα μοντέλο κατηγοριοποίησης είναι καλό ή όχι για την κατηγοριοποίηση στοιχείων χωρίς ετικέτα.

Ένας άλλος όρος που σχετίζεται με την ανάλυση των τροχιών είναι η ομαδοποίηση. Η ομαδοποίηση δεδομένων τροχιάς (clustering) διενεργείται όταν είναι επιθυμητό να

διαχωρίσουμε τα δεδομένα σε κάποιες κατηγορίες. Πιο συγκεκριμένα, η εφαρμογή της πραγματοποιείται για τη διάκριση στις αποκαλούμενες συστάδες ως προς κάποιες ιδιότητες τους. Ουσιαστικά, δεδομένα που ευρίσκονται στην ίδια ομάδα παρουσιάζουν κοινά χαρακτηριστικά που είναι ορατά είτε με κάποια απλή συνάρτηση είτε με κάποιο σύνθετο μετασχηματισμό. Επιπλέον, είναι σημαντικό να υπάρχει μια ένδειξη που να διαχωρίζει αυτές τις κατηγορίες από τις υπόλοιπες. Μερικές εφαρμογές της διαδικασίας αυτής αποτελούν τα μοντέλα, Hidden Markov Model (HMM), DBSCAN (density based spatial clustering of applications with noise), OPTICS (ordering points to identify the clustering structure), αλγόριθμοι όπως η Διαμέριση (partitioning), οι Ιεραρχικοί (hierarchical) και με βάση την πυκνότητα (density-based), BIRCH (balanced iterative reducing and clustering hierarchies) (Bian et al., 2019).

Η κατηγοριοποίηση σε δεδομένα τροχιάς αποσκοπεί στην εύρεση μιας τακτικής σύμφωνα με την οποία τοποθετούνται τα αντικείμενα σε συγκεκριμένες κλάσεις. Ειδικότερα, ενυπάρχουν προκαθορισμένες κλάσεις και ένα μέρος των αντικειμένων το οποίο έχει κατηγοριοποιηθεί ήδη σε αυτές. Αυτό το σύνολο αποτελεί το κομμάτι μέσω του οποίου εκπαιδεύεται ο αλγόριθμος μας (training set) με σκοπό την εύρεση ενός κατηγοριοποιητή. Ο τελευταίος συνδράμει στην «ετικετοποίηση» (labeling) των δεδομένων (στην περίπτωση μας δεδομένων τροχιάς), κάτι που επιτυγχάνεται λαμβάνοντας υπόψιν τα χαρακτηριστικά των δεδομένων των τροχιών. Ουσιαστικά, ο σκοπός είναι να «ταμπελοποιηθούν» (labeled) όλα τα δεδομένα. δεδομένου ότι είναι κατηγοριοποιημένο ήδη ένα σύνολο εκπαίδευσης εκ των προτέρων.

Στην πλειονότητα των αλγορίθμων για κατηγοριοποίηση τροχιάς χρησιμοποιείται μια κλασσική τακτική. Αρχικά, προκύπτουν κάποια χαρακτηριστικά, τα οποία στη συνέχεια επαναχρησιμοποιούνται βοηθώντας τον αλγόριθμο να δημιουργήσει ένα μοντέλο με καλύτερα αποτελέσματα. Αυτός ο σπόρος σε πρώτη φάση ανακύπτει με

βάση πρότερη γνώση που είναι γνωστή για τα δεδομένα ώστε η κατηγοριοποίηση να είναι η βέλτιστη δυνατή για τις κλάσεις και για την κάθε τροχιά ξεχωριστά. Επίσης, αυτές οι τιμές των ιδιοτήτων των τροχιών, συνήθως, παρουσιάζονται με τη μορφή διανυσμάτων (στην ενότητα 2.3 δίδεται ο αυστηρός ορισμός).

Συχνά πυκνά επιλέγονται και εφαρμόζονται αλγόριθμοι με σκοπό την κατηγοριοποίηση τροχιών. Ένας από αυτούς αφορά τα δένδρα απόφασης (decision tree algorithm), ο οποίος έχει χρησιμοποιηθεί σε άρθρο του (Zheng et al., 2010) για την κατηγοριοποίηση τροχιών σε μέσα μεταφοράς. Ουσιαστικά μέσα από ένα σύνολο κανόνων εφαρμοσμένων στα εκπαιδευτικά δεδομένα κατασκευάζεται ένας αλγόριθμος που αποφασίζει για την καλύτερη τοποθέτηση μιας δειγματικής παρατήρησης στην κλάση της. Σε πρώτη φάση υπάρχει μονάχα η αρχική τιμή του δεδομένου στο δέντρο. Έπειτα πραγματοποιούνται συγκρίσεις μεταξύ των δυνατών επιλογών και ακολουθείται ο κλάδος που βελτιστοποιεί το χαρακτηριστικό που εξετάζεται. Μετά γίνεται η μεταφορά στον επόμενο κόμβο όπου επαναλαμβάνεται η ίδια διεργασία με το αρχικό βήμα. Έτσι, εν τέλει, προκύπτει μια πορεία όπου τερματίζεται όταν είναι η καλύτερη δυνατή σε σχέση με οποιαδήποτε άλλη (Ferrero et al., 2020).

Τα δέντρα απόφασης μπορούν να διακριθούν ανάλογα με το output που πρόκειται να εξάγουν. Με αυτόν τρόπο υπάρχουν αυτά που καταλήγουν σε :

- μια κατηγορική μεταβλητή απόκρισης
- μια μεταβλητή που είναι συνεχής 2) δένδρο απόφασης συνεχούς μεταβλητής (το δένδρο απόφασης έχει μια συνεχή μεταβλητή στόχου).

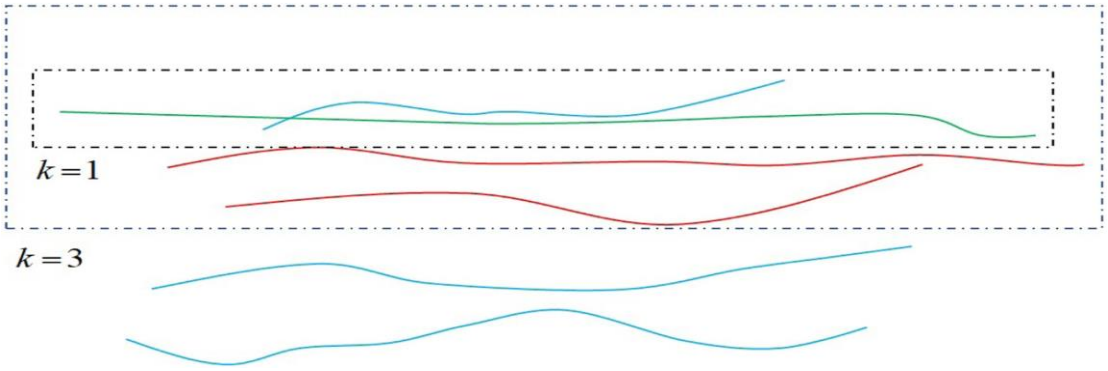
Αξίζει να αναφερθεί και το παράδειγμα του αλγορίθμου SVM (Support Vector Machine). Σε αυτή την εφαρμογή, ο κατηγοριοποιητής επιχειρεί να κατηγοριοποιήσει τα δεδομένα σε έναν χώρο με N διαστάσεις (κάθε διάσταση είναι και μια μεταβλητή -

χαρακτηριστικό) με τη μέγιστη ακρίβεια που μπορεί να επιτευχθεί αυτό. Στην περίπτωση που απαιτείται ο διαχωρισμός να υλοποιηθεί σε δύο κατηγορίες οι εφικτές λύσεις στο πρόβλημα είναι αρκετές. Όμως η βέλτιστη ευρίσκεται με βάση τη μεγιστοποίηση του κριτηρίου διάκρισης, δηλαδή επιλέγοντας δύο σημεία από διαφορετικές κλάσεις θα πρέπει να έχουν τη μέγιστη δυνατή «απόσταση» (ως προς κάποια προκαθορισμένη μετρική). Τουτέστιν, είναι λογικό επακόλουθο να πραγματοποιείται μια εξαιρετικής ακρίβειας κατηγοριοποίηση (Bian et al., 2019).

Ουκ ολίγες είναι οι περιπτώσεις που η κατηγοριοποίηση ακολουθεί ύστερα από άλλου είδους διαδικασία. Συνήθως εφαρμόζεται κάποια μέθοδος προετοιμασίας των δεδομένων όπως είναι η τμηματοποίηση ή ομαδοποίηση τους με σκοπό να επικαλεστεί η κατηγοριοποίηση με ορθότερο τρόπο και ακριβέστερα αποτελέσματα. Ακόμη, παρατηρούνται μέθοδοι εξόρυξης δεδομένων που καλούνται ως δευτερεύουσες. Σε αυτές περιλαμβάνονται διεργασίες όπως είναι το Pattern mining, το Outlier detection ή και το Prediction.

Σε πολλά πεδία η εξόρυξη δεδομένων τροχιάς αποτελεί βασικό τρόπο αντιμετώπισης καίριων ζητημάτων. Για παράδειγμα, αναλύοντας και ανακαλύπτοντας σχέσεις μεταξύ των πορειών των κινούμενων αντικειμένων, προκύπτουν κατηγοριοποιήσεις με βάση κρυφά χαρακτηριστικά και ιδιότητες των τελευταίων. Μια άλλη αντίστοιχη περίπτωση αποτελεί η διεξοδική ανάλυση τέτοιων πορειών ως προς την αναφορά τους σε κάποιες περιοχές, συνήθως μέσω check – in μέσω των μέσων κοινωνικής δικτύωσης, με σκοπό τη κατηγοριοποίηση της περιοχής ανάλογα με την επισκεψιμότητα και την επιδίωξη βαθμολόγησης της. Ακόμη μια εφαρμογή συναντάται για την εύρεση πιθανών συγκεντρώσεων με βάση την πορεία των τροχιών ατόμων αλλά ίσως και μια εκτίμηση του είδους της συγκέντρωση με βάση τις ιδιότητες τους και άλλες παραμέτρους για την ορθότερη κατηγοριοποίηση (Ferrero et al., 2020).

Ένας ακόμη, εναλλακτικός τρόπος κατηγοριοποίησης είναι και αυτός που εφαρμόζεται με κριτήριο τις αποστάσεις που εμφανίζουν τα δεδομένα. Ένας άλλος τρόπος για την κατηγοριοποίηση των τροχιών χρησιμοποιεί μέτρα απόστασης των δεδομένων στηριγμένα σε κάποια μετρική (δηλαδή έναν τρόπο μέτρησης της απόστασης). Η πιο συνήθη μορφή είναι η κατηγοριοποίηση πλησιέστερων γειτονικών σημείων (k-nn). Για την εκτέλεση αυτής της μεθόδου χρειάζονται τροχιές αποτελούμενες από διαδοχικά δεδομένα στα οποία συνιστάται μια επιλογή σύγκρισης τους μέσω αποστάσεων. Πάντως σαν τρόπος κατηγοριοποίησης λειτουργεί εξαιρετικά, μιας και μπορεί να διαχειριστεί έντονα διαφοροποιημένες μεταβλητές (Ferrero et al., 2020).



Σχήμα 7 : k-nn μέθοδος για κατηγοριοποίηση τροχιών.

Πηγή : Bian et al., 2019.

Τα πιο γνωστά μέτρα που δημιουργήθηκαν για την κατηγοριοποίηση δεδομένων που βασίζονται σε μια ακολουθίες χαρακτηριστικών είναι :

- Dynamic Time Warping (DTW) (Berndt and Clifford, 1994)
- Longest Common Subrocences (LCSS) (Vlachos et al., 2002)
- Edit Distance for Real Sequences (EDR) (Chen et al., 2005)
- Multidimensional Similarity Measure (MSM) (Furtado et al., 2015).

Σχετικά με το μέτρο DTW, αναφέρεται κυρίως σε χρονοσειρές και αριθμητικά δεδομένα. Η ιδέα του βασίζεται στην εύρεση της βέλτιστης γραμμής μεταξύ δυο ακολουθιών από τροχιές (μη γραμμικές). Ενίσχυση του παραπάνω μέτρου επήλθε μέσω των MD-DTW και DTWa (Holt et al., 2007; Shokoohi-Yekta et al.,2017). Στην πρώτη περίπτωση εφαρμόζεται μια μετρική που εξυπηρετεί πολλές διαστάσεις χωρίς να λαμβάνεται διαφορετικό μέτρο σε καθεμιά. Στην άλλη περίπτωση διακρίνεται μια καλύτερη προσέγγιση στην κατηγοριοποίηση (αφού υπάρχουν μετρικές, ορίζονται αυτόματα και σφάλματα).

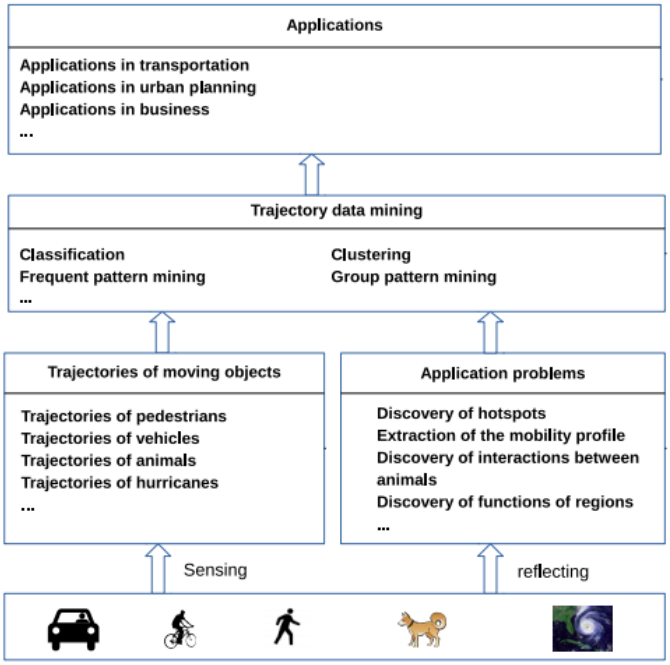
Η μετρική LCSS αναζητεί τη πιο μεγάλη όμοια ακολουθία μεταξύ μερικών τροχιών. Ουσιαστικά, με αυτό τον τρόπο ταιριάζει τις τροχιές που παρουσιάζουν κοινή υπακολουθία μέγιστου πλήθους χαρακτηριστικών. Όσον αφορά το EDR, προσπαθεί με δεομένη τη μια τροχιά μέσω διεργασιών να καταλήξει στην άλλη. Στη διαδικασία αυτή εν τέλει επιλέγεται κοινή ταμπέλα στα ζεύγη τροχιών με τις λιγότερες μετατροπές από τη μία στην άλλη (σαν να επιλέγεται η πιο απλή «απεικόνιση»). Ακόμη, το MSM υποθέτοντας ανεξαρτησία όλων των μεταβλητών χρησιμοποιείται για ετερογενείς διαστάσεις και μπορεί να συνυπολογίσει κυρίως σημασιολογικές ενδείξεις (C. A. Ferrero et al., 2020).

Γενικώς το βασικό ζήτημα σε κατηγοριοποιήσεις ανομοιόμορφων διαστάσεων είναι το γεγονός ότι δεν ενδείκνυται να δίδουμε την ίδια βαρύτητα σε όλες όσον αφορά τη χρήση μετρικών ή αποστάσεων. Λύση σε αυτό το πρόβλημα συνήθως πραγματοποιείται με τον καθαρισμό βαρών ανά διάσταση. Προφανώς, τα υψηλότερα εφαρμόζονται στις πιο σημαντικές μεταβλητές ενώ στις λιγότερο ενδιαφέρουσες για το θέμα που εξετάζεται, προσδίδονται αμελητέα βάρη. Σχετικά με τα προηγούμενα μέτρα : στο DTW επιβάλλονται βάρη ενώ στα LCSS, EDR καθορίζονται μόνο κατώτατα όρια.

Στο μέτρο MSM που φαίνεται να εξυπηρετεί περισσότερο τους σκοπούς της εργασίας μας, προκαθορίζονται και βάρη αλλά και κατώτατα όρια στη μέθοδο των αποστάσεων.

2.1.4 Μελέτες του παρελθόντος

Στο σημείο αυτό αναφέρονται μελέτες του παρελθόντος είτε για το ζήτημα που αναπτύσσεται στην υποκείμενη εργασία, *κατηγοριοποίηση κινούμενων αντικειμένων*, είτε σχετικών θεμάτων. Εξετάζουμε το έργο που έγινε από την πλευρά της εφαρμογής σύμφωνα με το πλαίσιο που φαίνεται στο ακόλουθο σχήμα. Υπάρχουν διάφοροι τύποι κινούμενων αντικειμένων που μπορούν να παρακολουθούνται. Η παρακολούθηση αυτών των αντικειμένων δημιουργεί ένα σύνολο τροχιών και μερικές ερωτήσεις, των οποίων οι λύσεις μπορούν να χρησιμοποιηθούν σε ορισμένα πεδία εφαρμογής. Καλούμε αυτά τα ερωτήματα προβλήματα εφαρμογής, επειδή οι λύσεις τους δεν είναι εφαρμογή από μόνες τους, αλλά οι γνώσεις που χρησιμοποιούνται σε πεδία εφαρμογής. (Mazimpaka, J. D., & Timpf, S et al., 2016)



Σχήμα 8 : Γενικό πλαίσιο της εξόρυξης δεδομένων τροχιάς βάσει εφαρμογών.

Πηγή : Mazimpaka, J. D., & Timpf, S., 2016.

