ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ

ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΧΡΗΣΤΩΝ ΚΟΙΝΩΝΙΚΩΝ ΔΙΚΤΥΩΝ ΘΕΣΗΣ ΒΑΣΕΙ ΧΩΡΟ-ΚΕΙΜΕΝΙΚΩΝ ΑΠΟΤΥΠΩΜΑΤΩΝ

Ευτυχία Κολάση

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

Νοέμβριος 2022

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

Τα μέλη της Επιτροπής ήταν:

- Πελέκης Νικόλαος (Επιβλέπων), Αναπληρωτής Καθηγητής

- Σωτήριος Μπερσίμης, Αναπληρωτής Καθηγητής

- Ελευθέριος Κοφίδης, Αναπληρωτής Καθηγητής

Η έγκριση της Διπλωματική 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 USERS OF LOCATION-BASED SOCIAL NETWORKS BASED ON SPATIO-TEXTUAL FOOTPRINTS

By

Eftychia Kolasi

MSc Dissertation

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

November 2022

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή της διπλωματικής εργασίας, κ.Πελέκη Νικόλαο, για τις γνώσεις, τις υποδείξεις και την καθοδήγηση που μου παρείχε κατά την διάρκεια συγγραφής της εργασίας.

Περίληψη

Με την εξέλιξη της τεχνολογίας και την ολοένα και την καθημερινή χρήση συσκευών με πρόσβαση στο διαδίκτυο, δημιουργείται ένας μεγάλος όγκος χωροχρονικών δεδομένων. Αυτό το είδος δεδομένων έχουν απασχολήσει ιδιαίτερα την επιστημονική κοινότητα τα τελευταία χρόνια. Τα δεδομένα κίνησης καταγράφουν την θέση ενός κινούμενου αντικειμένου κάθε χρονική στιγμή και αποτελούν αντικείμενο μελέτης της συμπεριφοράς όχι μόνο των ζώων και ανθρώπων αλλά και επίλυσης προβλημάτων όπως είναι η πρόβλεψη τροχιάς ενός ανεμοστρόβιλου και η επίλυση κυκλοφοριακών προβλημάτων.

Στην παρούσα εργασία επιχειρείται η ταξινόμηση χρηστών Twitter μέσα από τις τροχιές που δημιουργούν, κάνοντας χρήση αλγορίθμων Μηχανικής Μάθησης. Η προεπεξεργασία των δεδομένων βασίστηκε στην μέθοδο MasterMovelets (Ferrero et al., 2020) που ανακαλύπτει σχετικές υποτροχιές με διαφορετικές και ετερογενείς διαστάσεις και ποικίλου μήκους. Το αποτέλεσμα της μεθόδου είναι ένα σύνολο δεδομένων το οποίο χρησιμοποιείται στη συνέχεια ως δεδομένα εισόδου για τους αλγορίθμους ταξινόμησης. Τα δεδομένα που χρησιμοποιήθηκαν προέρχονται από tweets που πραγματοποιήθηκαν στην Σαντορίνη την περίοδο 2018-2019. Η ανάλυση έδειξε ότι ίσως η μέθοδος MasterMovelets να μην ταιριάζει στα δεδομένα μας καθώς οι ταξινομητές έδιναν χειρότερα αποτελέσματα από αυτά της εργασίας των Ferrero et al (2020). Στη συνέχεια επιχειρήθηκε η ταξινόμηση των χρηστών χωρίς να προηγηθεί κάποια περίπλοκη προεπεξεργασία η οποία όμως δεν έχει στόχο την σύγκριση με την μέθοδο MasterMovelets καθώς στην βάση τους διαφέρουν, η μία πραγματοποιεί εξόρυξη τροχιών πάνω στις οποίες γίνεται στη συνέχεια κατηγοριοποίηση ενώ η άλλη περνάει αμέσως στην κατηγοριοποίηση των σημείων.

Abstract

With the development of technology and the increasingly daily use of devices with internet access, a large amount of spatiotemporal data are being generated. This type of data has been of particular concern to the scientific community in recent years. Motion data records the position of a moving object at any moment and is the subject of studying the behavior of not only animals and humans but also solving problems such as predicting the trajectory of a tornado and solving traffic problems.

In this work, the classification of trajectories coming from Twitter users is attempted, using Machine Learning algorithms. Data pre-processing was based on the MasterMovelets method (Ferrero et al., 2020) which discovers relevant sub-trajectories with different and heterogeneous dimensions and varying lengths. The result of the pre-processing step is a data set which is then used as an input for the classification algorithms. The data used come from tweets in Santorini during the period 2018-2019. The analysis showed that the MasterMovelets method may not fit our data as the classifiers gave worse results than those in the work of Ferrero et al (2020). We attempted also classifying without having performed previously any complicated pre-processing. Our aim was not to compare the results coming from these methods as they are fundamentally different. MasterMovelts is based on trajectory mining, while the experimental method classifies trajectory points.

ΠΕΡΙΕΧΟΜΕΝΑ

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Εισαγωγή

Οι εξελίξεις στον τομέα της τεχνολογίας των πληροφοριών και των επικοινωνιών, ιδίως στον εντοπισμό θέσης κινητής τηλεφωνίας και στις ασύρματες επικοινωνίες, έχουν καταστήσει εύκολη τη δημιουργία δεδομένων κίνησης, δηλαδή δεδομένων που περιλαμβάνουν τη γεωγραφική θέση στη διάσταση του χρόνου. Η κατανόηση και ανάλυση των δεδομένων αυτών μπορεί να συνεισφέρει στην ανάπτυξη εφαρμογών υγειονομικής περίθαλψης, στον σχεδιασμό δημόσιων συγκοινωνιών, στις υπηρεσίες και διαφημίσεις βάσει τοποθεσίας, στην πρόβλεψη εγκλήματος, την πρόληψη επιδημιών κ.α.

Η κατηγοριοποίηση των τροχιών που παράγουν τα κινούμενα αντικείμενα αποτελεί ένα πεδίο έρευνας όπου πραγματοποιείται ανάλυση της συμπεριφοράς τους ώστε να δημιουργηθεί ένας ταξινομητής που θα είναι ικανός να διακρίνει τα διαφορετικά μοτίβα κίνησης. Στην εργασία στην οποία βασίστηκε η παρούσα διπλωματική, οι Ferrero et al.(2020) προτείνουν μια μέθοδο για την ανακάλυψη υποτροχιών χωρίς την ανάγκη ενός προκαθορισμένου κριτηρίου διαχωρισμού.

Στο πρώτο κεφάλαιο της εργασίας πραγματοποιείται μια εισαγωγή στα δεδομένα κίνησης, στα κινούμενα αντικείμενα, στις τροχιές. Επίσης παρουσιάζονται εργασίες που σχετίζονται με κατηγοριοποίηση τροχιών και γίνεται σύντομη ανάλυση της μεθοδολογίας. Στο δεύτερο κεφάλαιο γίνεται αναλυτική περιγραφή της μεθόδου MasterMovelets και των αλγορίθμων της. Στο τέταρτο κεφάλαιο παρουσιάζονται οι τεχνικές μηχανής μάθησης που χρησιμοποιήθηκαν στο πρακτικό μέρος της εργασίας το οποίο αποτελεί το πέμπτο κεφάλαιο.

Κεφάλαιο 1

Θεωρητικό Υπόβαθρο

Τα κινούμενα αντικείμενα δημιουργούν τροχιές, χρονικές ακολουθίες θέσεων που ορίζουν καμπύλες στο χώρο. Η μελέτη της τροχιάς που δημιουργούν τα κινούμενα αντικείμενα είναι θεμελιώδης για την κατανόηση της συμπεριφοράς τους. Δεν μας ενδιαφέρουν μόνο τα αρχικά και τελικά σημεία μιας τροχιάς, αλλά πιο πολύ η διαδρομή που ακολούθησε ένα τέτοιο αντικείμενο, δηλαδή τα ενδιάμεσα σημεία. Για παράδειγμα η ακρίβεια στην πρόβλεψη των τροχιών που ακολουθούν οι ανεμοστρόβιλοι μπορεί να ελαττώσει τις ζημιές που αυτοί προκαλούν και να σώσει ανθρώπινες ζωές.