Τα προβλήματα εφαρμογής επιλύονται χρησιμοποιώντας συγκεκριμένες μεθόδους εξόρυξης τροχιάς. Για παράδειγμα, εάν παρακολουθούνται οι κινούμενοι άνθρωποι παίρνουμε τις τροχιές τους και μερικές ερωτήσεις έρχονται στο μυαλό μας σχετικά με, για παράδειγμα, τα χαρακτηριστικά αυτών των ανθρώπων ή τον γεωγραφικό χώρο όπου μετακινούνται. Με την εξόρυξη αυτών των τροχιών, ενδέχεται να ανακαλύψουμε μέρη όπου πολλοί άνθρωποι σταματούν συχνά και μένουν για μια αρκετά μεγάλη διάρκεια. Αυτά τα μέρη χαρακτηρίζονται ως «hotspots» και οι γνώσεις σχετικά με «hotspots» μπορούν να χρησιμοποιηθούν, για παράδειγμα, σε επιχειρήσεις για διαφήμιση, σε δημόσια ασφάλεια για παρακολούθηση του πλήθους και σε μεταφορές για πρόταση παραλαβής ταξί. Επομένως, η ανακάλυψη «hotspots» είναι ένα πρόβλημα εφαρμογής του οποίου η λύση μπορεί να έχει διάφορες εφαρμογές.

Καθολικά χαρακτηριστικά εφαρμόζονται ευρέως για την κατηγοριοποίηση πολλών διεργασιών επί του θέματος, όπως στη μελέτη (Xiao et al., 2017), όπου διενεργείται κατηγοριοποίηση διάφορων μέσων μεταφοράς. Ιδιαίτερο ενδιαφέρον παρουσιάζουν και οι έρευνες που χρησιμοποιούν ειδικά χαρακτηριστικά για classification. Τέτοια μπορεί να είναι η ταχύτητα, επιτάχυνση, απόσταση που διανύουν τα κινητά αντικείμενα (Dodge et al., 2009; Zheng et al., 2010; Patel et al., 2012). Οι ανωτέρω μέθοδοι επικαλούνται κυρίως μέσω χωρικών συντεταγμένων και έτσι δεν απαιτείται καμία επεξεργασία των δεδομένων τους, τα οποία μπορούν να χρησιμοποιηθούν κατευθείαν για την κατηγοριοποίηση τους.

Μερικές ακόμη εφαρμογές στην περίπτωση ακατέργαστων δεδομένων (raw data) συναντώνται σε μελέτες για τυφώνες (Ferrero et al., 2018), σε σύνολα ζώων (Rowland et al, 1997), σε κατηγοριοποίηση οχημάτων (Frentzos et al., 2005) αλλά και μέσα μεταφοράς (Zheng et al., 2010). Μια καινοτόμος μέθοδος προτείνεται στο άρθρο (Ferrero et al., 2018), η οποία καλείται Movelets. Εκεί αναφέρεται μια παραμετρική

τεχνική, η οποία μεν πλεονεκτεί στην ύπαρξη πολλών διαστάσεων, αλλά δε, μειονεκτεί καθώς δεν έχει ικανότητα διάκρισης «σημαντικών» συνιστωσών του προβλήματος που μελετάται. Πιο συγκεκριμένα, είναι πιθανόν να εμφανίζονται χαρακτηριστικά που είναι πιο βασικά για την κατηγοριοποίηση του συγκεκριμένου προβλήματος και βασίζονται σε συγκεκριμένες συνιστώσες χωρίς να είναι εφικτό μέσω της μεθόδου να διαχωριστούν από τις υπόλοιπες.

Στο σημείο αυτό, γίνεται αναφορά σε μια ιδιαίτερη κατηγορία τροχιών, αυτές που εντοπίζονται στα κοινωνικά δίκτυα. Σε αυτές εκλείπει ο ποσοτικός χαρακτήρας καθώς έχουν αραιά δομή σχετικά με τη προηγούμενη μορφολογία που αναπτύχθηκε. Γι' αυτό και τα χαρακτηριστικά τους δεν δύναται να έχουν την ίδια αντιμετώπιση με τα δεδομένα που συλλέγονται από GPS. Βέβαια, μπορούν να αντληθούν δεδομένα με άλλο περιεχόμενο (σημασιολογικής επισήμανσης) με την ισχύ των οποίων αναλύονται ιδιότητες όπως η τοποθεσία, η κατηγορία του μέρους, η τιμή του, η βαθμολογία του, οι κριτικές του.

Μια σημαντική μελέτη έχει υλοποιηθεί γύρω από αυτό το ζήτημα (Gao et al., 2017). Διερευνήθηκαν τροχιές που αναζητήθηκαν από τα μέσα κοινωνικής δικτύωσης, αλλά χωρίς την εύρεση αλληλεπιδράσεων μεταξύ διαφορετικών σημασιολογικών διαστάσεων. Μάλιστα, στο άρθρο αυτό, προτάθηκε ένα μοντέλο κατηγοριοποίησης, το BiTULER, το οποίο μέσα από τεχνικές νευρωνικών δικτύων, αξιοποιεί τις εισόδους – λέξεις προκαλώντας μια μορφή κατηγοριοποίησης. Βέβαια, μειονεκτεί στον περιορισμό μη εξαγωγής αποτελεσμάτων σε κινησιολογικά χαρακτηριστικά των δεδομένων.

2.2 Μέθοδος TraClass

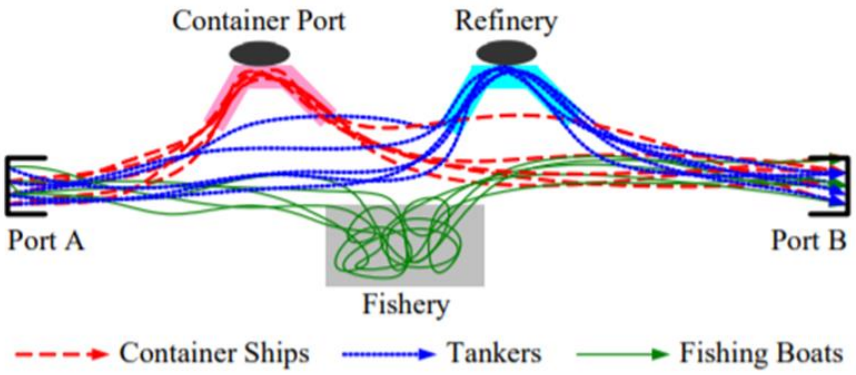
Στην παρούσα υπο-ενότητα παρουσιάζεται η μέθοδος TraClass. Η κατηγοριοποίηση τροχιάς (παραδείγματος χάρη, η κατασκευή ενός μοντέλου που αποσκοπεί στην πρόβλεψη των ετικετών κλάσης των κινούμενων αντικειμένων με βάση τις τροχιές τους και άλλα χαρακτηριστικά) έχει πολλές σημαντικές εφαρμογές. Υπάρχουν διάφορες μέθοδοι και τεχνικές, αλλά λόγω της χρήσης των σχημάτων ολόκληρων τροχιών για κατηγοριοποίηση, έχουν περιορισμένη ικανότητα κατηγοριοποίησης όταν εμφανίζονται διακριτά χαρακτηριστικά σε τμήματα των τροχιών ή δεν σχετίζονται με τα σχήματα των τροχιών. Αυτές οι περιπτώσεις συναντώνται συχνά σε μεγάλες τροχιές, οι οποίες απλώνονται σε μεγάλες γεωγραφικές περιοχές.

Δεδομένου ότι ένα βασικό έργο για μια αποτελεσματική κατηγοριοποίηση είναι η δημιουργία διακριτών χαρακτηριστικών, η μέθοδος TraClass (Lee et al., 2008b) δημιουργεί μια ιεραρχία χαρακτηριστικών, διαχωρίζοντας τις τροχιές και διερευνώντας δύο είδη ομαδοποίησης : α) με βάση την περιοχή (region-based) β) με βάση την τροχιά (trajectory-based). Το πρώτο είδος ομαδοποίησης συλλαμβάνει τα υψηλότερου επιπέδου χαρακτηριστικά βάσει περιοχής χωρίς την χρήση μοτίβων κίνησης. Το δεύτερο είδος ομαδοποίησης συλλαμβάνει τα χαμηλότερα επιπέδου χαρακτηριστικά με βάση την τροχιά, χρησιμοποιώντας μοτίβα κίνησης.

Η τεχνική TraClass ξεπερνάει τους περιορισμούς προηγούμενων μελετών, καθώς ο διαχωρισμός της τροχιάς καθιστά αναγνωρίσιμα τα διακριτά μέρη των τροχιών, και τα δύο είδη ομαδοποίησης συνεργάζονται έτσι ώστε να βρουν χαρακτηριστικά και βάσει των περιοχών αλλά και βάσει των τροχιών. Αρχικά, τα διακριτά χαρακτηριστικά είναι πιθανόν να εμφανίζονται σε μέρη της τροχιάς και όχι σε ολόκληρη την τροχιά.

Επιπλέον, τα διακριτά χαρακτηριστικά εμφανίζονται όχι μόνο σαν κοινά μοτίβα κίνησης, αλλά και σαν περιοχές (regions).

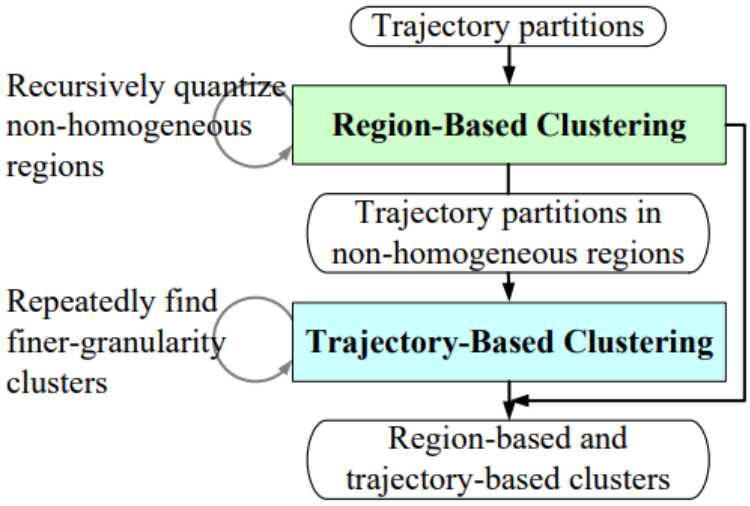
Πειραματικά αποτελέσματα έχουν αποδείξει ότι η μέθοδος TraClass δημιουργεί υψηλού επιπέδου χαρακτηριστικά (features) και επιτυγχάνει υψηλή ακρίβεια κατηγοριοποίησης πάνω σε πραγματικά δεδομένα τροχιάς. Για παράδειγμα με βάση το Σχήμα 9, δύναται να γίνει διακριτή η διαφορά ανάμεσα σε πλοία κοντέινερ και δεξαμενόπλοια λόγω διακρίσεων-μερών των τροχιών κοντά στη θύρα εμπορευματοκιβωτίων (Container Port) και στα διυλιστήρια (Refinery), ακόμη και αν μοιράζονται κοινές μεγάλες διαδρομές. Επίσης τα αλιευτικά σκάφη (Fishing Boats) δύναται να αναγνωριστούν από τις τροχιές τους κοντά στον αλιευτικό χώρο (Fishery) ακόμα κι αν δεν μοιράζονται κάποιο κοινό μονοπάτι. Η συνεργασία μεταξύ των δύο τύπων ομαδοποίησης οδηγεί στην ανακάλυψη και των δύο τύπων διακριτών χαρακτηριστικών, αυξάνοντας έτσι σημαντικά την ακρίβεια της κατηγοριοποίησης. Δεδομένου ότι ο διαχωρισμός της τροχιάς προηγείται της ομαδοποίησης, τα διακριτά μέρη των τροχιών καθίστανται αναγνωρίσιμα.



Σχήμα 9: Δύο είδη διακριτών χαρακτηριστικών για τροχιές.

Πηγή : Lee et al., 2008b.

Όπως έχει παρουσιαστεί από τον Gonzalez (2008), η μέθοδος TraClass, πριν από την ομαδοποίηση, χωρίζει την κάθε τροχιά σε ένα σύνολο κατατμήσεων τροχιάς. Πρώτον, η ομαδοποίηση με βάση την περιοχή πραγματοποιείται αναδρομικά, αρκεί να βρεθούν ομοιογενείς περιοχές λογικού μεγέθους. Τα διαχωριστικά τροχιάς που δεν καλύπτονται από ομοιογενείς περιοχές περνούν στο επόμενο βήμα. Δεύτερον, η ομαδοποίηση με βάση την τροχιά εκτελείται επανειλημμένα όσο βρίσκονται διαχωριστικές ομάδες. Προφανώς, αυτή η συνεργατική ιεραρχική ομαδοποίηση μας επιτρέπει να εντοπίζουμε περισσότερα χαρακτηριστικά υψηλής ποιότητας.



Σχήμα 10: Η διαδικασία της ομαδοποίησης με βάση την ιεραρχική περιοχή και την τροχιά.

Πηγή : Gonzalez et al., 2008.

Αξίζει να σημειωθεί (Gonzalez et al., (2008)), ότι η εξερεύνηση των δύο τύπων ομαδοποίησης δημιουργεί μια ιεραρχία χαρακτηριστικών, επειδή η ομαδοποίηση βάσει περιοχής εντοπίζει χαρακτηριστικά υψηλότερου επιπέδου (πιο γενικά) από ό,τι η ομαδοποίηση με βάση την τροχιά, επειδή δεν χρησιμοποιεί μοτίβα κίνησης. Επιπλέον, σε κάθε σύμπλεγμα, υπάρχει μια ιεραρχία από μεγαλύτερες ομάδες έως μικρότερες. Συνολικά, δημιουργούνται δυνατότητες από πάνω προς τα κάτω: οι λειτουργίες

υψηλότερου επιπέδου προτιμώνται από αυτές χαμηλότερου επιπέδου, δεδομένου ότι οι πρώτες είναι πιο αποτελεσματικές για την κατηγοριοποίηση και, ταυτόχρονα, φθηνότερες στη χρήση από την τελευταία. Χαρακτηριστικά παρουσιάζονται οι δυνατότητες και λειτουργίες που γίνονται σε κάθε επίπεδο:

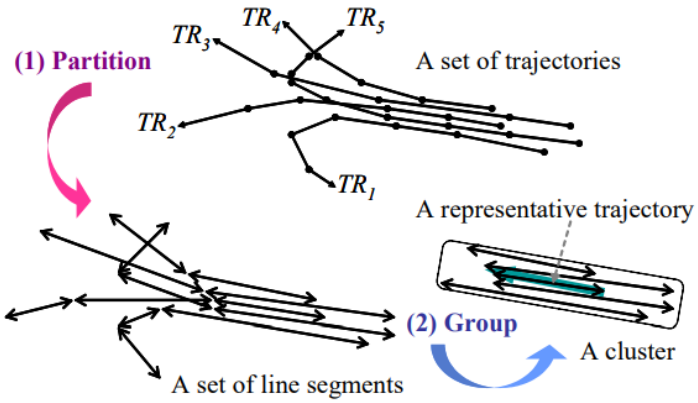
Α) Διαμέριση τροχιάς (Trajectory partitioning)

Η διαμέριση τροχιάς καθώς και η ομαδοποίηση βάσει τροχιάς βασίζονται στο πλαίσιο διαμέρισης και ομάδας (Lee et al., 2007), το οποίο αποτελείται από τις ακόλουθες δύο φάσεις:

(1) Η φάση διαμέρισης: Κάθε τροχιά χωρίζεται σε ένα σύνολο τμημάτων γραμμής (δηλαδή διαχωριστικά τροχιάς) κάθε φορά που η κατεύθυνση κίνησης αλλάζει γρήγορα.

(2) Φάση ομαδοποίησης: Παρόμοια τμήματα γραμμών ομαδοποιούνται σε ένα σύμπλεγμα χρησιμοποιώντας μια μέθοδο συμπλέγματος βάσει πυκνότητας ανάλογη με το DBSCAN. Μια συνάρτηση απόστασης για δύο τμήματα γραμμών έχει σχεδιαστεί για να καθορίσει την πυκνότητα. Οι βασικές έννοιες της ομαδοποίησης με βάση την πυκνότητα για σημεία αλλάζουν σε εκείνα τα τμήματα γραμμών.

Όπως έχει παρουσιαστεί από τον Gonzalez (2008), στο τελικό στάδιο της φάσης ομαδοποίησης, ένα μοντέλο που ονομάζεται «αντιπροσωπευτική τροχιά», το οποίο είναι μια ακολουθία σημείων όπως μια συνηθισμένη τροχιά, δημιουργείται για κάθε συστάδα. Πρόκειται για μια φανταστική τροχιά που δείχνει το κύριο μοτίβο κίνησης των διαμερισμάτων τροχιάς που ανήκουν στο σύμπλεγμα και λαμβάνεται με τον υπολογισμό των μέσων συντεταγμένων αυτών των κατατμήσεων τροχιάς. Η συνολική διαδικασία απεικονίζεται στο ακόλουθο σχήμα.

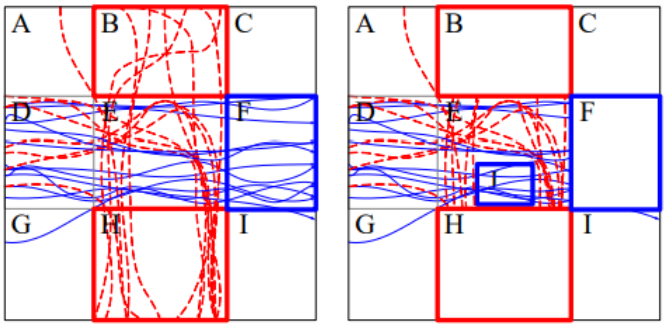


Σχήμα 11: Μια συνολική διαδικασία ομαδοποίησης τροχιάς στο πλαίσιο διαμέρισης και ομάδας.

Πηγή : Lee et al., 2008b.

Β) Ομαδοποίηση περιοχών (Region-Based Clustering)

Ένα σύμπλεγμα με βάση την περιοχή περιέχει πολλά τμήματα τροχιάς μιας μεγάλης κλάσης, αλλά πολύ λίγα διαχωριστικά τροχιών άλλων δευτερευουσών τάξεων. Παράδειγμα ομαδοποίησης βάσει περιοχών παρουσιάζεται στο ακόλουθο σχήμα.



Σχήμα 12: Παράδειγμα ομαδοποίησης βάσει περιοχών.

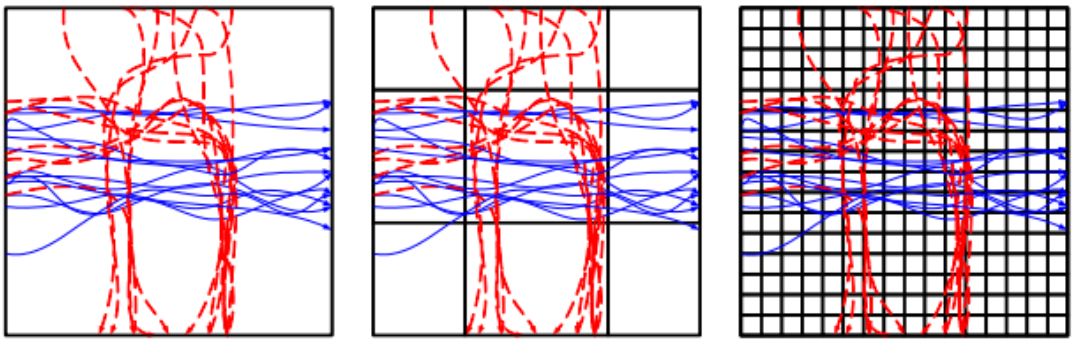
Πηγή : Lee et al., 2008b.

Για να βρεθούν όσο το δυνατόν περισσότερες ομοιογενείς περιοχές, η μέθοδος TraClass χρησιμοποιεί μια δομή πλέγματος πολλαπλών αναλύσεων, η οποία ποσοτικοποιεί τον χώρο του τομέα σε έναν πεπερασμένο αριθμό κελιών. Οι άξονες X

και Y χωρίζονται και στη συνέχεια δημιουργούνται κελιά διασχίζοντας τα διαχωριστικά των αξόνων X και Y. Μετά την ποσοτικοποίηση, κάθε περιοχή (δηλαδή, κελί) εξετάζεται για να προσδιοριστεί εάν είναι ομοιογενής.

Ένας καλός κβαντισμός του χώρου θα πρέπει να έχει δύο επιθυμητές ιδιότητες: ομοιογένεια και συνοπτικότητα. Ομοιογένεια σημαίνει ότι η κατανομή της τάξης σε κάθε περιοχή πρέπει να είναι όσο το δυνατόν πιο ομοιογενής. απαιτείται για την παραγωγή περισσότερων συστάδων βάσει περιοχής. Εν συντομία σημαίνει ότι ο αριθμός των περιοχών πρέπει να είναι όσο το δυνατόν μικρότερος.

Η ομοιογένεια και η συνοχή είναι μέτρα αντιπαλότητας. Εάν ολόκληρος ο χώρος περιορίζεται σε μία περιοχή όπως στο Σχήμα 13 (α), η ομοιογένεια μπορεί να γίνει χαμηλότερη, αλλά η συνοχή γίνεται υψηλότερη. Αντίθετα, εάν ο χώρος του τομέα κβαντοποιηθεί σε πολλές μικρές περιοχές έτσι ώστε να περικλείουν το πολύ ένα τμήμα τροχιάς όπως στο Σχήμα 13 (γ), η ομοιογένεια γίνεται υψηλότερη, αλλά η συνοχή γίνεται χαμηλότερη. Έτσι, είναι απαραίτητο να βρεθεί μια καλή ανταλλαγή μεταξύ των δύο ιδιοτήτων όπως στο Σχήμα 13 (β).



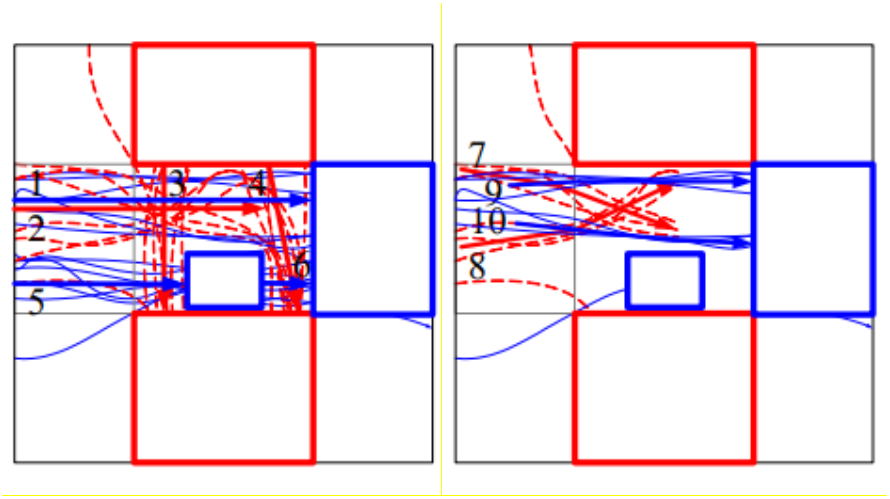
(α) (β) (γ)

Σχήμα 13: Σύγκριση ομοιογένειας και συνοχής μεταξύ δύο ακραίων περιπτώσεων.

Πηγή : Lee et al., 2008b.

C) Σύγκλιση με βάση την τροχιά (Trajectory-Based Clustering)

Όπως έχει παρουσιαστεί από τον Gonzalez (2008), ένα σύμπλεγμα με βάση την τροχιά είναι ένα σύνολο διαμερισμάτων τροχιάς που συνδέονται με την πυκνότητα της ίδιας κλάσης. Αυτός ο ορισμός είναι ουσιαστικά ο ίδιος με τον αρχικό ορισμό ενός συμπλέγματος τροχιάς. Εξ ορισμού, ένα σύμπλεγμα που βασίζεται σε τροχιά θα πρέπει να αναπτυχθεί χρησιμοποιώντας κατατμήσεις τροχιάς της ίδιας κλάσης. Το πραγματικό ενδιαφέρον εδώ είναι ότι ένα σύμπλεγμα προέρχεται από μία μόνο κατηγορία. Είναι προφανές ότι οι κοινές υπο-τροχιές που αποτελούνται από διαφορετικές κατηγορίες δεν έχουν καμία χρησιμότητα για την κατηγοριοποίηση. Παράδειγμα της σύγκλισης βάσει τροχιάς παρουσιάζονται παρακάτω.



Σχήμα 14: Παράδειγμα ομαδοποίησης βάσει τροχιών.

Πηγή : Lee et al., 2008b.

D) Επανάληψη εκτέλεσης και επιλογή τιμής παραμέτρου

Μόλις ένα σύμπλεγμα βάσει τροχιάς επιλεγεί, τα χωρίσματα τροχιάς στο σύμπλεγμα δεν λαμβάνονται υπόψη στη μετέπειτα ομαδοποίηση, δεδομένου ότι καλύπτονται ήδη από ένα σύμπλεγμα διακρίσεων. Έτσι, η μέθοδος TraClass (Lee et al., 2008b) εκτελεί

επανειλημμένα ομαδοποίηση μόνο σε ακάλυπτα χωρίσματα τροχιάς προκειμένου να δημιουργήσει μεγαλύτερο αριθμό μικρότερων συστάδων προσαρμόζοντας τις τιμές των παραμέτρων. Αυτή η επαναλαμβανόμενη ομαδοποίηση επιτρέπει την εύρεση περισσότερων (πιθανώς διακριτών) ομάδων. Δεδομένου ότι απαιτείται ένα εύλογο διάστημα τιμών παραμέτρων, ακολουθείται ευρετικός τρόπος για τον προσδιορισμό των αρχικών τιμών παραμέτρων και τον τρόπο προσαρμογής τους.

Ε) Δημιουργία συνδέσμου συμπλέγματος

Όπως έχει παρουσιαστεί από τον Gonzalez (2008), εκτός από την παραγωγή συστάδων που βασίζονται σε πολύ διακριτικές τροχιές, στον εν λόγω αλγόριθμο έχει αναπτυχθεί μια τεχνική για την περαιτέρω βελτίωση της ακρίβειας της κατηγοριοποίησης. Λόγω της διαδοχικής φύσης των δεδομένων τροχιάς, μια ακολουθία συστάδων είναι πιθανό να περιέχει σημαντικές πληροφορίες για χρήση στην κατηγοριοποίηση. Τα συνδυασμένα χαρακτηριστικά δημιουργούνται από ένα σύνολο συστάδων. Ένα συνδυασμένο χαρακτηριστικό ορίζεται ως μια ακολουθία συνδεδεμένων συμπλεγμάτων ή συνδυασμένων λειτουργιών.

Συμπερασματικά, η μέθοδος TraClass κάνει κατηγοριοποίηση σε απλά χωρο-χρονικά δεδομένα, σε αντίθεση με την μέθοδο MasterMovelets (Ferrero et al., 2020) η οποία είναι εμπλουτισμένη με πολλαπλές και ετερογενείς διαστάσεις.

2.3 Τεχνικές Μηχανικής Μάθησης

Στο υπο-κεφάλαιο αυτό αναλύονται οι τεχνικές της Μηχανικής Μάθησης που σχετίζονται άμεσα ή έμμεσα με το ζήτημα της κατηγοριοποίησης των δεδομένων δίνοντας έμφαση σε εκείνα που αναφέρονται ως τροχιές. Η Μηχανική Μάθηση

αποτελεί υποκατηγορία ενός ευρύτερου τομέα που καλείται Τεχνητή Νοημοσύνη (AI). Σε αυτά τα πλαίσια δίνεται η δυνατότητα στις μηχανές να εκπαιδεύονται και να μειώνουν την πιθανότητα σφαλμάτων σε ικανοποιητικό βαθμό με σκοπό να αντιλαμβάνονται ενδεχόμενα πρότυπα που ίσως ενυπάρχουν στα χαρακτηριστικά των δεδομένων.

Σημαντικό ρόλο διαδραματίζει η ποιότητα αλλά και ο σωστός τρόπος δειγματοληψίας ή ανάκτησης των εκάστοτε δεδομένων ώστε η εκπαίδευση να πραγματοποιηθεί με τον ορθότερο τρόπο. Τα μοτίβα που προκύπτουν μέσα από τους αλγορίθμους αυτούς ερμηνεύονται είτε ως συναρτήσεις είτε ως όρια αποφάσεων. Ακόμη, η διαδικασία αυτή προσδίδει μείωση του υπολογιστικού χρόνου και του όγκου των πληροφοριών, μιας και δεν απαιτούνται νέα δεδομένα για τις αναλύσεις του αλγορίθμου. Γενικά οι τρόποι με τους οποίους ενδείκνυται να προκληθούν κατηγοριοποιήσεις σε δεδομένα, πόσο μάλλον κινούμενα, βασίζονται σε τρεις θεμελιώδεις κατηγορίες αλγορίθμων (Bian et al., 2019).

Η πρώτη πτυχή αποτελείται από τις διαδικασίες που αποκαλούνται μη επιβλεπόμενες (unsupervised). Στην προκειμένη περίπτωση δεν απαιτούνται ετικέτες συνοδευτικές των υπό επεξεργασία δεδομένων. Σκοπός της τεχνικής αυτής είναι να αποκαλύψει κρυφές δομές και όχι να προβλέψει μια επιθυμητή έξοδο (output). Ακόμη, στα μη εποπτευόμενα μοντέλα ενυπάρχουν και δύο κατηγορίες ομαδοποίησης των δεδομένων : το clustering και το association.

Όσον αφορά το clustering (ομαδοποίηση) εφαρμόζεται συνήθως όταν είναι αναγκαίο να προκληθεί μια ομαδοποίηση των δεδομένων ως προς κάποιο χαρακτηριστικό (γνωστό ή μη). Μερικές μέθοδοι που ανήκουν σε αυτό το πεδίο είναι το PCA, K-means, DBSCAN, mix models (Ferrero et al., 2020).

Η επομένη κατηγορία αφορά τους λεγόμενους αλγόριθμους εποπτευόμενης μάθησης (supervised). Σε αυτούς τα δεδομένα απαιτείται να έχουν ετικέτες ή επισημάνσεις (αν όχι σε όλα, σε μερικά από αυτά). Ο τρόπος που λειτουργούν είναι πολύ συγκεκριμένος. Αρχικά εκπαιδεύονται από το μέρος των δεδομένων που έχει ετικέτες (labels). Έπειτα, χρησιμοποιούν αυτή τη γνώση με σκοπό να προβλέψουν σχετικά γεγονότα στο μέλλον. Όμως, το πλήθος των δεδομένων που θα περιληφθούν στο σύνολο εκπαίδευσης (train set) είναι σημαντικό να αποφασιστεί έπειτα από ελέγχους στο test set (που είναι το σύνολο πάνω στο οποίο ελέγχεται αν η εκπαίδευση είναι αποτελεσματική μέσω έλεγχο σφαλμάτων της εξόδου). Με αυτό τον τρόπο, το μοντέλο δύναται να παρέχει αναμενόμενη έξοδο ή ετικέτα για οποιαδήποτε νέα είσοδο υπό την προϋπόθεση «επαρκούς» εκπαίδευσης. Επιπλέον, μια γνωστή τεχνική ελέγχου της αποτελεσματικότητας του μοντέλου είναι η σύγκριση των αναμενόμενων αποτελεσμάτων με τα πραγματικά και συνήθως με βάση το σφάλμα (ground truth label) επαν-εκπαιδεύεται για να βελτιωθεί μέχρι να ελαχιστοποιηθεί αυτή η απόκλιση (back-propagation).

Οι επιβλεπόμενης μάθησης αλγόριθμοι συναντώνται συχνά σε μεθόδους παλινδρόμησης (regression) και κατηγοριοποίησης (classification). Με τον όρο παλινδρόμηση γίνεται αναφορά σε προβλήματα που η επιθυμητή έξοδος λαμβάνει πραγματικές συνεχείς τιμές (ποσοτικές) (συνήθεις εφαρμογές σε SVC, LDA SVR , regression random forests) ενώ στην περίπτωση της κατηγοριοποίησης το output είναι κατηγορίες. Στην τελευταία περίπτωση, λοιπόν, απαιτείται κωδικοποίηση των επιπέδων για τους σκοπούς της στατιστικής ανάλυσης οπότε θεωρητικά το σύνολο τιμών της εξόδου θα είναι ένα πεπερασμένο σύνολο τιμών, με την καθεμία να αντιστοιχεί σε μια υπαρκτή συνθήκη (Ferrero et al., 2020).

Η τελευταία, αλλά συνάμα αρκετά σημαντική, τεχνική μηχανικής μάθησης που αναπτύσσεται είναι η ημι εποπτευόμενη (semi-supervised). Ουσιαστικά είναι μια μεθόδευση που αντλεί στοιχεία τόσο από την επιβλεπόμενη μάθηση όσο και από τη μη επιβλεπόμενη μάθηση. Πιο συγκεκριμένα, χρησιμοποιούνται δεδομένα με labels αλλά και χωρίς labels με σκοπό την εκπαίδευση του αλγορίθμου. Το πλήθος των δεδομένων με ετικέτα συνήθως είναι μικρό και ανατροφοδοτείται συνεχώς και με δεδομένα χωρίς επισήμανση με σκοπό τη βελτιστοποίηση του μοντέλου. Η διαδικασία ολοκληρώνεται μόλις η ακρίβεια του τελευταίου μοντέλου υπερβαίνει εκείνη των αντίστοιχων του. Γενικώς, αποτελεί τη καλύτερη μέθοδο καθώς μειώνεται ο υπολογιστικός χρόνος και συγχρόνως διακρίνεται για την ακρίβεια της (σε σχέση με μια unsupervised) και δεν αντιμετωπίζεται το πρόβλημα της υπερπροσαρμογής (overfitting).

Η ημι εποπτευόμενη μάθηση εφαρμόζεται συχνά σε προβλήματα δεδομένων τροχιάς. Μερικά χαρακτηριστικά παραδείγματα αποτελούν οι περιπτώσεις κατηγοριοποίησης σε περιεχόμενα διαδικτύου αλλά και στη βιολογία με τις αλληλουχίες του DNA. Στην πρώτη περίπτωση η ετικετοποίηση (labeling) των ιστοσελίδων είναι μη εφικτή και αρκετά δύσκολη διαδικασία οπότε η προτίμηση αλγορίθμων ημι επίβλεψης είναι μονόδρομος. Μάλιστα, ακόμη και σε μηχανές αναζήτησης (όπως της Google) πραγματοποιείται χρήση τέτοιων μεθόδων ώστε να δημιουργηθεί μια μορφή κατάταξης των ιστοσελίδων με βάση της λέξεις κλειδιά που έχουν καταγραφεί στο πλαίσιο του search. Εκτεταμένη εφαρμογή τέτοιων αλγορίθμων υλοποιείται και στις αλληλουχίες του DNA, μιας και οι κλώνοι τους χαρακτηρίζονται για το μέγεθος τους οπότε η ανάπτυξη τέτοιων διαδικασιών αποτελεί επιτακτική αναγκαιότητα (Bian et al., 2019).