1.1. Χωρικοί Τύποι Δεδομένων

Τα χωρικά δεδομένα είναι εκείνα τα δεδομένα που σχετίζονται με τοποθεσίες (σημεία) ή περιοχές στο διάστημα. Ανάλογα με την εφαρμογή, τα δεδομένα μπορούν να περιγράψουν περιοχές σε 2D και 3D Ευκλείδειο χώρο ή ακόμα και άλλους τύπους χωρικών συντεταγμένων. Παραδείγματα τέτοιων δεδομένων μπορεί να είναι οι σημειακές θέσεις των αντικειμένων στο διάστημα, μια περιοχή ακανόνιστου σχήματος που περιγράφει μια λίμνη ή μια ευθεία γραμμή που δείχνει πού περνούν τα αυτοκίνητα. Τα χωρικά δεδομένα χρησιμοποιούνται για να περιγράψουν στοιχεία που αφορούν χωρικές διαστάσεις. Η περιγραφή των στοιχείων μπορεί να γίνει με τη χρήση:

- Σημείων
- Γραμμών,
- Πολυγώνων.

i. Σημεία

Τα σημεία χρησιμοποιούνται για να περιγράψουν σημειακές θέσεις στο χώρο. Γενικά, τα αντικείμενα στο χώρο καταλαμβάνουν κάποια περιοχή και δεν είναι σημειακά. Η θέση ενός αντικειμένου είναι ουσιαστικά ένα σύνολο σημείων στον χώρο που καταλαμβάνει. Για παράδειγμα, η θέση ενός κτιρίου, λαμβάνοντας υπόψη τον δισδιάστατο χώρο, είναι ισοδύναμη με το σχήμα που σχηματίζεται από την προβολή του κτιρίου στο επίπεδο. Εάν το ενδιαφέρον της εφαρμογής στραφεί σε θέσεις κτιρίου σε επίπεδο τετραγώνου, τότε ο χώρος που καταλαμβάνει το κτίριο έχει σημασία. Ωστόσο, οι περισσότερες εφαρμογές ενδιαφέρονται για τη θέση των αντικειμένων στην κλίμακα του χάρτη. Σε τέτοιες διαστάσεις, ο χώρος που καταλαμβάνει ένα αντικείμενο δεν είναι σημαντικός και αρκεί να υπολογιστεί ένα μόνο σημείο που αντιπροσωπεύει ολόκληρο το αντικείμενο. Άρα ο όγκος του αντικειμένου εκφυλίζεται σε μια σημειακή θέση.

ii. Γραμμές

Με τις γραμμές γίνεται αναπαράσταση περιοχών με κοινά χαρακτηριστικά όπως μια ισοϋψής καμπύλη σε ένα χάρτη ή τα σύνορα μεταξύ χωρών. Επίσης, για την αναπαράσταση της τροχιάς που ακολούθησε ένα αντικείμενο μπορεί να χρησιμοποιηθεί μια γραμμή. Οι γραμμές συνενώνουν μια ακολουθία από σημεία και καταδεικνύουν τη διαδρομή στην οποία κινήθηκε το αντικείμενο στο χώρο.

iii. Πολύγωνα

Για την αναπαράσταση περιοχών ή επιφανειών χρησιμοποιούνται πολύγωνα. Επίσης, όταν το μέγεθος ενός αντικειμένου έχει ιδιαίτερη σημασία τότε το αντικείμενο περιγράφεται από ένα πολύγωνο αντί ενός σημείου. Για παράδειγμα, με τη χρήση πολυγώνων μπορεί να αναπαρασταθεί η επιφάνεια που καλύπτει μια λίμνη ή ένα μεγάλο πλοίο που βρίσκεται στη θάλασσα.

1.2. Χωροχρονικά Δεδομένα

Χωροχρονικά δεδομένα θεωρούνται εκείνα που εκτός από την πληροφορία του χώρου, δηλαδή το σημείο στο οποίο βρίσκεται κάποιο αντικείμενο, έχουν και χρονική πληροφορία. Η χρήση της χωροχρονικών δεδομένων καθιστάται απαραίτητη όταν μας ενδιαφέρουν αντικείμενα τα οποία δεν είναι στατικά. Στατικά αντικείμενα είναι κτίριο ή μια λίμνη ενώ κινούμενα είναι τα ζώα, τα αεροπλάνα κ.α. Για παράδειγμα, έχοντας δεδομένα χώρου και χρόνου, μπορούμε να ξέρουμε μέσα από τις τροχιές δύο κινούμενων αντικειμένων, αν αυτά κάποια στιγμή συναντήθηκαν και με πόση διαφορά προσπέρασε το ένα το άλλο.

1.3. Κινούμενα Αντικείμενα

Τα κινούμενα αντικείμενα είναι μια γνωστή κατηγορία χωροχρονικών δεδομένων, η οποία αντιπροσωπεύει αντικείμενα των οποίων οι χωρικές θέσεις ή εκτάσεις μεταβάλλονται συνεχώς με την πάροδο του χρόνου. Παραδείγματα τέτοιων αντικειμένων είναι τα οχήματα, χρήστες κινητών τηλεφώνων, ανεμοστρόβιλοι, πετρελαιοκηλίδες στη θάλασσα, δασικές φωτιές, πολικές αρκούδες κ.α.. Τα ζώα είναι παράδειγμα ενός κινούμενου σημείου γιατί η θέση του αντικειμένου αλλάζει με την πάροδο του χρόνου, ενώ ένα αποψιλωμένο δάσος είναι παράδειγμα κινούμενης περιοχής αφού η έκτασή του μεταβάλλεται. Σχετικά με την διαχείριση των κινούμενων αντικειμένων στα συστήματα διαχείρισης βάσεων δεδομένων υπάρχουν δύο πρωτοβουλίες συστημάτων βάσεων δεδομένων κινούμενων αντικειμένων, Hermes και SECONDO. Επεκτείνουν ένα σύστημα τύπου DBMS SQL με νέους τύπους δεδομένων και λειτουργίες για την αναπαράσταση και τον χειρισμό των κινούμενων αντικειμένων, με βάση την άλγεβρα που προτείνεται από τους Erwig et al(2000).

1.4. Βάσεις Δεδομένων Κινούμενων Αντικειμένων

Δεδομένης της φύσης των κινούμενων αντικειμένων, οι συμβατικές βάσεις δεδομένων δε μπορούν να διαχειριστούν τα δεδομένα που προέρχονται από αυτά καθώς υποθέτουν ότι είναι στατικά. Σε αυτές περιπτώσεις χρησιμοποιούνται οι Βάσεις Δεδομένων Κινούμενων Αντικειμένων (Moving Object Databases – MOD). Τέτοιες Βάσεις Δεδομένων μπορούν να αποθηκεύσουν πληροφορίες σχετικές με τα κινούμενα αντικείμενα και επιτρέπουν την υποβολή ερωτημάτων (queries) σε αυτά.

Στις εφαρμογές MOD, κάθε αντικείμενο δεδομένων κινείται συνεχώς και αναφέρει συχνά τις τρέχουσες χωροχρονικές τιμές χαρακτηριστικών του (χωροχρονικές εγγραφές) που αντιπροσωπεύουν την τρέχουσα τοποθεσία, κατεύθυνση κίνησης και ταχύτητα στον διακομιστή βάσης δεδομένων. Οι MOD αποτελούν την βασική συνιστώσα οποιασδήποτε εφαρμογής με επίκεντρο την κινητικότητα.

Έχουν προταθεί δύο πρωτότυπες μηχανές Βάσεων Δεδομένων, η HERMES (Nikos Pelekis et al, 2015) και η SECONDO (Güting et al, 2006). Η HERMES ορίζει μια ισχυρή γλώσσα ερωτημάτων για βάσεις δεδομένων τροχιών, η οποία επιτρέπει την υποστήριξη εφαρμογών που επικεντρώνονται στην κινητικότητα, όπως οι Υπηρεσίες Βασισμένης στη Τοποθεσία (LBS). Επεκτείνει τον ορισμό δεδομένων και την γλώσσα χειρισμού των Object-Relational DBMS (ORDBMS) με χωροχρονική σημασιολογία και λειτουργικότητα βασισμένη σε προηγμένες τεχνικές χωροχρονικής ευρετηρίασης και επεξεργασίας ερωτημάτων. Η SECONDO είναι μια επεκτάσιμη πλατφόρμα DBMS κατάλληλη για την κατασκευή ερευνητικών πρωτοτύπων και για τη διδασκαλία της αρχιτεκτονικής και την υλοποίηση συστημάτων βάσεων δεδομένων. Δεν έχει σταθερό μοντέλο δεδομένων, αλλά είναι ανοιχτό για εφαρμογή νέου μοντέλου. Σκοπός του συστήματος SECONDO είναι να προσφέρει ένα γενικό πλαίσιο συστήματος βάσης δεδομένων που μπορεί να γεμίσει με υλοποιήσεις διαφόρων μοντέλων δεδομένων DBMS.

1.5. LBSN (Location Based Social Networks)

Τα μέσα κοινωνικής δικτύωσης είναι υπηρεσίες βασισμένες στο διαδίκτυο που επιτρέπουν στους χρήστες τους να δημιουργήσουν ημί-δημόσια προφίλ και παρέχουν μέσα επικοινωνίας μεταξύ των χρηστών μέσω σχολίων, προσωπικών ή άμεσων μηνυμάτων καθώς και κοινής χρήσης ψηφιακών μέσων, όπως φωτογραφίες ή βίντεο. Πηγαίνοντας ένα βήμα παρακάτω, τα μέσα κοινωνικής δικτύωσης που περιέχουν και πληροφορίες τοποθεσίας ονομάζονται Κοινωνικά

Δίκτυα Βασισμένα στην Τοποθεσία (LBSN). Τα δίκτυα αυτά εμφανίζουν γεωγραφικές πληροφορίες σε έναν χάρτη ή ως λίστα ενημερώσεων κατάστασης ταξινομημένες κατά γεωγραφική εγγύτητα σε αντίθεση με την παραδοσιακή έννοια της αντίστροφης χρονολογικής σειράς (Gordon and de Souza e Silva 2011). Αυτό το φαινόμενο έχει επίσης ονομαστεί Locative Mobile Social Networks. Είναι εμπορικές εφαρμογές, διαθέσιμες μέσω κινητού τηλεφώνου, που βοηθούν στη δημιουργία δικτύων κινητών κόμβων (σε αυτήν την περίπτωση κινούμενα αντικείμενα - χρήστες) εμφανίζοντας τη γεωγραφική θέση των χρηστών σε έναν χάρτη. Οι χρήστες έχουν τη δυνατότητα να εντοπίζουν ο ένας τον άλλον στον φυσικό χώρο και να αλληλεπιδρούν μεταξύ τους αναλόγως την σχετική απόσταση (de Souza e Silva και Frith 2010). Υπάρχουν δύο τρόποι διαμοιρασμού της γεωγραφικής τοποθεσίας στα LBSN. Ο πρώτος είναι το geotagging, το οποίο μετατρέπει τις φωτογραφίες, τα βίντεο, τις αναρτήσεις ή tweets σε γεωγραφικές πληροφορίες. Ο δεύτερος τρόπος είναι η κοινή χρήση των δραστηριοτήτων με την τρέχουσα τοποθεσία ή αλλιώς γεωκοινωνική δικτύωση. Για παράδειγμα στο Foursquare, ο χρήστης μπορεί να κάνει check-in σε συγκεκριμένη τοποθεσία και να μοιραστεί αυτήν την πληροφορία με τους φίλους του.

Η συνεχώς κοινοποιούμενη πληροφορία τοποθεσίας μπορεί να χρησιμοποιηθεί για την ανάλυση της συμπεριφοράς του ανθρώπου και την εξαγωγή συγκεντρωτικών ευρημάτων σχετικά με τα πρότυπα ανθρώπινης κινητικότητας και τη δομή των αστικών περιοχών. Μια έρευνα χωροχρονικών μοτίβων δραστηριοτήτων με δεδομένα από το Foursquare έδειξε ότι τα μοτίβα check-in που γίνονται τις καθημερινές ημέρες (Δευτέρα – Παρασκευή) διαφέρουν από εκείνα που πραγματοποιούνται το Σάββατο και την Κυριακή. Ακόμα, οι κυρίαρχες τοποθεσίες αλλάζουν κατά τη διάρκεια της ημέρας, τις πρωινές ώρες τα περισσότερα check-in συγκεντρώνονται σε χώρους με πολλή κίνηση ενώ τις απογευματινές ώρες η κυρίαρχη χωρική κατηγορία είναι το σπίτι. Οι κινήσεις ακολουθούν το μοτίβο πτήσης Lévy με πολλές διαδοχικές κινήσεις μικρής απόστασης που διακόπτονται από περιστασιακά ταξίδια μεγάλων αποστάσεων.

Τα LBSN προσφέρουν την δυνατότητα εύρεσης του Σημείου ενδιαφέροντος (Point of Interest – POI) που αντιστοιχεί στην τοποθεσία του χρήστη τη στιγμή που πραγματοποιήθηκε το check-in κάνοντας χρήση των συντεταγμένων. Ένας εύκολος τρόπος είναι να χρησιμοποιηθεί το OpenStreepMap. Το OSM είναι ένας χάρτης με ελεύθερη άδεια ο οποίος αναπτύσσεται από μια κοινότητα εθελοντών που συνεισφέρουν και διατηρούν δεδομένα σχετικά με δρόμους, μονοπάτια, καφετέριες, σιδηροδρομικούς σταθμούς, και πολλά περισσότερα, σε όλον τον κόσμο. Οι συνεισφέροντες χρησιμοποιούν αεροφωτογραφίες, συσκευές GPS, και τοπικούς χάρτες

χαμηλής τεχνολογίας για να σιγουρευτούν πως το OSM είναι ακριβής και ενημερωμένο στο μικρότερο δυνατό επίπεδο.

1.6. Twitter

Το Twitter είναι ένας ιστότοπος κοινωνικής δικτύωσης που επιτρέπει στους χρήστες του να στέλνουν και να διαβάζουν σύντομα μηνύματα (έως 280 χαρακτήρες) που ονομάζονται tweets. Οι μη συνδεδεμένοι χρήστες μπορούν επίσης να διαβάζουν μηνύματα, αλλά μόνο οι συνδεδεμένοι χρήστες μπορούν να δημοσιεύουν μηνύματα και να σχολιάζουν ή να κάνουν retweet.

Το Twitter είναι ένα LBSN το οποίο παρέχει έναν τεράστιο όγκο γεωγραφικών δεδομένων στα οποία μπορεί κανείς να έχει πρόσβαση μέσα από δημόσιο Application Programming Interface (API). Οι Gelernter και Mushegian (2011) εφάρμοσαν Named Entity Recognition για να βρουν πληροφορίες τοποθεσίας (χώρα, πόλη, οδός κλπ) για να εξάγουν αυτόματα μια αναφορά τοποθεσίας του αντίστοιχου tweet. Η μέθοδος εντόπισε σωστά μόλις το 34% των tweets κυρίως λόγω ορθογραφικών λαθών ή χρήσης συντομογραφιών. Οι Davis Jr. et al. (2011) ανέπτηξαν μια μέθοδο εύρεσης της τοποθεσίας του χρήστη από γνωστές τοποθεσίες των ακολούθων του χρήστη επιλέγοντας εκείνη την τοποθεσία που είχαν επισκεφτεί οι περισσότεροι.

Η παρακολούθηση της ροής των tweets επιτρέπει την εξαγωγή εκείνων που σχετίζονται με συγκεκριμένα συμβάντα. Για παράδειγμα, συγκρίσεις μεταξύ των μέσων ενημέρωσης για τον σεισμό στην πόλη Σετσουάν το 2008 και τα μηνύματα Twitter που σχετίζονταν με αυτόν έδειξαν ότι οι σχετικές αναφορές στο Twitter έγιναν μέσα σε δευτερόλεπτα και μεγάλος όγκος των tweets δημοσιεύτηκαν από χρήστες που βρίσκονταν κοντά στο συμβάν, ενώ η κάλυψη από τα μέσα ενημέρωσης ξεκινά από την επόμενη ημέρα. Ο σεισμός αναφέρθηκε ακόμη και πριν εντοπιστεί από το USGS (Unites States Geological Survey) (Li and Rao, 2010). Το φαινόμενο αυτό εκμεταλλεύτηκαν οι Earle et. al (2011) στην εργασία τους όπου ψάχνοντας την λέξη-κλειδί «σεισμός» στα tweets, προσπάθησαν να ανιχνεύσουν την εμφάνιση και το μέγεθος ενός σεισμού, συγκρίνοντας μακροπρόθεσμη συχνότητα λέξεων-κλειδιών με βραχυπρόθεσμη. Κατάφεραν να εντοπίσουν σημαντικό αριθμό σεισμών με την παρουσία ωστόσο πολλών λανθασμένα θετικών

συμβάντων τα οποία μπορούν να περιοριστούν, όπως υποστηρίζουν, με την χρήση αλγορίθμων επεξεργασίας φυσικής γλώσσας.

Οι χρήστες του Twitter μοιράζονται πολλές προσωπικές πληροφορίες, όπως είναι τα θέματα υγείας. Παρακολουθώντας την ροή των tweets και αναλύοντας το περιεχόμενό τους για συγκεκριμένες λέξεις-κλειδιά μπορεί κανείς τα εξάγει πληροφορίες σχετικά με τοπικές παθήσεις και την εξάπλωσή τους στον χώρο και χρόνο. Μελέτες έχουν δείξει την παρουσία συσχέτισης ανάμεσα στον αριθμό των tweets που σχετίζονται με την γρίπη με στατιστικές από επισήμους φορείς. Οι Collier et al.(2011) εντοπίζουν το ξέσπασμα γρίπης κατά παρόμοιο τρόπο με το Flu Trends της Google ενώ ο Culotta (2010) έδειξε ότι μια ανάλυση πολλαπλής παλινδρόμησης σε σχέση με τον αριθμό μηνυμάτων Twitter είναι κατάλληλη για τον εντοπισμό κρουσμάτων γρίπης.