2.4 Μετρικές αποστάσεων

Στο συγκεκριμένο υπο-κεφάλαιο επιχειρείται η επισκόπηση βασικών τρόπων μέτρησης «αποστάσεων» των δεδομένων. Ο σκοπός αυτών των τύπων είναι να δημιουργηθεί ένα μέτρο που θα συνδράμει στην προσπάθεια να αντιμετωπιστούν οι τροχιές με κοινό συγκριτικό κανόνα.

- *Ευκλείδεια απόσταση*, είναι ένα μέτρο που ορίζεται σε ευκλείδειο χώρο (όπως τον R^N) και ορίζεται με τον παρακάτω τρόπο στην περίπτωση δύο τροχιών i, j

  $D(Trajectory_i, Trajectory_j) = \frac{1}{N} \sum_{n=1}^{N} \left[\left(x_n^i - x_n^j\right)^2 + \left(y_n^i - y_n^j\right)^2 \right]^{\frac{1}{2}}$,

  όπου το N είναι το πλήθος των χαρακτηριστικών των δύο τροχιών και για σταθερό n κάθε φορά η σύγκριση πραγματοποιείται μεταξύ όμοιων συντεταγμένων ανά διάσταση x_n^i *με* x_n^j και y_n^i *με* y_n^j.

Για να βρεθούν διακριτικές υπο-τροχιές ορίζεται η απόσταση ανάμεσα σε δύο υποκατηγορίες. Αυτή η απόσταση πρέπει να λαμβάνει υπόψη τις διαφορετικές διαστάσεις. Δεδομένου ότι ένα στοιχείο (σημείο) μπορεί να έχει πολλαπλές και ετερογενείς διαστάσεις, τυπικά ορίζεται η έννοια της απόστασης μεταξύ των στοιχείων.

- **Διάνυσμα απόστασης μεταξύ δύο πολυδιάστατων στοιχείων** (Ferrero et al., 2020).. Δεδομένου ότι έχουμε δύο στοιχεία *ei* και *ej* που αντιπροσωπεύονται από *d* διαστάσεις, η απόσταση μεταξύ δύο πολυδιάστατων τα στοιχεία *dist(*ei , *ej)* επιστρέφουν ένα διάνυσμα απόστασης *V = (v1, v2, ..., vd),* όπου κάθε *vk = dist_ek (ei , ej)* είναι η απόσταση μεταξύ δύο στοιχείων στη διάσταση *k*, που σέβεται την ιδιότητα της συμμετρίας *dist_ek (ei , ej) = dist_ek (ej , ei)*. Η ιδέα πίσω από αυτόν τον ορισμό είναι να επιτρέπεται η χρήση μιας συνάρτησης απόστασης για κάθε διάσταση, και αποθήκευση των τιμών απόστασης όλων

των διαστάσεων σε ένα διάνυσμα απόστασης. Αυτό η προσέγγιση διαφέρει από τη μέθοδο Movelets (Ferrero et al. 2018), όπου η συνάρτηση *dist(ei, ej)* επιστρέφει μια τιμή απόστασης, χάνοντας τις λεπτομέρειες σχετικά με την απόσταση κατά μήκος των διαστάσεων. Για το λόγο αυτό, διατηρούνται οι τιμές της απόστασης μεταξύ των διαστάσεων σε ένα διάνυσμα απόστασης που είναι θεμελιώδες για τον υπολογισμό της απόστασης μεταξύ δύο υπο-τροχιών ίσου μήκους, διατηρώντας τις αποστάσεις σε κάθε διάσταση.

- **Διάνυσμα απόστασης μεταξύ δύο υπο-τροχιών ίσου μήκους** (Ferrero et al., 2020). Δεδομένου δύο υπο-τροχιών *s* και *r* και οι δύο μήκους *w* και *d* διαστάσεων, *dist_s(s,r),* υπολογίζεται η κατά ζεύγη απόσταση μεταξύ των στοιχείων υπο-τροχιάς σε διάνυσμα απόστασης *V = (v1, v2, . . . , vd),* όπου κάθε *vk* είναι η τιμή απόστασης μεταξύ *s* και *r* στη διάσταση *k.* Κάθε απόσταση *vk* σέβεται την ιδιότητα της συμμετρίας *dist_sk (s, r) = dist_sk (r , s).*

Η εύρεση της υπο-τροχιάς που είναι η πιο παρόμοια με μια δεδομένη υπο-τροχιά είναι ένα άλλο ουσιαστικό μέρος της μεθόδου. Η πιο παρόμοια υπο-τροχιά μιας τροχιάς *T* προς μια υπο-τροχιά *s* ονομάζεται καλύτερη ευθυγράμμιση (best alignment) και είναι η υπο-τροχιά *r* από το *T* με το ελάχιστη απόσταση σε *s*. Αυτή η σύγκριση δίνεται παρακάτω ως εξής:

- **Διάνυσμα απόστασης μεταξύ τροχιάς και υποτροχίας** (Ferrero et al., 2020). Δεδομένης μιας τροχιάς *T* και μιας υπο-τροχιάς *s* μήκους *w = |s|*, η απόσταση μεταξύ τους είναι η καλύτερη ευθυγράμμιση του *s* στο *T*, το οποίο ορίζεται από το $W^s_T = min_{r \in S^w_T}$ *(dist_s(s,r)),* όπου S^w_T είναι το σύνολο όλων των υποκατευθύνσεων μήκους *w* σε *T* , και η min() επιστρέφει το διάνυσμα

απόστασης της καλύτερης ευθυγράμμισης μεταξύ του s και όλων των υποκατευθύνσεων στο S^w_T.

3. Μέθοδος MasterMovelets

Τα τελευταία χρόνια η κατηγοριοποίηση τροχιάς έχει εφαρμοστεί σε πάρα πολλά πραγματικά προβλήματα, κυρίως όμως λαμβάνοντας υπόψιν τις διαστάσεις του χώρου και το χρόνου, ή τα χαρακτηριστικά τα οποία συνάγονται από αυτές τις διαστάσεις. Ωστόσο, η ακατάπαυστη συλλογή δεδομένων από τα ταχύτατα αναπτυσσόμενα κοινωνικά δίκτυα (Foursquare, Twitter, Facebook, Internet channels όπως το Weather Wunderground) , καθώς και η συνεχής πρόοδος στον σημασιολογικό εμπλουτισμό των δεδομένων κίνησης, επέφεραν την γέννηση ενός νέου τύπου δεδομένων τροχιάς. Τα χωρο-χρονικά σημεία μιας τροχιάς έχουν πλέον πολλαπλές και ετερογενείς σημασιολογικές διαστάσεις.

Η εκτόξευση των κοινωνικών δικτύων επέφεραν τον εμπλουτισμό των τροχιών με περισσότερη σημασιολογική και ετερογενή πληροφορία. Για παράδειγμα στο Σχήμα 15, η τροχιά ενός ατόμου ξεκινάει από το σπίτι του, έπειτα κατευθύνεται στο χώρο εργασίας του χρησιμοποιώντας το αυτοκίνητο του, και στην συνέχεια επισκέπτεται ένα εστιατόριο για να γευματίσει. Κατά τη διάρκεια της κίνησής του, οι καιρικές συνθήκες μεταβάλλονται δύο φορές, από βροχή σε συννεφιά και από συννεφιά σε ηλιοφάνεια, πράγμα το οποίο οδηγεί το άτομο να διανύσει μέρος της τροχιάς του είτε με το αυτοκίνητο (βροχή) είτε με τα πόδια (ηλιοφάνεια). Επιπρόσθετα, χρησιμοποιεί διάφορα κοινωνικά δίκτυα για να «ποστάρει» τα συναισθήματά του. Ένα τέτοιο είδος τροχιάς ονομάζεται τροχιά πολλαπλών πτυχών (Multiple Aspect Trajectory), όπου η τροχιά του ατόμου εμπλουτίζεται με πολλαπλές και ετερογενείς διαστάσεις δεδομένων, περιλαμβάνοντας χρονικά σημεία, συντεταγμένες, την κατηγορία του μέρους επίσκεψης, το όνομα του μέρους, την βαθμολογία του, τις τιμές του και τις διάφορες καιρικές συνθήκες.

Σχήμα 15 : Παράδειγμα τροχιάς πολλαπλών πτυχών.

Πηγή : Ferrero et al., 2020.

3.1 Ορισμός MasterMovelets

Σε αυτήν την ενότητα παρουσιάζεται ο αλγόριθμος MasterMovelets (Ferrero et al., 2020) (Multiple ASpect TrajEctoRy Movelets) ο οποίος θα χρησιμοποιηθεί στην παρούσα εργασία. Ο Ferrero πρότεινε την μέθοδο αυτή για την ανακάλυψη υπο-τροχιών χωρίς την ανάγκη ενός προκαθορισμένου κριτηρίου διαχωρισμού.

Η τεχνική MasterMovelets (Ferrero et al., 2020) υποστηρίζει πολλαπλές διαστάσεις, όπως ο χώρος, ο χρόνος, και οι σημασιολογικές πληροφορίες («semantics»). Το κύριο πρόβλημα που αντιμετωπίζει είναι ότι λαμβάνονται υπόψιν όλες οι διαστάσεις μαζί για την εύρεση σχετικών υπο-τροχιών. Για παράδειγμα, οι διαστάσεις συντήκονται όλες σε ένα διάνυσμα *(A = {α1,α2,…,αk})*. Αυτή η σύντηξη δύναται να αποκρύψει διακριτές διαστάσεις, οι οποίες αν εξεταστούν ξεχωριστά, θα είχαν την δυνατότητα να διακρίνουν την κλάση («class») καλύτερα. Η μέθοδος αυτή καλείται να λύσει το πρόβλημα της ευθυγράμμισης των κινητών, το οποίο ορίζεται ως εξής: «δεδομένης μιας υπο-τροχιάς και μιας τροχιάς, η καλύτερη ευθυγράμμιση της υπο-τροχιάς σε σχέση με την τροχιά συνίσταται στην εύρεση του πιο παρόμοιου (πλησιέστερου) τμήματος της τροχιάς προς την υπο-τροχιά» (Ferrero et al., 2020). Στις τροχιές πολλαπλών πτυχών, μια κλάση δύναται να χαρακτηριστεί με τον καλύτερο τρόπο από μια υπο-τροχιά χρησιμοποιώντας τις διαστάσεις του χώρου και του χρόνου.

Αντίθετα, μια άλλη κλάση μπορεί να χαρακτηριστεί καλύτερα από μια υπο-τροχιά με μία διάσταση. Γίνεται εμφανές και από το Σχήμα 15, ότι όσες περισσότερες διαστάσεις έχει μια τροχιά, τόσο πιο δύσκολο είναι να βρεθεί η υπο-τροχιά με τις καλύτερες διαστάσεις για την διάκριση των κλάσεων κίνησης.

Η μέθοδος MasterMovelets (Ferrero et al., 2020) που προτείνεται για την ανακάλυψη σχετικών υπο-τροχιών για κατηγοριοποίηση τροχιάς πολλαπλών πτυχών, βρίσκει αυτόματα τον καλύτερο συνδυασμό μήκους και διάστασης για την διάκριση της κλάσης. Η χρήση της τεχνικής αυτής είναι γενική και μπορεί να λάβει χώρα σε πολυδιάστατες εφαρμογές. Η κύρια διαφορά της με την «ακατέργαστη» τροχιά («raw trajectory») είναι ο αριθμός και ο τύπος των διαστάσεων. Μια τροχιά πολλαπλών πτυχών αποτελεί μια ακολουθία στοιχείων $(e_1, e_2,..., e_m)$, όπου κάθε στοιχείο έχει τις διαστάσεις x, y, t, A, όπου x και y αντιστοιχούν στη θέση του αντικειμένου στο διάστημα t και $A = \{\alpha_1,\alpha_2,...,\alpha_k\}$ είναι ένα σύνολο σημασιολογικών διαστάσεων και ονομάζονται πτυχές ή χαρακτηριστικά. Ως σημασιολογική διάσταση ορίζεται κάθε είδους πληροφορία που δεν είναι ούτε χωρική ούτε χρονική.

Η μέθοδος αυτή περιλαμβάνει την εξερεύνηση όλων των υποψηφίων υπο-τροχιών από ένα σύνολο δεδομένων τροχιάς και την επιλογή μόνο των κινητών, ώστε να εξετάσει κάθε υπο-τροχιά με την εύρεση της καλύτερης ευθυγράμμισης τους με όλες τις τροχιές στο σύνολο δεδομένων. Προκειμένου να βρεθεί η βέλτιστη ευθυγράμμιση, η μέθοδος αυτή υπολογίζει σε ένα τελικό στάδιο μια βαθμολογία συνάφειας.

Μετά από αυτό, τα επιλεγμένα κινητά χρησιμοποιούνται ως είσοδος για την εκπαίδευση ενός κατηγοριοποιητή τροχιάς. Το σύνολο δεδομένων τροχιάς που χρησιμοποιείται για την ανακάλυψη των επιλεγμένων κινητών και για την εκπαίδευση του κατηγοριοποιητή ονομάζεται σετ κατάρτισης τροχιάς (train set) και το σύνολο

δεδομένων που χρησιμοποιείται για την αξιολόγηση του κατηγοριοποιητή ονομάζεται σύνολο δοκιμής τροχιάς (test set).

3.2 Αλγοριθμική επεξήγηση της μεθόδου

Στην παρούσα ενότητα παρουσιάζεται η αλγοριθμική επεξήγηση της μεθόδου αυτής, καθώς είναι σημαντικό να παρουσιαστούν τα βήματα που γίνονται προκειμένου να μπορέσει ο αναγνώστης να κατανοήσει πως καταλήγει η μέθοδος στο επιθυμητό συμπέρασμα.

Ο αλγόριθμος της μεθόδου παρουσιάζεται στο Σχήμα 16 και περιγράφει λεπτομερώς τη μέθοδο MasterMovelets (Ferrero et al., 2020), η οποία έχει ως μοναδική είσοδο το σετ δεδομένων μιας τροχιάς *T*. Ο αλγόριθμος αυτός ξεκινά με την εξερεύνηση κάθε τροχιάς *T* στο σύνολο δεδομένων (γραμμές 2 έως 38 – Σχήμα 16). Η συνάρτηση Compute Element Distance Vectors (γραμμή 7 – Σχήμα 16) υπολογίζει την απόσταση μεταξύ όλων των στοιχείων μιας τροχιάς *T* και όλων των τροχιών στο **T**, και τα αποθηκεύει το αποτέλεσμα σε μια τετραδιάστατη συστοιχία, A_1 (γραμμή 4 – Σχήμα 16).

Το επόμενο βήμα συνίσταται στην εξερεύνηση όλων των μηκών των υπο-τροχιών, ένα προς ένα (γραμμές 5 έως 34 – Σχήμα 15). Για το μήκος της κάθε υπο-τροχιάς *w*, η συνάρτηση Compute Subtrajectory Distance Vectors (γραμμή 7 – Σχήμα 16) υπολογίζει την απόσταση μεταξύ όλων των υπο-τροχιών που ανήκουν στην τροχιά **T** προσθέτοντας απλά τις τιμές της αντίστοιχης συστοιχίας που προέκυψαν από το προηγούμενο βήμα χρησιμοποιώντας την ακόλουθη εξίσωση: $A_w - 1$. Το αποτέλεσμα αυτής της διαδικασίας αποθηκεύεται στη μεταβλητή A_w (γραμμή 7).

Στον βρόχο των γραμμών 9-33, για κάθε υπο-τροχιά μήκους w, ο αλγόριθμος χρησιμοποιεί το A_w για να ανακαλύψει τον καλύτερο συνδυασμό διαστάσεων. Σε αυτόν τον βρόχο, ο αλγόριθμος υπολογίζει πρώτα για κάθε υπο-τροχιά την απόσταση κατάταξης R μεταξύ όλων των υπο-τροχιών στη διάσταση k (γραμμές 10 έως 15).

```
   Input  : T // trajectory training set
   Output: movelets // set of relevant subtrajectories
 1 movelets ← ∅ ;
 2 for each trajectory T in T do
 3 |  candidates ← ∅;
 4 |  A_1 ← ComputeElementDistanceVectors (T, T) ;
 5 |  for subtrajectory length w from 1 to T.length do
 6 |  |  if w > 1 then
 7 |  |  |  A_w ← ComputeSubtrajectoryDistanceVectors (T, T, A_{w−1}, A_1, w) ;
 8 |  |  end
 9 |  |  for position j from 1 to (T.length − w + 1) do
10 |  |  |  R ← ∅ ;
11 |  |  |  for trajectory i from 1 to |T| do
12 |  |  |  |  for dimension d from 1 to |D| do
13 |  |  |  |  |  R[i, d, ..] ← Rank(A_w[i, j, d, ..]) ;
14 |  |  |  |  end
15 |  |  |  end
16 |  |  |  bestScore ← 0 ;
17 |  |  |  for each dimension combination C in C*_d do
18 |  |  |  |  W ← ∅;
19 |  |  |  |  for trajectory i from 1 to |T| do
20 |  |  |  |  |  W_i ← min MasterAlignment(R[i, C, ..], A_w[i, j, C, ..]) ;
21 |  |  |  |  |  W ← W ∪ (W_i, T[i].class) ;
22 |  |  |  |  end
23 |  |  |  |  relevance ← assess MasterRelevance(W, T.class) ;
24 |  |  |  |  if relevance.score > bestScore then
25 |  |  |  |  |  bestScore ← relevance.score ;
26 |  |  |  |  |  bestSp ← relevance.sp ;
27 |  |  |  |  |  bestW ← W ;
28 |  |  |  |  |  bestC ← C ;
29 |  |  |  |  end
30 |  |  |  end
31 |  |  |  M ← MoveletCandidate(T, j, w, bestC, bestW, bestScore, bestSp) ;
32 |  |  |  candidates ← candidates ∪ M ;
33 |  |  end
34 |  end
35 |  SortByRelevance (candidates) ;
36 |  RemoveSelfSimilar (candidates) ;
37 |  movelets ← movelets ∪ candidates ;
38 end
39 return movelets
```

Σχήμα 16 : Αλγόριθμος MasterMovelets.