1.7. Τροχιές

Σύμφωνα με τους Spaccapietra et. al (2008), οι τροχιές είναι μετρήσιμα ταξίδια που σχετίζονται με αντικείμενα τα οποία κινούνται μέσα στον χρόνο. Διαφορετικά είδη τροχιών μπορούν να εξαχθούν από ένα κινούμενο αντικείμενο. Αν, για παράδειγμα, μια εφαρμογή ενδιαφέρεται να μελετήσει την συμπεριφορά ενός ζώου, θα εξάγει τις τροχιές ομαδοποιώντας τις καθημερινές παρατηρήσεις ως τροχιά. Γνωρίζοντας την ιστορία της κίνησης αυτοκινήτων σε μία πόλη, μπορούν να εξαχθούν στοιχεία που θα βελτιώσουν την κυκλοφοριακή κίνηση και θα εντοπίσουν δρόμους με αυξημένη συμφόρηση. Παρομοίως, έχοντας καταγράψει την κίνηση ζώων σε ένα βιότοπο, μπορούν να εξαχθούν πληροφορίες για τις βιολογικές τους συνήθειες (λ.χ. διαδρομές που ακολουθούν κατά το κυνήγι) ή να εντοπιστούν περιπτώσεις ζώων που ξεφεύγουν από το βιότοπο. Η τροχιά ενός αντικειμένου συνιστά ακριβώς αυτό, την ιστορία της κίνησής του. Ζητούμενο είναι η καταγραφή της θέσεώς του στο χώρο με την πάροδο του χρόνου.

Οι κινητές συσκευές αφήνουν χωροχρονικά ίχνη τα οποία χαρακτηρίζουν την τροχιά ενός κινούμενου αντικειμένου. Τα δεδομένα τροχιάς κανονικά παράγονται ως σημεία, το οποίο δυσκολεύει την κατανόηση και ανάλυση τους καθώς πολύ συχνά δεν περιέχουν κάποια σημασιολογική πληροφορία.

Ως τροχιά T ορίζεται μια ακολουθία στοιχείων $< e_1,\ e_2, \dots, e_m >$, όπου το κάθε στοιχείο έχει ένα σύνολο l διαστάσεων $D = \{d_1, d_2, \dots, d_l\}$. (Ferrero et al. (2018))

- Ταξινόμηση τροχιών

  Έστω ένα σύνολο τροχιών με ένα σύνολο ζευγών σημείων

  $\mathbf{T} = \{(T_1, class_{T_1}), (T_2, class_{T_2}), \dots, (T_1, class_{T_n})\}$, όπου το κάθε ζευγάρι περιέχει μια τροχιά και την κλάση (ή τάξη), η ταξινόμηση τροχιών είναι το έργο της εκμάθησης μιας συνάρτησης f η οποία αντιστοιχίζει κάθε τροχιά T_i του $\mathbf{T}$ σε μια από τις τάξεις με όσο δυνατόν υψηλότερη ακρίβεια.

  Σε αντίθεση με τις συμβατικές μεθόδους ταξινόμησης, στην ταξινόμηση τροχιών το κύριο πρόβλημα είναι η εύρεση των βέλτιστων χαρακτηριστικών των τροχιών που θα τροφοδοτηθούν στον ταξινομητή, τα οποία χαρακτηριστικά δεν είναι απαραίτητο να προέρχονται από ολόκληρη την τροχιά αλλά και από τμήμα της (υπό-τροχιά)

ΣΧΗΜΑ 1 ΠΑΡΑΔΕΙΓΜΑ ΑΚΑΤΕΡΓΑΣΤΗΣ ΤΡΟΧΙΑΣ, ΠΗΓΗ: FERRERO ET AL (2020)

- Υπό-τροχιά

  Δεδομένης μιας τροχιάς T μήκους m, μια υπό-τροχιά $s = < e_a, \dots, e_b >$ είναι μία συνεχόμενη υπό-ακολουθία του T ποz su ξεκινάει με το στοιχείο e_a και τελειώνει με το στοιχείο e_b, όπου $1 \leq a \leq m$ και $a \leq b \leq m$. Η υπό-τροχιά s μπορεί να αναπαρασταθεί με όλες τις διαστάσεις του D ή με ένα υποσύνολο του D τέτοιο ώστε $D' \subseteq D$. Το μήκος της υπό-τροχιάς ορίζεται ως $w = |s|$.

Τα δεδομένα τροχιάς είναι περίπλοκα λόγω των πολλαπλών διαστάσεων τους. Μια ακατέργαστη τροχιά που παράγεται από μία συσκευή GPS (Global Positioning System) και αποτελεί την πιο απλή μορφή δεδομένων τροχιάς, είναι μια αλληλουχία χωρο-χρονικών σημείων με τη μορφή

(x, y, t), όπου το x και y αναπαριστούν την χωρική θέση του κινούμενου αντικειμένου την χρονική στιγμή t.

1.8. Ταξινόμηση Τροχιών

Η ανάπτυξη συσκευών παρακολούθησης οδήγησε στην συλλογή μεγάλου όγκου δεδομένων τροχιάς κινούμενων αντικειμένων, γεγονός το οποίο καθιστά την εξαγωγή χρήσιμής πληροφορίας επιτακτική. Η ταξινόμηση τροχιών είναι ένας αποτελεσματικός τρόπος ανάλυσης δεδομένων τροχιάς και έχει εφαρμοστεί στην αναγνώριση προτύπων, την ανάλυση δεδομένων, τη μηχανική μάθηση κ.λπ. Η ταξινόμηση των τροχιών χρησιμοποιείται και για τη λήψη πληροφορίας σχετική με τον χρόνο και τον χώρο από δεδομένα τροχιάς που σημαίνει ότι είναι παρούσα σε ορισμένα πεδία εφαρμογών όπως είναι η πρόβλεψη της κίνησης, η παρακολούθηση κυκλοφορίας, η κατανόηση δραστηριότητας, η ανίχνευση ανωμαλιών, η πρόγνωση καιρού και η γεωγραφία. Η άντληση δεδομένων κίνησης μπορεί να γίνει από διάφορες πηγές με ποικίλες μορφές. Για παράδειγμα τα δεδομένα που προέρχονται από το συσκευές GPS παράγουν μια τροχιά παρακολουθώντας την κίνησης των αντικειμένων. Δεδομένα τροχιάς μπορούν να αντληθούν και από δεδομένα εικόνας ή βίντεο όπου χρησιμοποιούνται πληροφορίες εικόνας και χωροχρονικές.

Ημι-εποπτευόμενες και εποπτευόμενες μέθοδοι ταξινόμησης τροχιών έχουν σκοπό την ανάθεση κάποιας ετικέτας στα δεδομένα δοκιμής(test set) από ένα προκαθορισμένο με ετικέτες σύνολο δεδομένων εκπαίδευσης (training set). Οι σημασιολογικές πληροφορίες που εμπλουτίζουν τις τροχιές μπορούν να μειώσουν τον όγκο του συνόλου δεδομένων εκπαίδευσης και να αυξήσουν την απόδοση του αλγορίθμου.

Η ταξινόμηση των τροχιών βρίσκει εφαρμογή σε πολλά πεδία. Ένα τέτοιο παράδειγμα είναι η ταξινόμηση της κινητικής συμπεριφοράς των ζώων μέσω της κατοικίας στο χώρο και τον χρόνο με την μέθοδο Residence in Space and Time (RTS) που προτάθηκε από τους Torres LG et al. (2017) όπου γίνεται ταξινόμηση των μοτίβων συμπεριφοράς των ζώων σε δεδομένα κίνησης έχοντας βασιστεί στην ιδέα ότι οι καταστάσεις συμπεριφοράς μπορούν να διαιρεθούν με την ποσότητα και του χώρου και χρόνου που καταλαμβάνεται σε μια περιοχή σταθερής κλίμακας. Χρησιμοποιώντας κανονικοποιημένες τιμές του χρόνου διαμονής και της απόστασης διαμονής σε μια σταθερή ακτίνα αναζήτησης, το RST είναι σε θέση να διαφοροποιεί μοτίβα συμπεριφοράς

που είναι χρονοβόρα (π.χ. ανάπαυση), έντασης χρόνου και απόστασης (π.χ. αναζήτηση περιορισμένης περιοχής) και διέλευσης (σύντομη χρόνος και απόσταση). Ένα δεύτερο παράδειγμα είναι η ταξινόμηση τροχιών πλοίων, με τους Tao Guo και Lei Xie (2022) να προτείνουν μια μέθοδο ομαδοποίησης τροχιών πλοίων για ύδατα με πυκνή και περίπλοκη ροή που λαμβάνει τα δεδομένα τροχιάς πλοίων διαφόρων συστάδων στην υδάτινη περιοχή. Στη συνέχεια χρησιμοποιώντας ως βάση αυτά τα ομαδοποιημένα δεδομένα μελετούν την ανίχνευση ανωμαλιών τροχιάς πλοίου μια ταξινόμηση καναλιού, έτσι ώστε να παρέχουμε υποστήριξη αποφάσεων για έξυπνη διαχείριση κινδύνου και έλεγχο των τμημάτων ελέγχου κυκλοφορίας πλοίων.

1.9. Σημασιολογικά Εμπλουτισμένες Τροχιές

Οι σημασιολογικά εμπλουτισμένες τροχιές είναι μια εναλλακτική αναπαράσταση μιας διαδρομής κίνησης ενός κινούμενου αντικειμένου. Μια σημασιολογικά εμπλουτισμένη τροχιά θα μπορούσε να οριστεί ως μια αλληλουχία επεισοδίων στάσεων (stop) και μετακινήσεων (move), με κάθε ένα από αυτά τα επεισόδια να είναι συνυφασμένο με μεταδεδομένα (ετικέτες-tags). Στάσεις είναι τα σημεία της τροχιάς του αντικειμένου όπου το αντικείμενο παραμένει στάσιμο σε ένα σημείο. Μετακινήσεις είναι τα τμήματα της τροχιάς μεταξύ δύο στάσεων όπου το αντικείμενο κινείται. Ετικέτες είναι τα μεταδεδομένα που σχετίζονται με τις στάσεις και τις μετακινήσεις. Στην εικόνα 2, παρουσιάζεται ένα παράδειγμα σημασιολογικά εμπλουτισμένης τροχιάς όπου ένα άτομο ξεκινάει από το σπίτι του, μετά πηγαίνει στη δουλειά με αυτοκίνητο και τέλος πηγαίνει για φαγητό. Κατά τη διάρκεια των διαδρομών του παρατηρούμε τις αλλαγές του καιρού (συννεφιά, ήλιος, βροχή) και τον τρόπο της μετακίνησης του από το ένα μέρος στο άλλο (αυτοκίνητο, περπάτημα). Επιπλέον, το κινούμενο αντικείμενο είναι και χρήστης κοινωνικών δικτύων και δημοσιεύει τις σκέψεις/συναισθήματά του σε αυτά. Όλες αυτές οι επιπλέον πληροφορίες, δηλαδή ο καιρός, ο τρόπος μετακίνησης, το μέρος επίσκεψης, βαθμολογία που έδωσε ο χρήστης στο μέρος κ.α. αποτελούν τις σημασιολογικές πληροφορίες που εμπλουτίζουν τις τροχιές. Λόγω της παρουσίας των πολλαπλών και ετερογενών διαστάσεων οι τροχιές αυτές αποκαλούνται και τροχιές πολλαπλών πτυχών (Ferrero et al. 2020).

ΣΧΗΜΑ 2 ΠΑΡΑΔΕΙΓΜΑ ΣΗΜΑΣΙΟΛΟΓΙΚΑ ΕΜΠΛΟΥΤΙΣΜΕΝΗΣ ΤΡΟΧΙΑΣ, ΠΗΓΗ: FERRERO ET AL(2020)

Η διαφορά ανάμεσα στις τροχιές πολλαπλών πτυχών και τις ακατέργαστες τροχιές είναι το πλήθος και είδος των διαστάσεων. Μια τροχιά πολλαπλών πτυχών είναι μια αλληλουχία στοιχείων $< e_1, e_2, \dots, e_m >$, όπου το καθένα από αυτά έχει διαστάσεις x, y, t, A όπου x και y αντιστοιχούν στη γεωγραφική θέση του αντικειμένου (συντεταγμένες) στη χρονική στιγμή t, και $A = \{a_1, a_2, \dots, a_k\}$ είναι ένα σύνολο σημασιολογικών διαστάσεων που είναι οποιοδήποτε είδος πληροφορίας που δεν είναι ούτε χωρική αλλά ούτε και χρονική.

1.10. Βιβλιογραφική Ανασκόπηση

Η ανάλυση τροχιών έχει χρησιμοποιηθεί με ποικίλους τρόπους και για διαφορετικούς σκοπούς. Στις περισσότερες περιπτώσεις, ο σκοπός ήταν η αναγνώριση ανωμαλιών και έκτροπων παρατηρήσεων, η πρόβλεψη μελλοντικών καταστάσεων των υπό εξέταση κινούμενων αντικειμένων, ταξινόμηση των αντικειμένων σε εκ των προτέρων ορισμένες τάξεις.

Οι περισσότερες από τις υπάρχουσες προσεγγίσεις βασίζονται σε ακατέργαστα δεδομένα κίνησης όπως είναι οι συντεταγμένες (γεωγραφικό πλάτος και μήκος) και ο χρόνος. Υπάρχουν όμως και πιο σύγχρονες προσεγγίσεις οι οποίες κάνουν χρήση και σημασιολογικά εμπλουτισμένων δεδομένων. Τα σημασιολογικά δεδομένα είναι οποιοδήποτε είδος πληροφορίας που δεν αναφέρεται σε χωροχρονικά δεδομένα. Η πρόκληση με αυτό τον τύπο δεδομένων έγκειται στην σωστή επιλογή των χαρακτηριστικών που θα μπορέσουν να πετύχουν την καλύτερη δυνατή ακρίβεια στο πρόβλημα της ταξινόμησης.

Στην εργασία τους οι Rossi Luca και Musolesi Mirco(2014) προτείνουν μια σειρά από τεχνικές για την ταυτοποίηση των χρηστών κοινωνικών δικτύων χρησιμοποιώντας τα check-in τους, ενώ ταυτόχρονα εγείρεται ο κίνδυνος διαρροής προσωπικών δεδομένων όχι μόνο των χρηστών που πραγματοποιούν το check-in αλλά και των φίλων τους που έχουν συμπεριληφθεί σε αυτό (tagging). Η πρώτη τεχνική βασίζεται τα χωροχρονικά δεδομένα GPS των χρηστών. Η δεύτερη βασίζεται στην συχνότητα που ένας χρήστης επισκέπτεται κάποια μέρη καθώς τα σημεία GPS είναι ομαδοποιημένα γύρω από συγκεκριμένες τοποθεσίες οπότε η ταυτοποίηση του μπορεί να γίνει βάσει της επισκεψιμότητας παρά με την χρήση των τροχιών των σημείων GPS. Εδώ προστίθεται και η παράμετρος της κοινωνικής εξομάλυνσης, η επίδραση δηλαδή που μπορούν να έχουν οι κοινωνικοί δεσμοί σε έναν χρήστη καθώς είναι πιο πιθανό να επισκεφτεί αυτός ο χρήστης μια τοποθεσία αν προηγουμένως την επισκέφτηκε κάποιος φίλος του. Αναπτύσσεται μία ακόμη μέθοδος ταυτοποίησης βασισμένη στο πολυωνυμικό Μπεϋζιανό μοντέλο κάνοντας την υπόθεση ότι κάθε check-in είναι ανεξάρτητο από τα υπόλοιπα, το οποίο μπορεί επίσης να εμπλουτιστεί με την χρονική παράμετρο δεδομένου ότι οι χρήστες τείνουν να κάνουν check-in σε παρόμοιες τοποθεσίες σε παρόμοιες χρονικές στιγμές. Προτείνεται και ένα τρίτο μοντέλο, το υβριδικό, όπου χωρικά δεδομένα και δεδομένα συχνότητας χρησιμοποιούνται μαζί σε ένα μοντέλο. Οι Li et al (2007), χρησιμοποιήσαν έναν ταξινομητή βασισμένο σε κανόνες για να κάνουν ιεραρχική ταξινόμηση χαρακτηριστικών. Στην εργασία τους παρουσίασαν μια τεχνική για την ανίχνευση ανωμαλιών κινούμενων αντικειμένων, που ονομάζεται ROAM (Rule and Motif-Based Anomaly Detection in Massive Moving Object Data Sets). Οι τροχιές εκφράζονται ως ένα σύνολο διακριτών θραυσμάτων που ονομάζονται μοτίβα, τα