Πηγή : Ferrero et al., 2020.

Η μέθοδος αυτή έχει δύο υπο-μεθόδους – συναρτήσεις, τις MasterAlignment και MasterRelevance. Οι συναρτήσεις αυτές ενεργοποιούνται μόλις υπολογιστεί το *R* προκειμένου να διερευνηθεί κάθε συνδυασμός διαστάσεων *C* (γραμμές 17 έως 30 – Σχήμα 16). Σε αυτόν τον βρόχο, βρίσκει το διάνυσμα απόστασης της καλύτερης ευθυγράμμισης μεταξύ κάθε υπο-τροχιάς, χρησιμοποιώντας μια συγκεκριμένη μέθοδο για πολυδιάστατη ευθυγράμμιση μιας υπο-τροχιάς, που ονομάζεται MasterAlignment. Η μέθοδος αυτή, που θα αναλυθεί εκτενέστερα στην ενότητα 3.2, παράγει ως έξοδο και αποθηκεύει το διάνυσμα απόστασης (γραμμές 18 έως 22). Μετά τον υπολογισμό του διανύσματος απόστασης, ο αλγόριθμος μετρά τη συνάφεια κάθε υπο-τροχιάς με βάση την τιμή αυτή χρησιμοποιώντας μια συγκεκριμένη συνάρτηση που ονομάζεται MasterRelevance.

Όπως σημειώνεται από τον Ferrero (2020), βρίσκοντας τον συνδυασμό διαστάσεων με την υψηλότερη βαθμολογία συνάφειας, ο αλγόριθμος διατηρεί επίσης τα σημεία διαχωρισμού, τα διανύσματα απόστασης και τον συνδυασμό διαστάσεων (γραμμές 23 έως 29 – Σχήμα 16). Στη συνέχεια, ορίζει την υποψήφια υπο-τροχιά ως την υπο-τροχιά με τον πιο σχετικό συνδυασμό διαστάσεων και την αποθηκεύει στα υποψήφια σύνολα (γραμμές 31 και 32 – Σχήμα 16). Ακολουθώντας τον εξωτερικό βρόχο, ταξινομεί τις υποψήφιες τροχιές ανάλογα με τη συνάφειά τους (γραμμές 35 έως 36 – Σχήμα 16), χρησιμοποιώντας την μέτρηση που παρουσιάζεται στην ενότητα 3.3.

3.3 Πολυδιάστατη ευθυγράμμιση μιας υποτροπής σε τροχιά (MasterAlignment)

Όπως σημειώνεται από τον Ferrero (2020), η συνάρτηση MasterAlignment χρησιμοποιείται για τη πολυδιάστατη ευθυγράμμιση μιας υπο-τροχιάς και έχει ως

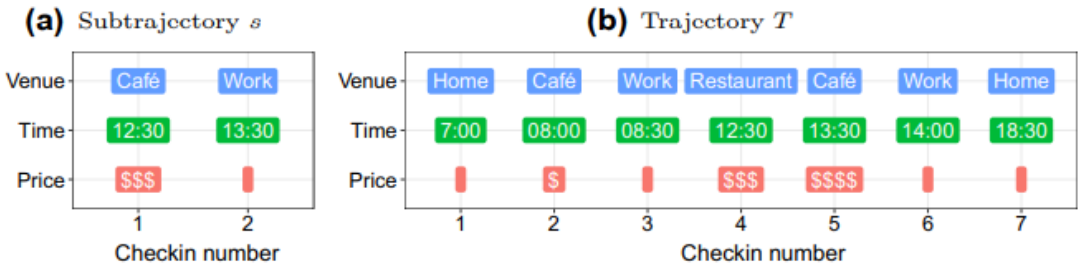
είσοδο δύο παραμέτρους, τις *V* και *R*. Οι παράμετροι αυτοί σχετίζονται με τις τιμές της απόστασης και την κατάταξη της απόστασης, αντιστοίχως.

Στην αρχή του αλγορίθμου αρχικοποιείται το σύνολο μέσης κατάταξης *Y*, ο αριθμός διαστάσεων, ο οποίος όπως παρουσιάστηκε είναι σημαντικός για την παραγωγή της βαθμολογίας συνάφειας, και η αρχική θέση της ελάχιστης μέσης κατάταξης, η οποία ταυτίζεται με 1. Έπειτα, υπολογίζεται η μέση κατάταξη για την κάθε διάσταση. Στην επαναληπτική διαδικασία αυτή προστίθενται οι τιμές κατάταξης κατά μήκος των διαστάσεων και λαμβάνεται η μέση τιμή τους, η οποία διατηρείται στην αντίστοιχη θέση του συνόλου *Y*.

Αφού υπολογίζεται η μέση κατάταξη, συγκρίνεται με την τρέχουσα μέση κατάταξη, και υπολογίζεται η ελάχιστη μέση κατάταξη. Εάν η τρέχουσα μέση κατάταξη είναι χαμηλότερη, αλλάζει η θέση της ελάχιστης μέσης κατάταξης, αλλιώς παραμένει ως έχει. Επιπλέον, ο αλγόριθμος ολοκληρώνεται με την απόκριση του διανύσματος απόστασης της καλύτερης ευθυγράμμισης βάσει αυτής της θέσης. Μάλιστα, αυτό είναι και το διάνυσμα απόστασης μεταξύ της υπο-τροχιάς και της τροχιάς (Ferrero et al., 2020).

Για να γίνει πιο κατανοητός ο τρόπος με τον οποίο λειτουργεί η μέθοδος για να πετύχει τον σκοπό της δίνεται το ακόλουθο παράδειγμα (Ferrero, 2020). Ενώ η συνήθης λύση που δίνεται σε παρόμοιες περιπτώσεις είναι η μετατροπή κάθε διανύσματος απόστασης σε μία τιμή απόστασης εφαρμόζοντας μια συνάρτηση, αυτή η λύση φέρνει δύο σημαντικά μειονεκτήματα. Το πρώτο είναι ο σχεδιασμός μιας συνάρτησης μετασχηματισμού ώστε να ενσωματώνει τις αποστάσεις, οι οποίες εξαρτώνται από το πεδίο και η δεύτερη είναι η απώλεια πληροφορίας σχετικά με την απόσταση σε κάθε διάσταση. Για να γίνει πιο κατανοητό αυτό το σενάριο, το ακόλουθο σχήμα (α) δείχνει

ένα παράδειγμα μιας υπο-τροχιάς και (β) μια ολοκληρωμένη τροχιά *T*. Η υπο-τροχιά που πρέπει να ευθυγραμμιστεί είναι η: «Οι χρήστες που επισκέπτονται ένα Café με τιμή $$$ περίπου στις 12:30 π.μ. και μετά πηγαίνουν στη δουλειά περίπου στις 13:30 μ.μ.». (Ferrero et al., 2020)



Σχήμα 17 : Παράδειγμα μιας υπο-τροχιάς και μιας τροχιάς *T*.

Πηγή : Ferrero et al., 2020.

Σχήμα 18 : Η καλύτερη ευθυγράμμιση υπο-τροχιάς επισημαίνεται στην τροχιά

Πηγή : Ferrero et al., 2020.

Όπως σημειώνεται από τον Ferrero (2020), και παρουσιάζεται στο Σχήμα 17 ο χρήστης της τροχιάς *T* δεν εκτελεί την ακριβή ακολουθία συνεχώς, λαμβάνοντας υπόψη όλες τις διαστάσεις. Για παράδειγμα, ο χρήστης μπορεί να πάει στο χώρο του καφέ και μετά στη δουλειά, ξεκινώντας από τη θέση 2 (Σχήμα 17) και μετά να πάει στην θέση 5 (Σχήμα 17). Επιπλέον, στην αρχική θέση 4 (Σχήμα 17) ο χρήστης της τροχιάς *T* πραγματοποιεί check-in στις 12:30 μ.μ. Σύμφωνα με τον Ferrero (2020), η καλύτερη ευθυγράμμιση αντιπροσωπεύεται από την ακολουθία των check-in

ξεκινώντας από τη θέση 5, επειδή εκτός από την ακολουθία των χώρων που είναι η ίδια (Café, Work), οι διαστάσεις Time και Price είναι επίσης πολύ παρόμοιες χρονικά.

Με βάση το παράδειγμα που έχει δοθεί στα Σχήματα 17 και 18, δίνεται η ανάλυση του με βάση τη μέθοδο αυτή. Ο πίνακας της ακόλουθης εικόνας παρουσιάζει τις τιμές των διανυσμάτων απόστασης του παραδείγματος τα οποία εξάγονται σαν αποτελέσματα. Η στήλη στην αρχική θέση 1, δείχνει οτι το διάνυσμα απόστασης V_1 μεταξύ της υπο-τροχιάς s = (Café, 12:30, $$$), (Work, 13:30, ∅) και της ακολουθίας του T στην αρχική θέση 1, αντιπροσωπεύεται από r_1 = (Home, 07:00, ∅), (Café, 08: 00, $). Για τη διάσταση Venue, η απόσταση είναι 2, επειδή τα Café, Work διαφέρουν από την ακολουθία Home, Café και στα δύο check-in (θέσεις 2 και 4 όπως αναφέρθηκε στην ενότητα 3.1). Στη διάσταση Time, η απόσταση είναι 660, επειδή το άθροισμα των διαφορών χρόνου σε λεπτά είναι: | (12: 30-07: 00) | = 330 λεπτά και | (13: 30-08: 00) | = 330 λεπτά, συνολικά 660. Στη διάσταση τιμής, η απόσταση είναι 4 μονάδες τιμών, επειδή το άθροισμα της διαφοράς μεταξύ των τιμών είναι: | ($$$ - ∅) | = 3 μονάδες τιμών και | (∅ - $) | = 1 μονάδα τιμής, δηλαδή συνολικά 4 μονάδες τιμών. Η μέθοδος εκτελεί τον ίδιο υπολογισμό απόστασης για τις επόμενες θέσεις εκκίνησης.

(a) Distance values

Distance	Starting position					
	1	2	3	4	5	6
Venue	2	0	2	2	0	2
Time	660	570	300	0	90	390
Price	4	2	6	4	1	3
Vector	$\mathbf{V}_1$	$\mathbf{V}_2$	$\mathbf{V}_3$	$\mathbf{V}_4$	$\mathbf{V}_5$	$\mathbf{V}_6$

(b) Distance rankings

Ranking	Starting position					
	1	2	3	4	5	6
Venue	4.5	1.5	4.5	4.5	1.5	4.5
Time	6	5	3	1	2	4
Price	4.5	2	6	4.5	1	3
Avg. rank	5.0	2.8	4.5	3.3	1.5	3.8

Σχήμα 19 : Εύρεση των καλύτερων ευθυγραμμίσεων από τα διανύσματα απόστασης.

Πηγή : Ferrero et al., 2020.

Μετά από την παραγωγή του παραπάνω πίνακα, κατατάσσονται οι τιμές απόστασης για κάθε διάσταση, τα αποτελέσματα του οποίου φαίνονται στο Σχήμα 19 (b) και δείχνει τις τιμές κατάταξης. Για παράδειγμα, για τη διάσταση ώρας (Time) οι τιμές απόστασης είναι (660, 570, 300, 0, 90, 390) και οι τιμές κατάταξης είναι (6, 5, 3, 1, 2, 4). Αυτή η κατάταξη υποδεικνύει ότι η απόσταση 660 έχει τη χειρότερη τιμή κατάταξης, 6 και η απόσταση 0 έχει την καλύτερη τιμή κατάταξης, 1. Σημειώνεται ότι για τη διάσταση Venue η μέθοδος υποστηρίζει επίσης κλασματικές τάξεις σε περίπτωση ισοπαλίας, όπως 1,5 σε αρχικές θέσεις 2 και 5. (Ferrero et al., 2020)

Στη συνέχεια, η μέθοδος υπολογίζει τη μέση κατάταξη σε κάθε αρχική θέση (τελευταία σειρά στο Σχήμα 19β), με αποτέλεσμα (5.0, 2.8, 4.5, 3.3, 1.5, 3.8). Έτσι, το MasterAlignment ((Ferrero et al., 2020)) θεωρεί την καλύτερη ευθυγράμμιση ως τη χαμηλότερη μέση κατάταξη, δηλαδή 1,5 (υπογραμμίζεται στον στο Σχήμα 19β) και αντιστοιχεί στην 5η αρχική θέση. Τέλος, η μέθοδος επιστρέφει το διάνυσμα απόστασης $v5$ = (0, 90, 1), που αντιπροσωπεύει τις αποστάσεις (της καλύτερης ευθυγράμμισης) μεταξύ της υπο-τροχιάς s και της τροχιάς *T*.

3.4 Μέτρηση συνάφειας για πολυδιάστατους υποψηφίους υπο-τροχιάς (MasterRelevance)

Η μέθοδος MasterMovelets (Ferrero et al., 2020) ασχολείται με το πρόβλημα των αποστάσεων σε πολλές διαστάσεις και βρίσκει τα σημεία διαχωρισμού που μεγιστοποιούν τη συνάφεια της υποψήφιας υπο-τροχιάς. Όπως παρουσιάστηκε στην ενότητα 3.1, υπάρχουν περιπτώσεις όπου υπάρχουν πολυδιάστατοι υποψήφιοι υπο-τροχιών και αλληλοεπικαλύψεις. Δύο υποψήφιοι είναι ίδιοι εάν αλληλεπικαλύπτονται

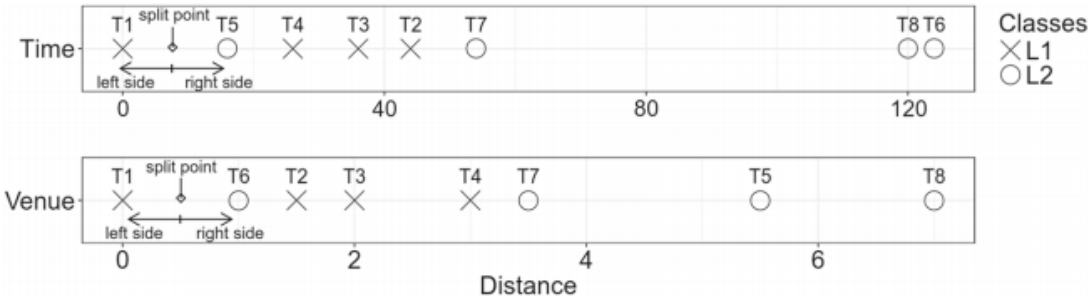
σε τουλάχιστον ένα στοιχείο τροχιάς και ο αλγόριθμος διατηρεί τον υψηλότερο σχετικό υποψήφιο. Δύο βασικά σημεία για την ανακάλυψη κινητών σε πολλαπλές πτυχές είναι: η εύρεση της καλύτερης ευθυγράμμισης της υπο-τροχιάς που εκτελείται με τη μέθοδο MasterAlignment και η μέτρηση της συνάφειας της υπο-τροχιάς, που εκτελείται με τη μέθοδο MasterRelevance. (Ferrero et al., 2020)

Όπως σημειώνεται από τον Ferrero (2020), η συνάφεια μιας υπο-τροχιάς σχετίζεται με τον αριθμό των τροχιών της ίδιας κλάσης που εκτελούν παρόμοια κίνηση. Αναλύονται οι αποστάσεις της βέλτιστης ευθυγράμμισης μεταξύ μιας υπο-τροχιάς και όλων των τροχιών στο σύνολο δεδομένων προκειμένου να προσδιοριστεί το ποιες τροχιές της ίδιας κλάσης εκτελούν παρόμοια κίνηση. Η πιο κοινή προσέγγιση συνίσταται στην αναπαράσταση των καλύτερων αποστάσεων σε μια σειρά και την εύρεση ενός σημείου διαχωρισμού για το διαχωρισμό των αποστάσεων σε δύο ομάδες (Σχήμα 20): (α) την πλησιέστερη (αριστερή πλευρά) και (β) την πιο απομακρυσμένη (δεξιά πλευρά), όπου η πλησιέστερη εκτελεί παρόμοια κίνηση και η πιο απομακρυσμένη όχι.

Για το σκοπό αυτό έχουν προταθεί αρκετές τεχνικές για την εύρεση του διαχωριστικού σημείου, όπως το μέγιστο κέρδος πληροφοριών (Ye and Keogh 2011), το Kruskal-Wallis και το Mood's Median (Lines and Bagnall 2012) και το Left Side Pure (LSP) (Ferrero et al . 2018). Ωστόσο, αυτές οι τεχνικές περιορίζονται στην εύρεση του διαχωριστικού σημείου σε μια διαστατική σειρά.