οποία σχηματίζουν έναν πολυδιάστατο χώρο χαρακτηριστικών για κάθε δείγμα. Ένας ταξινομητής που βασίζεται σε κανόνες, στη συνέχεια αναλαμβάνει την ιεραρχική εξερεύνηση του χώρου των χαρακτηριστικών για να βρει τις αποτελεσματικές περιοχές που ορίζουν μια ανωμαλία. Παρομοίως, οι Lee et al (2008) παρουσίασαν έναν αλγόριθμο που ανακαλύπτει τις έκτροπες τιμές και βασίζεται στην κατάτμηση της τροχιάς. Η ιδέα είναι παρόμοια με την προαναφερθείσα εργασία, με την διαφορά ότι σε αυτήν την περίπτωση, ο στόχος ήταν να ανακαλυφθούν και οι έκτροπες υπό-τροχιές. Εδώ, η ανακάλυψη γίνεται με μια υβριδική προσέγγιση βασισμένη σε απόσταση και πυκνότητα. Ιδιαίτερο ενδιαφέρον παρουσιάζει και η μέθοδος Traclass που προτάθηκε από τους ίδιους συγγραφείς, η οποία αποτελεί μια μέθοδο δημιουργίας ιεραρχίας χαρακτηριστικών μέσω του διαχωρισμού των τροχιών και διερευνά δύο είδη ομαδοποίησης: (1) βάσει της περιοχής και (2) βάσει της τροχιάς. Η ομαδοποίηση που βασίζεται στην περιοχή συλλαμβάνει υψηλότερου επιπέδου χαρακτηριστικά δίχως τη χρήση μοτίβων κίνησης ενώ αυτή που βασίζεται στην τροχιά συλλαμβάνει χαρακτηριστικά χαμηλότερου επιπέδου χρησιμοποιώντας μοτίβα κίνησης. Σε αντίθεση με προηγούμενες μεθόδους, η TraClass ξεπερνά το πρόβλημα της παρουσίας διακριτών χαρακτηριστικών σε σημεία των τροχιών ή που δεν σχετίζονται με τα σχήματα των τροχιών, καθώς με την κατάτμηση των τροχιών καθιστά τα τμήματα τους αναγνωρίσιμα. Μια ακόμα μέθοδος ομαδοποίησης είναι αυτή των Holst και Jonasson(2014) σχετικά με την ταξινόμηση κινήσεων σκι. Στην προσέγγιση αυτή, τα δεδομένα μοντελοποιούνται κάνοντας χρήση μιας αλυσίδας Markov πολυμεταβλητών κανονικών κατανομών πάνω στα οποία γίνεται η ταξινόμηση. Σκοπός ήταν να ταξινομηθούν και να ανιχνευτούν διάφορες τεχνικές σκι που ονομάζονται gears. Σε αυτήν την περίπτωση τα δεδομένα που αναλύθηκαν προέρχονται από ένα επιταχυνσιόμετρο που είναι τοποθετημένο στον αθλητή. Οι Petry et al.(2020) προτείνουν την μέθοδο MARC για την ταξινόμηση σημασιολογικά εμπλουτισμένων τροχιών πολλαπλών πτυχών μέσω του χώρου, του χρόνου και των σημασιολογικών embeddings. Η μέθοδος είναι μια προσέγγιση που βασίζεται στα Αναδρομικά Νευρωνικά Δίκτυα (RNN) για την ταξινόμηση των τροχιών πολλαπλών πτυχών μέσω ενός πολυμεταβλητού embedding layer το οποίο επιτρέπει την κωδικοποίηση ετερογενών διαστάσεων που σχετίζονται με κάθε σημείο της τροχιάς, δεδομένης της αραιότητας και ετερογένειας των δεδομένων. Η μέθοδος MARC δέχεται ως είσοδο μια τροχιά και δίνει ως έξοδο την αντίστοιχη κλάση. Επειδή οι τροχιές πολλαπλών πτυχών έχουν πολλές διαστάσεις, χρησιμοποιείται ένα embedding layer πολλαπλών πτυχών για την κωδικοποίησή τους. Μετά την κωδικοποίηση των σημείων των τροχιών, αυτές δίνονται ως δεδομένα εισόδου στα LSTM Αναδρομικά Νευρωνικά Δίκτυα (RNN) . Τα LSTM διασφαλίζουν την αποθήκευση και πρόσβαση στις πληροφορίες ακόμα και με την πάροδο μεγάλων χρονικών περιόδων ή πολλών βημάτων. Αυτό

σημαίνει ότι τα LSTM μπορούν μοντελάρουν τις σχέσεις ανάμεσα σε διαφορετικά σημεία της τροχιάς και των χαρακτηριστικών τους, ακόμα και αν υπάρχει μεγάλη απόσταση μεταξύ τους. Αναδρομικά Νευρωνικά Δίκτυα για την ταξινόμηση τροχιών χρηστών κοινωνικών δικτύων χρησιμοποιούν και οι Gao, Qiang et al (2017). Η μέθοδος που προτείνουν ονομάζεται TULER και πρόκειται για ένα ημι-εποπτευόμενο μοντέλο μάθησης που βασίζεται σε αναδρομικά νευρωνικά δίκτυα (RNN), το οποίο εκμεταλλεύεται τα χωροχρονικά δεδομένα για να συλλάβει την υποκείμενη σημασιολογία των μοντέλων κινητικότητας των χρηστών. Διαθέτοντας ένα σύνολο τροχιών χωρίς τους χρήστες που τις παρήγαγαν, το TULER αρχικά χωρίζει τις τροχιές σε υπό-τροχιές, οι οποίες στη συνέχεια κωδικοποιούνται χρησιμοποιώντας trajectory embedding λόγω της κατανομής νόμου δύναμης, που όπως παρατηρήθηκε, ακολουθούν τα check-in που πραγματοποιούνται στις διάφορες τοποθεσίες. Τέλος, αφού λύσει το πρόβλημα ύπαρξης πυκνών check-in ακόμα και μετά τον διαχωρισμό σε υπο-τροχιές , χρησιμοποιώντας ορισμένα μοντέλα RNN, δηλαδή τα stacked LSTM, GRU και Bidirectional RNN , ταξινομεί τις τροχιές. Παρομοίως και η μέθοδος DeepeST των Freitas, Nicksson et al. (2021) χρησιμοποιώντας Αναδρομικά Νευρωνικά Δίκτυα προσπαθεί να κάνει ταξινόμηση τροχιών. Η DeepeST είναι μια παραμετρική μέθοδος ταξινόμησης τροχιών, η οποία βρίσκει την κατηγορία από ένα μεγάλο πλήθος υπό-τροχιών προερχόμενες από υπηρεσίες GPS και δεδομένα check-in. Η κατηγορία μπορεί να είναι κάποιο μεταφορικό μέσο, εγκληματική δραστηριότητα ή και ο χρήστης στον οποίο ανήκει η τροχιά. Η μέθοδος χρησιμοποιεί RNN LSTM λόγω του διαφορετικού μήκους των ακολουθιών που χρησιμοποιούνται ως δεδομένα εισόδου και καθώς μπορούν να διαχειριστούν μακροπρόθεσμες εξαρτήσεις σε μια ακολουθία. Τα RNN BLSTM (Bi-directional LSTM) χρησιμοποιούνται καθώς διαχειρίζονται απεριόριστο πλήθος γενικού πλαισίου και από τις δύο κατευθύνσεις μιας υπό-τροχιάς και δεν έχουν το πρόβλημα του περιορισμένου γενικού πλαισίου που αντιμετωπίζουν τα feed-forward μοντέλα. Ο στόχος αυτής της μεθόδου είναι διττός : (1) Trajectory-User Linking (TUL), δηλαδή αντιστοίχιση των χρηστών στις τροχιές τους και (2) αναγνώριση εγκληματικών μοτίβων για τη σύνδεση εγκληματικών δραστηριοτήτων με τις υπό-τροχιές που προέρχονται από το GPS που είναι ανεξάρτητες από τις τροχιές των χρηστών. Η DeepeST καταφέρνει να ελαχιστοποιήσει την πολυπλοκότητα του υπολογισμού καθώς λειτουργεί με λίγες διαστάσεις ώστε να βρει την υποκείμενη κατηγορία από τα δεδομένα των υπό-τροχιών. Η ανάλυση τροχιών οχημάτων και κίνησης είναι επίσης ένα κομμάτι της ανάλυσης δεδομένων κίνησης που έχει πολύ ενδιαφέρον ειδικά στις περιπτώσεις που εφαρμόζονται στην πραγματική ζωή. Οι Kumaran et al. (2021) παρουσίασαν την εργασία τους που είχε σχέση με τον εντοπισμό ανωμαλιών και την ταξινόμηση κίνησης πάνω σε βίντεο παρακολούθησης της κυκλοφορίας. Ορίζουν μια τροχιά ως δεδομένα χρονοσειράς με θέσεις αντικειμένων που είναι

ευρετηριασμένα με χρονική σειρά. Οι τροχιές μπορεί να είναι ποικίλου μήκους, γεγονός που καθιστά την ταξινόμηση τους πρόκληση καθώς ταξινομητές που βασίζονται σε νευρωνικά δίκτυα δέχονται ως είσοδο δεδομένα σταθερού μήκους. Ένας τρόπος να επιλυθεί το πρόβλημα είναι να μετατραπούν τα δεδομένα σε τροχιές σταθερού μήκους είτε με συμπλήρωση είτε με δειγματοληψία που μοιάζει με την κβαντοποίηση η οποία οδηγεί σε απώλεια πληροφορίας. Αρχικά εισάγουν μια υψηλού επιπέδου αναπαράσταση των τροχιών αντικειμένων χρησιμοποιώντας μια μορφή βασισμένη σε κλίση χρώματος. Στο επόμενο στάδιο για την δημιουργία ετικετών (labels) στην ταξινόμηση χρησιμοποιήθηκε ένας ημί-εποπτευόμενος τρόπος για να εντοπίσει τις τροχιές των κινούμενων αντικειμένων που εξήχθησαν με τη βοήθεια της Temporal Uknown Incremental Ομαδοποίησης. Οι ανώμαλες τροχιές διαχωρίστηκαν χρησιμοποιώντας την στοχαστική εμφύτευση γειτόνων (Distributed Stochastic Neighbor Embedding). Τέλος, εφαρμόζοντας ένα υβρίδιο CNN και Variational Autoencoder μπόρεσαν να ανιχνεύσουν ακραίες τιμές και να ταξινομήσουν τις τροχιές με υψηλή ακρίβεια. Δεδομένα κίνησης οχημάτων χρησιμοποιήσαν και οι Khosroshahi et al.(2016). Προτείνουν μια μέθοδο για ταξινόμηση συμπεριφοράς των οχημάτων χρησιμοποιώντας τρισδιάστατα σημάδια τροχιάς και ένα μοντέλο Μακροπρόθεσμης Μνήμης (LSTM). Ως μελέτη περίπτωσης επιχείρησαν να ταξινομήσουν τους ελιγμούς των γύρω οχημάτων σε διασταυρώσεις τεσσάρων κατευθύνσεων. Οι μετρήσεις LIDAR, GPS, IMU χρησιμοποιούνται για την εξαγωγή των περιβάλλοντων τροχιών με αντιστάθμιση ego-motion από κλιπ δεδομένων στο σημείο αναφοράς KITTI.

Κεφάλαιο 2

MasterMovelets

2.1. Εισαγωγή

Οι Ferrero, C.A., Petry, L.M., Alvares, L.O. et a.l. (2020) προτείνουν την μέθοδο MasterMovelets, μια χωρίς παραμέτρους μέθοδο για την ανακάλυψη των πιο σχετικών υποτροχιών με διαφορετικές και ετερογενείς διαστάσεις και ποικίλου μήκους για προβλήματα ταξινόμησης τροχιών. Ως ταξινόμηση τροχιών, ορίζεται η τεχνική της εξόρυξης δεδομένων για την πρόβλεψη των κλάσεων κινούμενων αντικειμένων, βάσει των τροχιών τους. Εδώ γίνεται χρήση ενός νέου είδους τροχιάς που ονομάζεται τροχιά πολλαπλών πτυχών (Multiple Aspect Trajectory), η οποία σε σχέση με την απλή τροχιά, που ορίζεται ως μια ακολουθία στοιχείων, όπου το κάθε στοιχείο έχει τις διαστάσεις του χώρου (x, y) και του χρόνου (t) , κάνει χρήση και των σημασιολογικών διαστάσεων (A) που επίσης ονομάζονται πτυχές ή χαρακτηριστικά. Σημασιολογικές διαστάσεις είναι οποιοδήποτε είδος πληροφορίας που δεν αναφέρεται στον χώρο ή τον χρόνο. Τέτοια πληροφορία μπορεί να είναι το μέρος που επισκέφτηκε κάποιος, το είδος του μέρους, η τιμή που πλήρωσε, η βαθμολογία που έβαλε καθώς και ο καιρός τη δεδομένη εκείνη στιγμή. Το πλήθος των ετερογενών διαστάσεων που έχουν οι τροχιές πολλαπλών πτυχών αυξάνουν την δυσκολία της ταξινόμησης των τροχιών καθώς (α) δε μπορεί να γίνει η αναπαράσταση των διαστάσεων των δεδομένων με μία μόνο τιμή λόγω της φύσης της κάθε διάστασης και επειδή απαιτούνται διαφορετικές συναρτήσεις απόστασης για να γίνει η σύγκριση των μεταβλητών/χαρακτηριστικών και (β) κάθε τροχιά πρέπει να χωριστεί σε υποτροχιές προκειμένου να βρεθούν μοτίβα, επειδή μια ολόκληρη τροχιά ειναι πιθανό να μη μπορεί να διαφοροποιηθεί από κάποια άλλη ως προς την κλάση (Lee et al. 2008).

Σε πολλές εφαρμογές οι σχετικές υποτροχιές μπορούν να διαφοροποιήσουν την κλάση καλύτερα απ' ότι οι χωρο-χρονικές μεταβλητές αλλά το πρόβλημα έγγειται στο πως θα ανακαλυφθούν αυτές οι υποτροχιές και ποιές είναι οι πιο κατάλληλες διαστάσεις και οι συνδυασμοί τους (Ferrero et al. 2018).

Η μέθοδος MOVELETS, που επίσης κάνει χρήση πολλαπλών διαστάσεων (x, y, t, A) και προτάθηκε απο τους Ferrero et al. (2018), ανακαλύπτει τις σχετικές υποτροχιές χωρίς όμως τη χρήση κάποιου splitting criterion και ξεπέρασε σε ακρίβεια όλες τις προηγούμενες μεθόδους που έκαναν χρήση μόνο των διαστάσεων του χρόνου και του χώρου. Το πρόβλημα αυτής της μεθόδου

είναι ότι οι διαστάσεις εκπροσωπούνται από μια μόνο μεταβλητή, το οποίο σημαίνει ότι κάποιες διαστάσεις που ίσως μπορούσαν να διαφοροποιήσουν την κλάση δεν λαμβάνονται υπόψιν.

Η μέθοδος MasterMovelets αναπτύχθηκε συγκεκριμένα για τροχιές πολλαπλών πτυχών. Η μέθοδος αυτή εξάγει $movelets$, υποτροχιές δηλαδή που διαφοροποιούν καλύτερα κάθε κλάση και χρησιμοποιούνται ως δεδομένα εισόδου στους ταξινομητές. Αναζητά όλες τις πιθανές υποτροχιές (ή υποακολουθίες) οποιουδήποτε μεγέθους και όλους τους πιθανούς συνδυασμούς διαστάσεων, ενώ ψάχνει τις βέλτιστες που μπορούν να αντιπροσωπεύσουν κάθε κλάση

2.2. Μετρικές Αποστάσεων

Στην εργασία τους οι Ferrero et al, κάνουν χρήση μετρικών αποστάσεων με σκοπό την εύρεση διακριτών υποτροχιών. Για την καλύτερη κατανόηση του αλγορίθμου MasterMovelets οι ορισμοί των αποστάσεων αυτών παρατίθενται παρακάτω:

- ***Διάνυσμα απόστασης ανάμεσα σε δύο πολυδιάστατα στοιχεία***
  Έστω δύο στοιχεία e_i και e_j με d διαστάσεις. Η απόσταση των στοιχείων αυτών $dist(e_i, e_j)$ επιστρέφει ένα διάνυσμα απόστασης $V = (v_1, v_2, \dots, v_d)$, όπου κάθε $v_k = dist\_e_k(e_i, e_j)$ είναι η απόσταση ανάμεσα σε δύο σημεία στην διάσταση k, ενώ ισχύει η ιδιότητα της συμμετρίας $dist_{e_k(e_i,e_j)} = dist\_e_k(e_j, e_i)$.

Η παραπάνω απόσταση είναι διαφορετική στην εργασία MOVELETS (Ferrero et. al 2018) όπου η συνάρτηση $dist(e_i, e_j)$ επέστρεφε μία τιμή, χάνοντας με αυτόν τον τρόπο τις πληροφορίες για την απόσταση σε όλες τις διαστάσεις.

- ***Διάνυσμα απόστασης ανάμεσα σε δύο υποτροχιές ίσου μήκους***
  Έστω δύο υποτροχιές s και r μήκους w και διαστάσεων d, η συνάρτηση $dist\_s(s, r)$ υπολογίζει την απόσταση κατά ζευγάρια ανάμεσα στα στοιχεία της υποτροχιάς και ένα διάνυσμα απόστασης $V = (v_1, v_2, \dots, v_d)$, όπου κάθε v_k είναι η τιμή της απόστασης ανάμεσα στις υποτροχιές s και r στην διάσταση k, η οποία υπολογίστηκε με μία συνάρτηση πάνω στις w αποστάσεις μεταξύ των δύο υποτροχιών στην διάσταση k. Για κάθε απόσταση v_k ισχύει η ιδιότητα της συμμετρίας $dist\_sk\ (s, r) = dist\_sk\ (r, s)$.

Ένα πολύ σημαντικό μέρος της μεθόδου MasterMovelets είναι η εύρεση εκείνης της υποτροχιάς που είναι παρόμοια με μια άλλη υποτροχιά. Η υποτροχιά που μοιάζει περισσότερο σε μια υποτροχιά s μιας τροχιάς T ονομάζεται βέλτιστη ευθυγράμμιση και αποτελεί την υποτροχιά r της τροχιάς T με την ελάχιστη απόσταση από την s. (Ferrero et al, 2018)

- ***Διάνυσμα απόστασης ανάμεσα σε μια τροχιά και μια υποτροχιά***
  Έστω μία τροχιά $\boldsymbol{T}$ και μια υποτροχιά $\boldsymbol{s}$ μήκους $\boldsymbol{w} = |\boldsymbol{s}|$, η απόσταση ανάμεσα τους είναι η βέλτιστη ευθυγράμμιση της $\boldsymbol{s}$ στην $\boldsymbol{T}$, που ορίζεται από $W_T^s = min_{r \in S_T^w}(dist\_s(s,r))$, όπου S_T^w είναι το σύνολο όλων των υποτροχιών μήκους w στην T, και το min () δίνει το διάνυσμα απόστασης της βέλτιστης ευθυγράμμισης ανάμεσα στην s και όλων των υποτροχιών S_T^w.