Προκειμένου να γίνει κατανοητός ο τρόπος μέτρησης της συνάφειας, χρησιμοποιείται το παράδειγμα των σχημάτων 20 και 21. Σύμφωνα με το παράδειγμα τα T_1, T_2, ..., T_8 των κατηγοριών L_1 και L_2, που αντιπροσωπεύονται από τις διαστάσεις Time και Venue, όπου τα T_1, T_2, T_3 και T_4 είναι των κατηγοριών L_1 και T_5, T_6, T_7 και

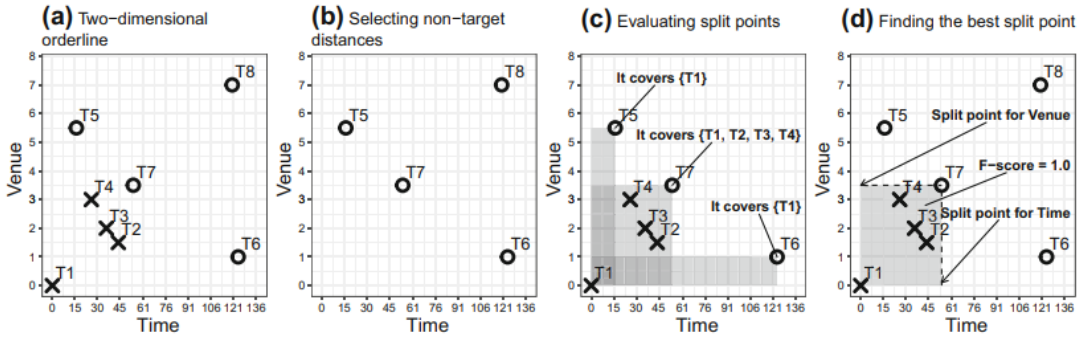
το T_8 ανήκει στην κατηγορία L_2 και ένα υποψήφιο κινητό (movelet candidate) M εξάγεται από τον T_1 της κατηγορίας L_1. (Ferrero et al., 2020)



Σχήμα 20 : Γραμμές παραγγελίας για χρόνο και χώρο διαστάσεων.

Πηγή : Ferrero et al., 2020.

Το Σχήμα 20 δείχνει τις γραμμές παραγγελίας, όπου κάθε σημείο δείχνει την τιμή απόστασης προς την τροχιά i^{th} για κάθε διάσταση και τα σύμβολα X και O είναι οι κλάσεις. Η τροχιά T_1 είναι η πρώτη τιμή απόστασης σε κάθε διάσταση επειδή το υποψήφιο (movelet candidate) M προέρχεται από το T_1.



Σχήμα 21 : Παράδειγμα εύρεσης διαχωριστικών σημείων σε μια πολυδιάστατη σειρά παραγγελιών.

Πηγή : Ferrero et al., 2020.

Το Σχήμα 21α δείχνει τις τιμές απόστασης των υπο-τροχιών που παρουσιάζονται στο Σχ. 20, σε ένα διάγραμμα διασποράς, όπου κάθε σημείο δείχνει τις αποστάσεις και στις δύο διαστάσεις, Χρόνος και Χώρος, στην τετμημένη και τεταγμένη, αντίστοιχα. Το πρώτο βήμα συνίσταται στην επιλογή μόνο των σημείων της κατηγορίας L_2 (που είναι κλάση στόχος) και στη συνέχεια στο κλάδεμα των σημείων με μεγαλύτερες τιμές

απόστασης από κάποιες άλλες και στις δύο διαστάσεις Time and Venue, που ονομάζονται καλυμμένα σημεία. (Ferrero et al., 2020)

Στο Σχ. 21b το σημείο του T_8 κλαδεύεται επειδή έχει μεγαλύτερες τιμές απόστασης από το T_7 και στις δύο διαστάσεις. Στο δεύτερο βήμα, η μέθοδος MasterRelevance αξιολογεί τα μη κομμένα σημεία ανάλογα με τον αριθμό των σημείων κάθε κλάσης που καλύπτονται, λαμβάνοντας υπόψη και τις δύο διαστάσεις. Το Σχήμα 21c δείχνει ότι χρησιμοποιώντας τις τιμές σημείου των T_5 ή T_6 ως σημεία διαχωρισμού καλύπτουν μόνο το T_1 της κλάσης L_1, αλλά χρησιμοποιώντας τις τιμές σημείου του T_7 καλύπτουν τα T_1, T_2, T_3 και T_4. Στο τελευταίο βήμα επιλέγει τους πόντους διαίρεσης που έχουν την υψηλότερη βαθμολογία συνάφειας. Για να υπολογιστεί η βαθμολογία συνάφειας χρησιμοποιείται το F-μέτρο, που είναι ο αρμονικός μέσος όρος της ακρίβειας και της ανάκλησης.

Σε αυτό το πλαίσιο, η ακρίβεια είναι η αναλογία των σημείων που καλύπτονται από τα σημεία διαίρεσης που ανήκουν στην κατηγορία στόχου και η ανάκληση είναι η αναλογία των σημείων που καλύπτονται από τα σημεία διαίρεσης που ανήκουν στην κατηγορία στόχου σε σχέση με όλα τα σημεία της κατηγορίας στόχου. Όπως φαίνεται στο Σχ. 21d, τα καλύτερα σημεία διαίρεσης είναι οι τιμές που ορίζονται από το σημείο $T7$ και το σκορ είναι 1.

4. Πειραματικό μέρος

4.1 Περιγραφή των δεδομένων

Στην προσπάθεια να αναλυθούν πλήρως οι θεωρητικές βάσεις που χτίστηκαν στο μεγαλύτερο μέρος της βιβλιογραφικής επισκόπησης, ακολουθεί η περιγραφή των δεδομένων που θα χρησιμοποιηθούν. Όπως έγινε αντιληπτό και στις προηγούμενες ενότητες, οι μέθοδοι που επικαλούνται, αντικατοπτρίζονται στην κατηγοριοποίηση των παρατηρήσεων με βάση τις σημάνσεις τους, δηλαδή χαρακτηριστικά που δεν αφορούν απλώς τις κινήσεις τους, αλλά αποτελούν μια επιπλέον πληροφορία που μπορεί να συνδράμει στην ανάδειξη ενός χαρακτηριστικού τους.

Με σκοπό την καλύτερη δυνατή συνοχή στην εμπειρική ανάλυση της μελέτης μας, η βάση δεδομένων που διαδραματίζει βασικό ρόλο στις τεχνικές μας, αποτελεί κομμάτι του αστικού περιβάλλοντος.

Με την έννοια του αστικού περιβάλλοντος, εννοείται ένα σύνολο που αντικατοπτρίζει τις κινήσεις ανθρώπων σε εξωτερικούς χώρους. Οι ενέργειες αυτές μπορούν να εξαχθούν με την επισήμανση του τόπου που βρίσκονται με τη βοήθεια κοινωνικών μέσων δικτύωσης («check – in») ή μέσω GPS κατασκευασμένων με τέτοιον τρόπο ώστε να μη συλλέγουν απλώς γεωγραφικά δεδομένα. Οι πληροφορίες που μπορεί κανείς να συλλέξει με αυτό τον τρόπο, πέρα από τις γεωγραφικές συντεταγμένες του τόπου επισήμανσης, είναι αρκετές. Για παράδειγμα, μπορούν να βρεθούν συσχετίσεις μεταξύ των ατόμων που αναφέρουν πως τους αρέσει η δημοσίευση αυτή («like»).

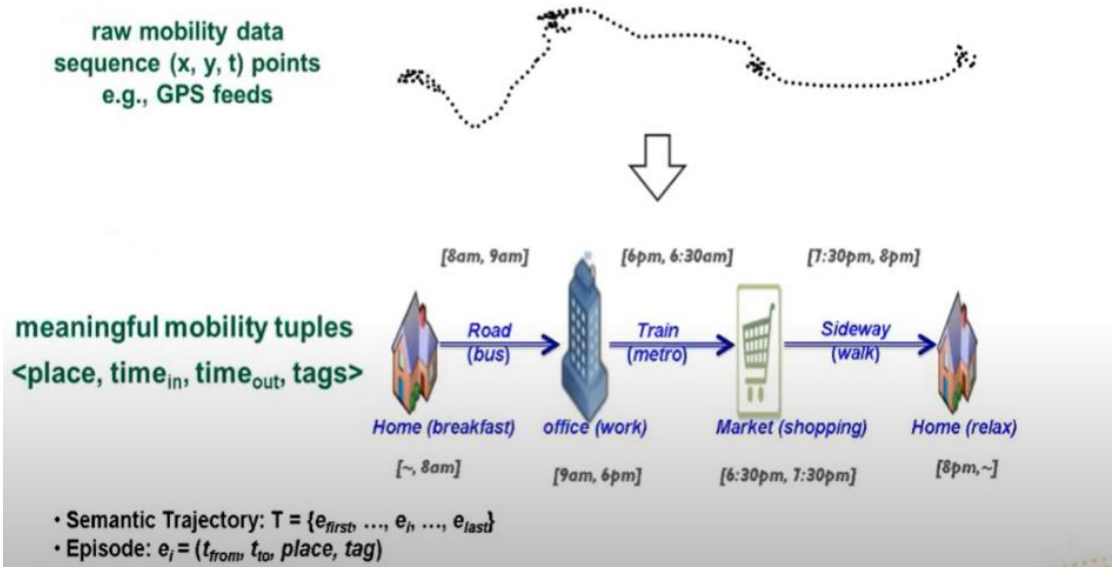
Αυτού του είδους η διασύνδεση μπορεί να θεωρήσει κανείς ότι είναι και τελείως τυχαία, αλλά σίγουρα αποτελεί έναν έμμεσο τρόπο διαφήμισης του υποκείμενου τόπου ή τοποθεσίας που επισκέπτεται το άτομο που προέβη σε αυτή τη δημοσίευση μέσω του

προσωπικού του λογαριασμού στο αντίστοιχο ηλεκτρονικό μέσο κοινωνικής δικτύωσης. Τουτέστιν, ο έλεγχος μιας υποτιθέμενης κατηγοριοποίησης θα μπορούσε να συνεισφέρει σε έναν τύπο βαθμολόγησης του τόπου ή να προσδώσει ταμπέλες («labels») σε άτομα με παρόμοιες προτιμήσεις ότι είναι πιθανόν να επισκεφτούν αυτό τον τόπο.

Στόχος στην παρούσα εργασία είναι να χρησιμοποιηθούν μέθοδοι κυρίως επιβλεπόμενης μάθησης με σκοπό την ανάλυση του περιβάλλοντος που αναφέρθηκε στις προηγούμενες παραγράφους.

Κατά την τελευταία δεκαετία, ο τομέας της εξόρυξης δεδομένων κίνησης εμφανίστηκε παρέχοντας πολλές αποτελεσματικές μεθόδους για την ανακάλυψη διαισθητικών προτύπων που αντιπροσωπεύουν τη συλλογική συμπεριφορά τροχιών κινούμενων αντικειμένων. Τα μεγάλα σύνολα κινούμενων δεδομένων συνήθως δεν είναι διαθέσιμα, αλλά και όταν είναι, οι βασικές πληροφορίες αλήθειας «ground truth information» τέτοιων μοτίβων εκλείπουν. Παρόλο που πρόσφατα έγιναν διαθέσιμα μερικά σύνολα δεδομένων πραγματικής τροχιάς, αυτά δεν επαρκούν για την πειραματική αξιολόγηση των διαφόρων προτάσεων, επομένως, οι ερευνητές προτιμούν τις συνθετικές γεννήτριες τροχιάς. Όμως, οι γεννήτριες συνθετικών δεδομένων που χρησιμοποιούνται δεν εγγυώνται την ύπαρξη τέτοιων μοτίβων κινητικότητας, καθώς αδυνατούν να προσομοιώσουν τις σημασιολογικές πληροφορίες μίας κίνησης. Αυτή η περίπτωση είναι προβληματική επειδή, αφενός, τα πραγματικά σύνολα δεδομένων είναι συνήθως μικρά, γεγονός που δυσκολεύει τα πειράματα κλιμάκωσης και, από την άλλη πλευρά, οι συνθετικές γεννήτριες δεδομένων δεν έχουν σχεδιαστεί για να παράγουν τροχιές με μοτίβο κινητικότητας. Παρακινούμενος από αυτήν την παρατήρηση, παρουσιάζεται το σύνολο δεδομένων «Hermoupolis», μια αποτελεσματική γεννήτρια συνθετικών τροχιών κινούμενων αντικειμένων που έχει ως

κύριο στόχο τα σύνολα δεδομένων που προκύπτουν να υποστηρίζουν διάφορους τύπους κινητικότητας (συμπλέγματα, σμήνη, συνοδείες κ.λπ.). Δημιουργούν λοιπόν σύνολα δεδομένων με βασικές πληροφορίες αλήθειας («ground truth information»). Το σύνολο δεδομένων «Hermoupolis» αποτελεί μια επέκταση της γνωστής Brinkhoff γεννήτριας τροχιών.



Σχήμα 22 : Διαφορά απλής τροχιάς με σημασιολογικής τροχίας.

Πηγή : http://infolab.cs.unipi.gr/pubs/pkdd2013/.

Στο σύνολο δεδομένων μπορεί κανείς να διακρίνει στοιχεία που αποτελούν κινήσεις σε αστικό περιβάλλον. Όμως η παραγωγή των συγκεκριμένων παρατηρήσεων έχει διαδραματιστεί μέσα από έναν αλγόριθμο που παράγει συνθετικά δεδομένα, ανάλογα με το αν το συγκεκριμένο πρόσωπο μπορεί να κινηθεί με τα πόδια ή με κάποιο διαφορετικό τρόπο μετακίνησης (λεωφορείο, μετρό, ποδήλατο, αυτοκίνητο κλπ). Πρόκειται για εξομοίωση κίνησης ενός σεναρίου μιας ημέρας όπου τα 100 αντικείμενα

ακολουθούν 4 διαφορετικές κλάσεις κίνησης (πρότυπα κίνησης) στην περιοχή της Αθήνας.

Στο παρακάτω πίνακα παρουσιάζονται οι μεταβλητές που περιέχονται στο αρχικό σύνολο δεδομένων, το οποίο αρχικά δεν έχει υποστεί καμία επεξεργασία.

Πίνακας 1. Μεταβλητές που περιέχονται στο αρχικό σύνολο δεδομένων

ΣΤΗΛΕΣ	ΠΕΡΙΓΡΑΦΗ
scenarioID	Το αναγνωριστικό του σεναρίου
MOid	Το αναγνωριστικό του κινούμενου αντικειμένου
MPid	Το αναγνωριστικό της κλάσης όπου ανήκει το κινούμενο αντικείμενο
edgeID	το αναγνωριστικό της ακμής του δικτύου όπου κινείται το αντικείμενο
realX realY	Οι συντεταγμένες όπου βρίσκεται κάθε στιγμή το αντικείμενο σε σύστημα προβολής 2100 (SRID)
realTime	Ο απόλυτος χρόνος της εξομοίωσης
episodesems	Το αναγνωριστικό του τρέχοντος επεισοδίου κίνησης όπου βρίσκεται το αντικείμενο : ο χρόνος σε μορφή έτος-μήνα-ημέρα ώρα σε 24η:λεπτά:δεύτερα ,το είδος του επεισοδίου (STOP / MOVE), το μέρος αν είναι σε STOP ή το μέσο μετακίνησης αν είναι σε MOVE, η ενέργεια που κάνει το αντικείμενο (π.χ RELAXING)

Στο σύνολο δεδομένων Hermoupolis μπορεί κανείς να διακρίνει στοιχεία που αποτελούν κινήσεις σε αστικό περιβάλλον, όπως σχολιάστηκε στις ανωτέρω παραγράφους. Αξίζει να αναφερθεί άλλη μια φορά ότι η παραγωγή των συγκεκριμένων παρατηρήσεων έχει διαδραματιστεί μέσα από έναν αλγόριθμο. Ο τελευταίος, λοιπόν, παράγει «φτιαχτά» δεδομένα ανάλογα με το αν το συγκεκριμένο πρόσωπο μπορεί να κινηθεί με τα πόδια ή με κάποιο διαφορετικό τρόπο μεταφοράς.

Πιο συγκεκριμένα, στο συγκεκριμένο σετ υπάρχουν 420.511 παρατηρήσεις, για τις οποίες θα προβούμε σε αναλύσεις σε σχέση με κάποιες ιδιότητες τους για τις οποίες λαμβάνονται πληροφορίες μέσα από τα χαρακτηριστικά που έχει προσφέρει ο αλγόριθμος.

4.2 Μεθοδολογία

4.2.1 Επεξεργασία των Δεδομένων

Η επεξεργασία του συνόλου των δεδομένων πραγματοποιήθηκε μέσω της γλώσσας R. Πρόκειται για μία γλώσσα προγραμματισμού και ένα περιβάλλον για στατιστικούς υπολογισμούς και γραφικές παραστάσεις. Αποτελεί ένα έργο GNU παρόμοιο με τη γλώσσα S και δύναται να θεωρηθεί μια διαφορετική υλοποίηση της γλώσσας S. Η γλώσσα R διατίθεται ως ελεύθερο λογισμικό και αποτελεί μια ολοκληρωμένη «σουίτα» εγκαταστάσεων λογισμικού για χειρισμό δεδομένων («data manipulation»), υπολογισμό και απεικόνιση γραφημάτων.

Θα βασιστούμε στον κώδικα που παρέχεται δημόσια στο αποθετήριο Github: https://github.com/anfer86/dmkd_masterMovelets_results για να εκτελέσουμε την μέθοδο MasterMovelets (Ferrero et al., 2020) στο δικό μας σύνολο δεδομένων Hermoupolis. Αυτή η μέθοδος απαιτεί διάφορα διακριτά βήματα για να εκτελεστεί, τα οποία παρουσιάζονται παρακάτω αναλυτικά.

Καταρχάς πραγματοποιείται η εγκατάσταση των ακόλουθων πακέτων των οποίων η γενική χρήση τους αναφέρεται στον κάτωθι πίνακα:

Πίνακας 2. Πακέτα που χρησιμοποιήθηκαν στην προ-επεξεργασία των δεδομένων

ΠΑΚΕΤΑ	ΧΡΗΣΗ
caTools	Περιέχει βασικές συναρτήσεις
tidyr	Σκοπός του είναι να δημιουργήσει οργανωμένα δεδομένα
dplyr	Χρησιμεύει στη χειραγώγηση δεδομένων (π.x. mutate , select , filter, summarise)
lubridate	Χρήσιμο πακέτο που βοηθά σε ζητήματα αναφορικά με ημερομηνίες και χρόνο
data.table	Χρήσιμο για γρήγορη συγκέντρωση μεγάλων δεδομένων, γρήγορη προσθήκη/τροποποίηση/διαγραφή στηλών κλπ.

Απαιτείται για να «τρέξει» ο αλγόριθμος MasterMovelets (Ferrero et al., 2020) να μετασχηματιστούν τα δεδομένα του συνόλου Hermoupolis κατ' εικόνα και καθ' ομοίωσίν με τα αντίστοιχα του Ferrero, (2020).

Αρχικά τρέχουμε το R script split_dataset.R, το οποίο λαμβάνει τα δεδομένα και τα χωρίζει σε υποσύνολα train και test. Σημειώνεται ότι χρησιμοποιήθηκε R έκδοση 4.1.1. Ο χωρισμός των δεδομένων είναι σημαντικός για την αξιολόγηση, καθώς μας ενδιαφέρει η απόδοση σε νέα δεδομένα που δεν έχει «δει» ο αλγόριθμος προηγουμένως κατά την εκπαίδευση. Εδώ θέσαμε αναλογία 80% train και 20% test (5-Fold Validation). Πιο συγκεκριμένα, σε αυτό το βήμα δημιουργούνται οι φάκελοι train και test με αρχεία κειμένου με κατάληξη .r2. Τέλος, δημιουργούμε ένα zip αρχείο για κάθε φάκελο με όλα τα αρχεία που περιέχει. Επίσης, έχουμε δημιουργήσει ένα αρχείο json

για την περιγραφή των δεδομένων σύμφωνα με τα δεδομένα που δίνονται ως υπόδειγμα.

Συνεπώς, από το σύνολο των δεδομένων πραγματοποιείται ο διαχωρισμός της στήλης «episodesems» - η οποία περιέχει τις σημασιολογικές πληροφορίες που θα χρησιμοποιηθούν - στις επιμέρους στήλες: «episodeID», «timeInt» , «type», «action», «activity». Ακόμα, πραγματοποιείται η συνένωση των στηλών «realX»,«realY» στην νέα στήλη «realCoords», δηλαδή των συντεταγμένων , έτσι ώστε να έρθουν στην ίδια μορφή με την αντίστοιχη στήλη («Space») (Ferrero et al., 2020) στα αντίστοιχα σύνολα δεδομένων (Gowalla, Brightkite).

Πίνακας 3. Περιγραφή διάστασης τροχιάς του Hermoupolis

ΜΕΤΑΒΛΗΤΕΣ	ΔΙΑΣΤΑΣΗ	ΤΥΠΟΣ ΜΕΤΑΒΛΗΤΩΝ
realCoords	Space	Σύνθετη (realX realY)
timeInt	Time	Χρονική
type	Nominal	Ονομαστική
action	Nominal	Ονομαστική
activity	Nominal	Ονομαστική

Έπειτα το νέο σύνολο δεδομένων, το οποίο πλέον αποτελείται από τις στήλες «realCoords», «time» , «type», «action», «activity», χωρίζεται σε test και train set. Όπως έχει αναφερθεί και σε προηγούμενη ενότητα, ως train set ορίζεται το σετ των δεδομένων των οποίων οι κλάσεις είναι γνωστές από τον αλγόριθμο κατηγοριοποίησης, ο οποίος χρησιμοποιεί το test set για την πρόκληση ενός νέου μοντέλου κατηγοριοποίησης. Ως test set ορίζεται ως το σετ των δεδομένων των οποίων η κλάση δεν είναι γνωστή και την οποία ο αλγόριθμος κατηγοριοποίησης καλείται να αναγνωρίσει. Το 80% του συνόλου των δεδομένων χωρίζεται τυχαία σε train set και το υπόλοιπο 20% σε test set. Στη στατιστική γενικά όσο περισσότερα δεδομένα χρησιμοποιούνται, τόσο πιο ακριβείς είναι και οι εκτιμήσεις που προκύπτουν. Από

αυτήν την άποψη μπορεί να φαίνεται ότι ο καλύτερος τρόπος για να επιτευχθεί η καλύτερη εκτίμηση είναι η χρησιμοποίηση όλων των διαθέσιμων δεδομένων. Αυτή θεωρείται καλή ιδέα αν και μόνο αν είμαστε απόλυτα σίγουροι ότι το μοντέλο περιγράφει ακριβώς το αντίστοιχο φαινόμενο. Τότε είναι που υπερφορτώνεται το μοντέλο (overfitting) και θεωρείται ότι το μοντέλο περιγράφει όλα τα δεδομένα τέλεια χωρίς όμως στην πραγματικότητα αυτό να ισχύει. Για να αποφευχθεί η υπερφόρτωση του μοντέλου συνίσταται λοιπόν ο διαχωρισμός του συνόλου των δεδομένων σε test & train set. Αρχικά χρησιμοποιούνται τα δεδομένα εκπαίδευσης για να καθοριστούν οι παράμετροι του μοντέλου και αφετέρου γίνεται σύγκριση των προβλέψεων του μοντέλου για όλα τα δεδομένα δοκιμών («test set») με αυτά που παρατηρήθηκαν και χρησιμοποιείται αυτή η σύγκριση για να μετρηθεί η ακρίβεια του μοντέλου.

4.2.2 Πειραματικό Μέρος

Το επόμενο βήμα είναι η εκτέλεση του κώδικα MasterMovelets (Ferrero et al., 2020), ο οποίος δίνεται ως βιβλιοθήκη jar της γλώσσας προγραμματισμού Java. Μιας και εργαζόμαστε σε Windows, με βάση το bash αρχείο που δίνεται, δημιουργούμε την παρακάτω εντολή, την οποία την εκτελούμε στο φάκελο που βρίσκεται το αρχείο jar:

```
java -Xmx220g -Xms1g -jar MasterMoveletsByClass.jar -curpath
../datasets/hermoupolis/ -respath ../results/hermoupolis_final
-descfile ../datasets/descriptions/hermoupolis.json -nt 4 -
cache true -ms 1 -Ms -1 -ed true -mnf -1 -samples 1 -sampleSize
0.5 -medium "none" -output "discrete" -lowm "false" >
"../results/output_master_hermoupolis_final.txt"
```

Το πρόγραμμα αυτό δημιουργεί δύο αρχεία σε μορφή csv, train και test, για κάθε κλάση ξεχωριστά, τα οποία περιέχουν τα movelets. Μια καταγραφή των ενεργειών του προγράμματος φαίνεται στο αρχείο output_master_hermoupolis_mnf_-1__Ms_-1.txt.

Στη συνέχεια, με το R script MergeDatasets.R ενώνουμε τα αρχεία κάθε κλάσης, ώστε να έχουμε ένα αρχείο train.csv και test.csv.

Τέλος, εκτελούμε κατηγοριοποίηση με τυχαίο δάσος (Random Forest - RF). Ο αλγόριθμος τυχαίο δάσος παίρνει το όνομα του από το γεγονός ότι χρησιμοποιεί πολλά δένδρα αποφάσεων για να εξάγει το αποτέλεσμα. Το Τυχαίο Δάσος μια μέθοδος κατηγοριοποίησης, η οποία χρησιμοποιεί ένα μεγάλο αριθμό από ταξινομητικά και παλινδρομικά δένδρα (Classification and Regression Trees - CART) με σκοπό να παρέχει υψηλότερη ακρίβεια σε σχέση με ένα μεμονωμένο δένδρο αποφάσεων. Άλλα πλεονεκτήματα είναι ότι τα δέντρα επιτρέπουν εύκολα την ερμηνεία των δεδομένων μέσω της σημαντικότητας των μεταβλητών. Επίσης, μπορούν να χειριστούν αποτελεσματικά μεγάλα σύνολα δεδομένων.

Ένα δέντρο είναι χτισμένο με τη λογική από την κορυφή προς τα κάτω (top-down) με αναδρομικό διαχωρισμό του χώρου των μεταβλητών. Σε κάθε βήμα, προσδιορίζεται ένα διαχωριστικό σημείο για κάθε μεταβλητή, έτσι ώστε να ελαχιστοποιείται ένα συγκεκριμένο κριτήριο στα δύο νέα υποσύνολα και η μεταβλητή που έχει ως αποτέλεσμα την ελάχιστη τιμή επιλέγεται να διαχωριστεί σε αυτόν τον κόμβο. Αυτή η διαδικασία ολοκληρώνεται όταν ένας κόμβος περιέχει έναν ελάχιστο αριθμό δειγμάτων ή όταν το κριτήριο δεν μπορεί να ελαχιστοποιηθεί παραπάνω από ένα όριο. Το προεπιλεγμένο κριτήριο τερματισμού είναι ο δείκτης Gini, ο οποίος δείχνει κατά πόσο τα δείγματα που καταλήγουν σε έναν κόμβο είναι «καθαρά», δηλαδή ανήκουν σε μία μόνο κλάση. Στη συνέχεια, το δέντρο μπορεί να κλαδευτεί για απλοποίηση και για να

αποφευχθεί η υπερ-εκπαίδευση, βελτιώνοντας την ακρίβεια της δοκιμής, ωστόσο στο Random Forest δεν πραγματοποιείται αυτή η λειτουργία.

Τα Random Forest δημιουργούν ένα μεγάλο αριθμό δένδρων χωρίς κλάδεμα, τα οποία διαφέρουν αρκετά μεταξύ τους λόγω της τυχαίας κατασκευής τους. Έτσι, τα δένδρα δεν συσχετίζονται μεταξύ τους, οπότε τα Random Forest μπορούν να αποφύγουν την υπερ-προσαρμογή (overfit) στα δεδομένα εκπαίδευσης και μπορούν να επιτύχουν μεγαλύτερη ακρίβεια από ένα μεμονωμένο δέντρο. Η τυχαιότητα χρησιμοποιείται σε δύο σημεία: α) κάθε δέντρο εκπαιδεύεται σε ένα διαφορετικό υποσύνολο των δεδομένων (bagging), με ομοιόμορφη δειγματοληψία από τα αρχικά δείγματα, με αποτέλεσμα να περιέχει κατά μέσο όρο το 63% όλων των δειγμάτων και β) οι μεταβλητές που εξετάζονται για την καλύτερη διαίρεση σε κάθε κόμβο επιλέγονται τυχαία (παράμετρος mtry). Τέλος, η κλάση στην έξοδο καθορίζεται από την πλειοψηφία της εξόδου των μεμονωμένων δέντρων. Σε κάθε δένδρο δίνεται τυχαία ένα υποσύνολο των δεδομένων, ώστε κάθε δένδρο να είναι διαφορετικό και να αποφευχθεί η υπερπροσαρμογή (overfit) στα δεδομένα εκπαίδευσης. Εδώ χρησιμοποιήσαμε 100 δένδρα και την προεπιλεγμένη τιμή του mtry που είναι ίση με την τετραγωνική ρίζα του αριθμού των χαρακτηριστικών.

Επίσης, για σύγκριση της απόδοσης, δοκιμάσαμε τον αλγόριθμο των Κ πλησιέστερων γειτόνων (K-nearest neighbors, KNN). Αυτός είναι ένας πολύ απλός και αρκετά διαδεδομένος αλγόριθμος ταξινόμησης. Για κάθε δείγμα των δεδομένων δοκιμής, βρίσκει τους Κ κοντινότερους γείτονες από τα δεδομένα εκπαίδευσης. Η εύρεση των κοντινότερων γειτόνων πραγματοποιείται με κάποια μετρική, όπως η Ευκλείδεια απόσταση. Στη συνέχεια βρίσκεται ποια κατηγορία από τους Κ γείτονες έχει την πλειοψηφία και επιστρέφεται ως έξοδος. Σε αυτή την εργασία οι παράμετροι του KNN τέθηκαν ως αριθμός γειτόνων Κ=5 και ως απόσταση η Ευκλείδεια.

Τέλος, για αναλύσουμε σε βάθος τα αποτελέσματα της μεθόδου MasterMovelets (Ferrero et al., 2020), μιας και έχουμε πάρα πολλά χαρακτηριστικά, πραγματοποιήσαμε ανάλυση κυρίων συνιστωσών (principal components analysis – PCA). Η PCA βρίσκει ένα νέο σύνολο ορθοκανονικών μεταβλητών που είναι γραμμικός συνδυασμός των παλιών μεταβλητών, με σκοπό οι νέες μεταβλητές να έχουν τη μεγαλύτερη διασπορά. Πιο συγκεκριμένα, για την πρώτη συνιστώσα, σε ένα πίνακα X, όπου η κάθε στήλη *j* είναι ένα διάνυσμα x_j θέλουμε να βρούμε ένα διάνυσμα *w* ώστε:

$$w_1 = arg\ max\ (\Sigma_j\ (x_j\ w_1)^2\ )$$

με τον περιορισμό $|w_1|=1$.

Στη συνέχεια αφαιρούμε την πρώτη συνιστώσα από τα δεδομένα ως εξής:

$$X_2 = X - Xw_1\ w_1^T$$

Μετά, επαναλαμβάνουμε τη διαδικασία στον πίνακα X_2 για να βρούμε την επόμενη συνιστώσα w_2 κλπ.

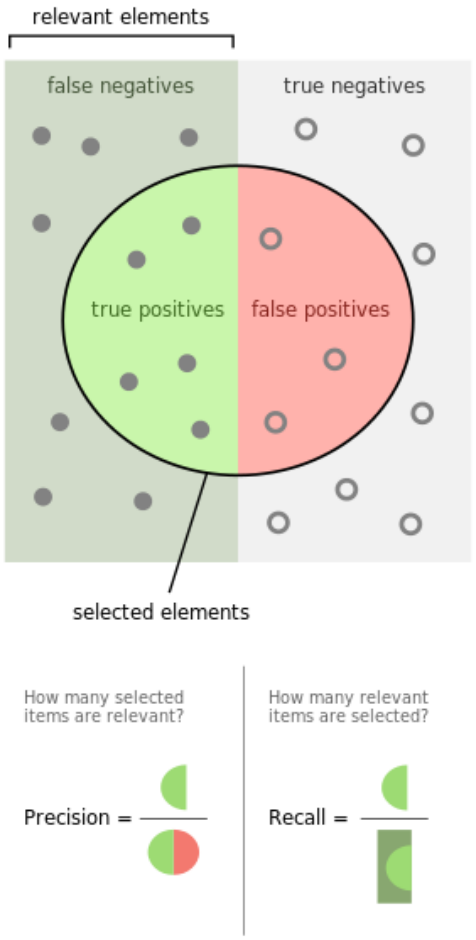
Αποδεικνύεται ότι τα w_k είναι τα ιδιοδιανύσματα του X διατεταγμένα σύμφωνα με το μέγεθος των αντίστοιχων ιδιοτιμών.

Ο κώδικας για το βήμα της ταξινόμησης και της ανάλυσης κύριων συνιστωσών είναι γραμμένος σε γλώσσα προγραμματισμού Python έκδοση 3.8.3.

4.2.3 Αποτελέσματα

Αρχικά αναφέρουμε ότι ο αλγόριθμος MasterMovelets (Ferrero et al., 2020) προσδιόρισε 2828 Movelets από 26546 υποψήφια, ενώ μετά από φιλτράρισμα έμειναν 440 Movelets. Ο χρόνος εκτέλεσης ήταν 454.117 δευτερόλεπτα.

Για την αξιολόγηση των αποτελεσμάτων του αλγορίθμου κατηγοριοποίησης χρησιμοποιούμε τις μετρικές precision, recall, F1 score και accuracy. Η μετρική precision είναι το ποσοστό των δειγμάτων που ταξινομήθηκαν σωστά σε μια κλάση ως προς τα συνολικά δείγματα που ο αλγόριθμος ταξινομεί σε αυτή την κλάση. Η μετρική recall είναι το ποσοστό των δειγμάτων που ταξινομήθηκαν σωστά σε μια κλάση ως προς το αριθμό των δειγμάτων που ανήκουν πραγματικά στην κλάση. Μια γραφική αναπαράσταση των δύο αυτών μετρικών φαίνεται στην εικόνα 1.



Εικόνα 23. Ορισμός μετρικών precision και recall. Πηγή: https://en.wikipedia.org/wiki/Precision_and_recall#/media/File:Precisionrecall.svg

Το F1 score συνδυάζει τις δύο προηγούμενες μετρικές και ορίζεται ως:

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

Η μετρική accuracy δείχνει συνολικά το ποσοστό των σωστών ταξινομημένων δειγμάτων.

Τα αποτελέσματα της κατηγοριοποίησης με RandomForest για το υποσύνολο test φαίνονται παρακάτω στον Πίνακα 1. Η στήλη support δείχνει πόσα δείγματα ανήκουν σε κάθε κλάση. Η γραμμή μέσος όρος δείχνει το μέσο όρο για τις 4 κλάσεις, ενώ ο ζυγισμένος μέσος όρος λαμβάνει υπόψη το μέγεθος της κάθε κλάσης.

Πίνακας 4. Αποτελέσματα κατηγοριοποίησης για τον αλγόριθμο Τυχαίο Δάσος.

Κλάση	**Precision**	**Recall**	**F1 score**	**Accuracy**	**Support**
0	0,87	1	0,93	-	13
1	1	0,91	0,95	-	11
2	1	1	1	-	9
3	0,88	0,78	0,82	-	9
Μέσος όρος	0,94	0,92	0,93	-	-
Ζυγισμένος μέσος όρος	0,93	0,93	0,93	0,93	-

Παρατηρούμε ότι το αποτέλεσμα είναι αρκετά υψηλό σε όλες τις μετρικές στο σύνολο test. Αυτό σημαίνει ότι τα δεδομένα ταξινομούνται με πολύ λίστα σφάλματα. Η ακρίβεια που συνοψίζει το αποτέλεσμα για τις 4 κλάσεις έχει τιμή 93%.

Επιπλέον, δοκιμάσαμε τον K-Nearest_Neighbor (KNN) σαν μέθοδο κατηγοριοποίησης. Τα αποτελέσματα φαίνονται στον πίνακα 2.

Πίνακας 5. Αποτελέσματα κατηγοριοποίησης για τον αλγόριθμο KNN.

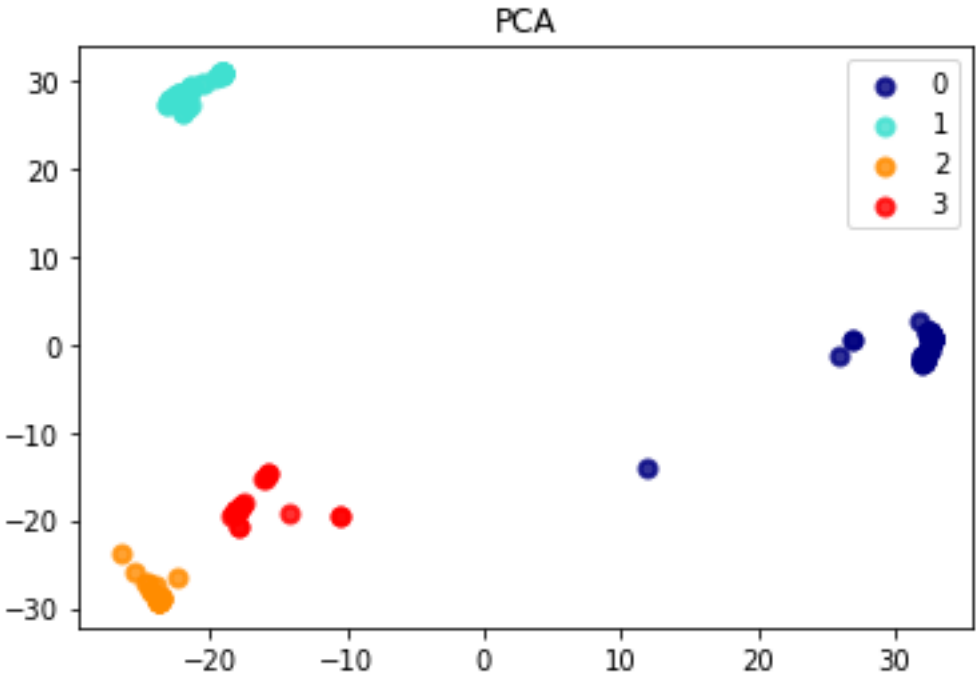
Κλάση	**Precision**	**Recall**	**F1 score**	**Accuracy**	**Support**
0	1	1	1	-	13
1	1	0,91	0,95	-	11
2	1	1	1	-	9
3	0,9	1	0,95	-	9
Μέσος όρος	0,97	0,98	0,97	-	-
Ζυγισμένος μέσος όρος	0,98	0,98	0,98	0,98	-

Παρατηρούμε ότι ο KNN έχει υψηλότερη ακρίβεια από τον αλγόριθμο Τυχαίο Δάσος. Συγκεκριμένα, η ακρίβεια είναι 98%. Αυτό οφείλεται στην καλύτερη κατηγοριοποίηση στις κλάσεις 0 και 3, που προηγουμένως είχαν χαμηλότερες τιμές στις μετρικές precision και recall.

Οπότε, διαπιστώνουμε ότι ο αλγόριθμος MasterMovelets (Ferrero et al., 2020) είναι αποτελεσματικός στην κατηγοριοποίηση των τροχιών του συνόλου δεδομένων Hermoupolis. Επίσης, παίρνουμε πολύ καλά αποτελέσματα με δύο διαφορετικούς αλγορίθμους κατηγοριοποίησης. Επιπλέον, ένας πιθανός λόγος που έχουμε τόσο καλή κατηγοριοποίηση είναι ότι έχουμε πάρα πολλά χαρακτηριστικά, πολύ περισσότερα από

τα δείγματα, καθώς για 100 αντικείμενα έχουμε 1457 χαρακτηριστικά, οπότε είναι σχετικά εύκολο να βρεθεί ένας συνδυασμός χαρακτηριστικών που να δίνει υψηλή ακρίβεια κατηγοριοποίησης.

Τέλος, για αναλύσουμε σε βάθος τα αποτελέσματα της μεθόδου MasterMovelets (Ferrero et al., 2020), μιας και έχουμε πάρα πολλά χαρακτηριστικά, πραγματοποιήσαμε ανάλυση κυρίων συνιστωσών (principal components analysis – PCA). Συγκεκριμένα, πήραμε τις δύο κύριες συνιστώσες για απεικόνιση σε δύο άξονες, οι οποίες εξηγούν το 54.2% και 34.9% της διακύμανσης των δεδομένων. Οπότε, αθροιστικά, μόνο οι δύο αυτές συνιστώσες εξηγούν περίπου το 89% της διακύμανσης των δεδομένων. Οι συνιστώσες και η κατηγορίες των δεδομένων φαίνονται στην Εικόνα 2. Οπότε, είναι φανερό ότι το αποτέλεσμα της MasterMovelets (Ferrero et al., 2020) παράγει δεδομένα που είναι εύκολο να διαχωριστούν σε ομάδες στο χώρο, οπότε έτσι προκύπτει η υψηλή ακρίβεια κατηγοριοποίησης.



Εικόνα 24. Ανάλυση κυρίων συνιστωσών.

5. Μελλοντικές επεκτάσεις

Η αυξανόμενη χρήση συσκευών που γνωρίζουν την τοποθεσία έχει οδηγήσει σε αυξανόμενη διαθεσιμότητα δεδομένων τροχιάς που αντιπροσωπεύουν την κίνηση κινούμενων αντικειμένων. Η δυνατότητα αυτών των δεδομένων στην επίλυση σημαντικών ερευνητικών προβλημάτων έχει αυξήσει το ενδιαφέρον των ερευνητών για μεθόδους ανάλυσης γι 'αυτά. Συγκεκριμένα, εμπνευσμένοι από την πρόοδο στην εξόρυξη κλασικών μεγάλων συνόλων δεδομένων, οι ερευνητές έχουν αναπτύξει διαφορετικές μεθόδους εξόρυξης δεδομένων για τροχιές. Ωστόσο, δεν έχει διερευνηθεί η παροχή μιας ολοκληρωμένης οπτικής επίλυσης προβλημάτων και μεθόδων στην εξόρυξη τροχιάς, καθώς και οι εφαρμογές που βασίζονται σε λύσεις σε αυτά τα προβλήματα.

Στο θεωρητικό τμήμα παρουσιάσαμε διάφορες σύγχρονες μεθόδους για αποτελεσματική ανάλυση τροχιών. Στο πρακτικό τμήμα υλοποιήσαμε κατηγοριοποίηση υπο-τροχιών με βάση τη μέθοδο MasterMovelets (Ferrero et al., 2020). Η εφαρμογή έγινε στο σύνολο δεδομένων Hermoupolis, το οποίο περιλαμβάνει προσομοίωση κίνησης 100 αντικειμένων στην Αθήνα με 4 πρότυπα κίνησης. Η μέθοδος MasterMovelets (Ferrero et al., 2020) σε συνδυασμό με κατηγοριοποίηση χρησιμοποιώντας δύο διαφορετικές μεθόδους, KNN και RandomForest, κατάφερε να ξεχωρίσει τις 4 κατηγορίες τροχιών με πολύ υψηλή ακρίβεια, 98% και 93% αντίστοιχα. Αυτό αποδεικνύει την αποτελεσματικότητά της σε δεδομένα τροχιών με διαφορετικού τύπου υπο-τροχιές.

Η εργασία που παρουσιάζεται μπορεί να επεκταθεί σε μια συγκριτική μελέτη αλγορίθμων που εφαρμόζει μια συγκεκριμένη μέθοδο εξόρυξης τροχιάς σε

συγκεκριμένα προβλήματα εφαρμογής με δυνατότητα παροχής συστάσεων για την επιλογή μεταξύ των αλγορίθμων. Μία από τις τρέχουσες τάσεις στην ανάλυση δεδομένων κίνησης είναι να συσχετιστεί η κίνηση με το περιβάλλον ενσωμάτωσής της. Λαμβάνοντας υπόψη αυτήν την τάση, η παρούσα εργασία μπορεί να επεκταθεί με την προσαρμογή αλγορίθμων που θα εφαρμόζουν ορισμένες από τις παρουσιαζόμενες μεθόδους για την εκτέλεση εξόρυξης δεδομένων τροχιάς με γνώμονα το πλαίσιο για επιλεγμένα προβλήματα εφαρμογών, και την σύγκριση των αποτελεσμάτων τους.

6. Βιβλιογραφία

Alvares, L. O., Bogorny, V., Kuijpers, B., Moelans, B., Fern, J. A., Macedo, E. D., & Palma, A. T. (2007). Towards semantic trajectory knowledge discovery. Data Mining and Knowledge Discovery, 12.

Atev, S., Masoud, O. & Papanikolopoulos, N. (2006). Learning traffic patterns at intersections by spectral clustering of motion trajectories. *In Proceedings of the International Conference on Intelligent Robots and Systems (IROS'06)*. 4851–4856.

Bashir, F. I. & Khokhar, A. A. & Schonfeld, D. (2007). Object trajectory-based activity classification and recognition using hidden Markov models. *IEEE Trans. Image Process.* 16, 7, 1912‑1919.

Bashir, F. I. & Khokhar, A. A. & Schonfeld, D. (2007). Real-time motion trajectory-based indexing and retrieval of video sequences. *IEEE Trans. Multimedia* 9, 1, 58‑65.

Bashir, F. I., Khokhar, A. A. & Schonfeld, D. (2007). Object trajectory-based activity classification and recognition using hidden Markov models. *IEEE Trans. Image Process. 16*, 7, 1912–1919.

Bian, J., Tian, D., Tang, Y., & Tao, D. (2019). Trajectory Data Classification: A Review. *ACM Trans. Intell. Syst. Technol.* 10, 4, Article 33, pp 33.

Cai G., Lee K. & Lee, I. (2016). A framework for mining semantic-level tourist movement behaviors from geo-tagged photos. *In Proceedings of the Australasian Joint Conference on Artificial Intelligence. Springer*, 519–524.

Chen, Z., Shen, H. T., Zhou, X., Zheng, Y. & Xie, X. (2010). Searching trajectories by locations: An efficiency study. *In Proceedings of the ACM SIGMOD International Conference on Management of Data. ACM*, 255–266.

Choi, D. W., Pei, J., & Heinis, T. (2017). Efficient mining of regional movement patterns in semantic trajectories. *Proc. VLDB Endow. 10*, 13, 2073–2084.

Dai, Q., Jiang, Y. J., Xue, X., Liu, W. & Ngo, C. W. (2012). Trajectory-based modeling of human actions with motion reference points. *In Proceedings of the European Conference on Computer Vision. Springer*, 425– 438.

Feng, Z., & Zhu, Y. (2016). A Survey on Trajectory Data Mining: Techniques and Applications. *IEEE Access*, 4, 2056-2067.

Ferreira, N., Klosowski, J. T., Scheidegger, C. E. & Silva, C. T. (2013). Vector field k-means: Clustering trajectories by fitting multiple vector fields. *In Computer Graphics Forum*, Vol. 32. Wiley Online Library, 201–210.

Ferrero, C.A., Alvares, L.O., Zalewsky, W., & Bogorny, V. (2018). Movelets: exploring relevant subtrajectories for robust trajectory classification. In: Proceedings of the 33rd ACM SAC, ACM, Pau, France, p.p. 1–8

Ferrero, C. A., Petry, L. M., Alvares, L. O., da Silva, C. L., & Zalewski, W., & Bogorny, V. (2020). MASTERMOVELETS: discovering heterogeneous movelets for multiple aspect trajectory classification. *Data Mining and Knowledge Discovery*, 34, p.p 652–680.

Furletti, B., Cintia P., Renso, C. & Spinsanti, L. (2013). Inferring human activities from GPS tracks. *UrbComp '13: Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, 5, pp. 1–8. doi: 10.1145/2505821.2505830

Gurung S., Lin, D., Jiang, W., Hurson, A.,& Zhang, R. (2014). Traffic information publication with privacy preservation. *ACM Trans. on Intell. Syst. Technol. 5*, 3, pp. 44.

Hu, H. & Feng, J. & Zhou, J. (2015). Exploiting unsupervised and supervised constraints for subspace clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* 37, 8, 1542‑1557.

Hu, W. & Li, X. & Tian, G. & Maybank, S. & Zhang, Z. (2013). An incremental DPMM-based method for trajectory clustering, modeling, and retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 5 (2013), 1051‑1065.

Hung, C. C., Peng, W. C., & Lee, W. C. (2015). Clustering and aggregating clues of trajectories for mining trajectory patterns and routes. The VLDB Journal, 24(2), 169-192.

Ishikawa, S., & Uemura, H., & Mikolajczyk, K. (2008). Feature tracking and motion compensation for action recognition. *In Proceedings of the British Machine Vision Conference (BMVC'08)*. 1–10.

Kong, X., Li, M., Ma, K., Tian, K., Wang, M., Ning Z. & Xia, F. (2018). Big trajectory data: A survey of applications and services. *IEEE Access*, 6, 58295–58306.

Kumar, S., Dai, Y. & Li, H. (2017). Spatio-temporal union of subspaces for multi-body non-rigid structure-from-motion. *Pattern Recogn*. 71, 428–443.

Li, X., Han, J., Lee, J. G. & Gonzalez, H. (2007). Traffic density-based discovery of hot routes in road networks. *In Proceedings of the International Symposium on Spatial and Temporal Databases. Springer*, 441–459.

Liou, W. G., Hsieh, C. Y. & Lin, W. Y. (2011). Trajectory-based sign language recognition using Discriminant Analysis in higher-dimensional feature space. In *Proceedings of the IEEE International Conference on Multimedia and Expo*. IEEE, 1‑4.

Lu, E. H. C., & Tseng, V. S. (2009, May). Mining cluster-based mobile sequential patterns in location-based service environments. In *2009 Tenth International Conference on Mobile Data Management: Systems, Services and Middleware* (pp. 273-278). IEEE.

Mazimpaka, J. D. & Timpf, S. (2016). Trajectory data mining: A review of methods and applications. *J. Spatial Info. Sci. 2016,* 61–99.

Mo, Y., Wu, D. & Du, Y. (2015). Application of trajectory clustering and regionalization to ocean eddies in the South China Sea. *In Proceedings of the 2nd IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM'15)*. IEEE, 45–48.

Monreale, A., Pinelli, F., Trasarti, R., & Giannotti, F. (2009, June). Wherenext: a location predictor on trajectory pattern mining. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 637-646).

Morris, B. T. & Trivedi, M. M. (2008). A survey of vision-based trajectory learning and analysis for surveillance. *IEEE Trans. Circ. Syst. Video Technol. 18,* 8, 1114–1127.

Morris, B. T. & Trivedi, M. M. (2008). A survey of vision-based trajectory learning and analysis for surveillance. *IEEE Trans. Circ. Syst. Video Technol.* 18, 8, 1114‑1127.

Naftel, A. & Khalid, S. (2006). Motion trajectory learning in the DFT-coefficient feature space. In *Proceedings of the 4th IEEE International Conference on Computer Vision Systems (ICVS'06)*. IEEE, 47‑47.

Nikam, S. S. (2015). A comparative study of classification techniques in data mining algorithms. Oriental journal of computer science & technology, 8(1), 13-19.

Schenk, J. & Rigoll, G. (2006). Novel hybrid NN/HMM modelling techniques for on-line handwriting recognition. In *Proceedings of the 10th International Workshop on Frontiers in Handwriting Recognition*. Suvisoft.

Wang, H., Oneata, D., Verbeek, J. & Schmid, C. (2016). A robust and efficient video representation for action recognition. *Int. J. Comput*. Vision 119, 3, 219–238.

Yao, T., Wang, Z., Xie, Z., Gao, J. & Feng, D. D. (2017). Learning universal multiview dictionary for human action recognition. *Pattern Recogn*. 64, 236–244.

Ying, J. J. C., Lee, W. C., Weng, T. C. & Tseng, V. S. (2011). Semantic trajectory mining for location prediction. *In Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 34–43.

Ying, J. J. C., Lu, E. H. C., Lee, W. C., Weng, T. C. & Tseng, V. S. (2010). Mining user similarity from semantic trajectories. *In Proceedings of the 2nd ACM*

SIGSPATIAL International Workshop on Location-based Social Networks. ACM, 19–26.

Yoo, J. S., Shekhar, S., Smith, S., & Kumquat, J. P. (2004). A partial join approach for mining colocation patterns. *In Proceedings of the 12th Annual ACM International Workshop on Geographic Information Systems*. ACM, 241–249.

Yuan G., Sun, P., Zhao, J., Li D. & Wang, C. (2017). A review of moving object trajectory clustering algorithms. *Artific. Intell. Rev*. 47, 1, 123–144.

Zheng, B., Yuan, N. J., Zheng, K., Xie, X., Sadiq, S. & Zhou, X. (2015). Approximate keyword search in semantic trajectory database. *In Proceedings of the IEEE 31st International Conference on Data Engineering (ICDE'15)*. IEEE, 975–986.

Zheng, Y., Chen, Y., Li, Q., Xie, X., & Ma, W. Y. (2010). Understanding transportation modes based on gps data for web applications. *ACM Trans Web TWEB.* 4(1), 1–36.

https://en.wikipedia.org/wiki/Precision_and_recall#/media/File:Precisionrecall.svg

http://infolab.cs.unipi.gr/pubs/pkdd2013/