2.3. Ο Αλγόριθμος

Ο αλγόριθμος Mastermovelets ανακαλύπτει τα ετερογενή $movelets$. Σε πρώτη φάση ο αλγόριθμος, έχοντας ως δεδομένα εισόδου το σύνολο των τροχιών του training set, εξερευνά κάθε τροχιά T. Η συνάρτηση $ComputeElementDistanceVectors()$ υπολογίζει τις αποστάσεις ανάμεσα σε όλα τα στοιχεία των τροχιών Τ και ανάμεσα σε όλες τις τροχιές και τις αποθηκεύει σε μια συστοιχία (array) τεσσάρων διαστάσεων, A_1. Κάθε τιμή $A_1[i,j,d,k]$ είναι η απόσταση ανάμεσα στο στοίχειο j της τροχιάς T και του σημείου k λαμβάνοντας υπόψιν την διάστασης d.

Στη συνέχεια γίνεται εξερεύνηση των μηκών όλων των υποτροχιών. Για το μήκος w μιας υποτροχιάς, η συνάρτηση $ComputeSubtrajectoryDistanceVectors()$ υπολογίζει την απόσταση άναμεσα σε όλες τις υποτροχιές που ανήκουν στην ίδια τροχιά του training set προσθέτοντας τις τιμές των A_{W1}και A_1 οι οποίες αποθηκεύονται στην μεταβλητή A_w.

Στον βρόγχο που ακολουθεί, για κάθε υποτροχιά μήκους w μιας τροχιάς T, ο αλγόριθμος χρησιμοποιεί την μεταβλητή Aw για να ανακαλύψει τον καλύτερο συνδυασμό διαστάσεων και το προσθέτει στο σύνολο των υποψήφιων $movelets$. Συγκεκριμένα, αφού υπολογιστεί η απόσταση κατάταξης R για κάθε υποτροχιά μεταξύ όλων των τροχιών στην διάσταση k, ο αλγόριθμος εξερευνά κάθε συνδυασμό διαστάσεων C. Σε αυτόν τον βρόγχο βρίσκει το διάνυσμα αποστάσεων της καλύτερης ευθυγράμμισης (best alignment) ανάμεσα σε κάθε υποτροχιά της τροχιάς T σε κάθε τροχιά T_i κάνοντας χρήση της μεθόδου Master Alignment και αποθηκεύει το διάνυσμα απόστασης στο W. Αφού υπολογιστούν τα διανύσματα απόστασης, μετράται η συνάφεια κάθε υποτροχιάς, βάσει αυτών των διανυσμάτων, με τη χρήση της συνάρτησης Master Relevance. Η εύρεση του συνδυασμού διαστάσεων με το υψηλότερο σκορ συνάφειας, επιτρέπει στον αλγόριθμο να διατηρήσει τα σημεία διαχωρισμού (split points), τα διανύσματα αποστάσεων και τον συνδυασμό διαστάσεων. Αφού η υποτροχιά με τον πιο συναφή συνδυασμό διαστάσεων γίνει υποψήφια τροχιά, αποθηκεύεται

στο σύνολο των candidates. Στη συνέχεια, στον εξωτερικό βρόγχο, αφαιρούνται οι υποψήφιες τροχιές που επικαλύπτονται σε ένα τουλάχιστον στοιχείο τους από κάποια άλλη τροχιά και έχουν χαμηλότερο σκορ συνάφειας. Τέλος, προστίθενται οι εναπομείνουσες υποψήφιες τροχιές στο σύνολο των movelets.

Algorithm 1: MASTERMOVELETS

```
Input  : T // trajectory training set
Output: movelets // set of relevant subtrajectories
movelets ← ∅ ;
for each trajectory T in T do
    candidates ← ∅;
    A_1 ← ComputeElementDistanceVectors (T, T) ;
    for subtrajectory length w from 1 to T.length do
        if w > 1 then
            A_w ← ComputeSubtrajectoryDistanceVectors (T, T, A_{w−1}, A_1, w) ;
        end
        for position j from 1 to (T.length − w + 1) do
            R ← ∅ ;
            for trajectory i from 1 to |T| do
                for dimension d from 1 to |D| do
                    R[i, d, ..] ← Rank(A_w[i, j, d, ..]) ;
                end
            end
            bestScore ← 0 ;
            for each dimension combination C in C*_d do
                W ← ∅;
                for trajectory i from 1 to |T| do
                    W_i ← min MASTERALIGNMENT(R[i, C, ..], A_w[i, j, C, ..]) ;
                    W ← W ∪ (W_i, T[i].class) ;
                end
                relevance ← assess MASTERRELEVANCE(W, T.class) ;
                if relevance.score > bestScore then
                    bestScore ← relevance.score ;
                    bestSp ← relevance.sp ;
                    bestW ← W ;
                    bestC ← C ;
                end
            end
            M ← MoveletCandidate(T, j, w, bestC, bestW, bestScore, bestSp) ;
            candidates ← candidates ∪ M ;
        end
    end
    SortByRelevance (candidates) ;
    RemoveSelfSimilar (candidates) ;
    movelets ← movelets ∪ candidates ;
end
return movelets
```

ΣΧΗΜΑ 3 ΑΛΓΟΡΙΘΜΟΣ MASTERMOVELETS, ΠΗΓΗ: FERRERO ET AL (2020)

2.4. Master Alignment

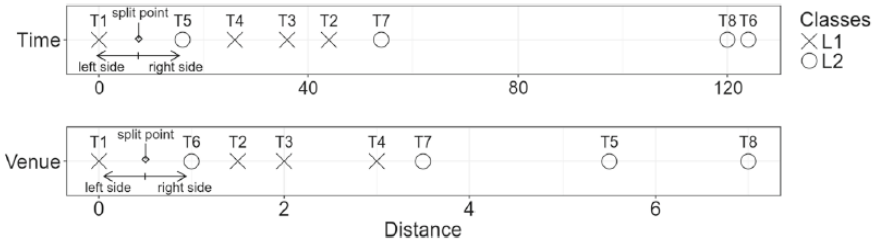
Η διαδικασία εντοπισμού του τμήματος της τροχιάς που μοιάζει περισσότερο σε μια δεδομένη υποτροχιά με την παρουσία πολλαπλών και διαφορετικών διαστάσεων ονομάζεται MasterAlignment.

Ο αλγόριθμος τροφοδοτείται με την παράμετρο V που περιλαμβάνει τις αποστάσεις των ευθυγραμμίσεων των τροχιών και την R που έχει τις κατατάξεις των ευθυγραμμίσεων των τροχιών

και δίνει ως αποτέλεσμα ένα σύνολο W που περιλαμβάνει τις αποστάσεις των καλύτερων ευθυγραμμίσεων των τροχιών. Δεδομένων των μεταβλητών Y (μέση κατάταξη), l (αριθμός διαστάσεων) και $posMinAvgRank$ (αρχική θέση ελάχιστης μέσης κατάταξης) ο αλγόριθμος υπολογίζει την μέση κατάταξη για κάθε διάσταση. Στον βρόγχο, αθροίζονται οι τιμές κατάταξης κατά μήκος των διαστάσεων και μετά υπολογίζεται μέση τιμή αυτού του αθροίσματος για το σύνολο των διαστάσεων η οποία αποθηκεύεται στο σύνολο Y. Η παραπάνω μέση τιμή συγκρίνεται με την ελάχιστη που έχει βρεθεί και γίνεται η ανάθεση της πιο μικρής στην μεταβλητή $posMinAvgRank$. Τέλος, βάσει αύτης της θέση, το διάνυσμα απόστασης της καλύτερης πολυδιάστασης ευθυγράμμισης αποθηκεύεται στο διάνυσμα W.

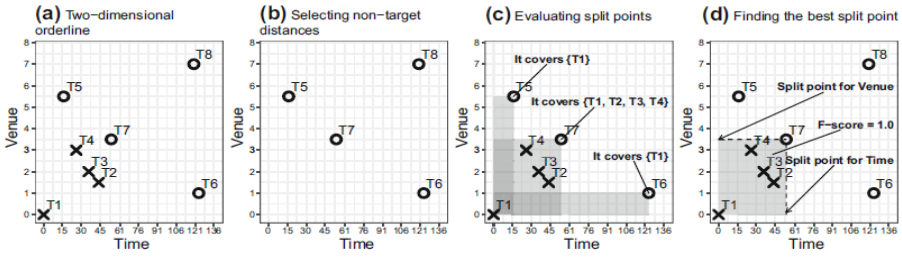
2.5. Master Relevance

Η συνάφεια μιας υποτροχιάς σχετίζεται με τον αριθμό των τροχιών της ίδιας τάξης που εκτελούν παρόμοια κίνηση. Ενώ προηγούμενες προσεγγίσεις έκαναν εκτίμηση ενός split point για κάθε διάσταση ξεχωριστά, σε αυτήν τη προσέγγιση τα split points που υπολογίζονται μεγιστοποιούν την συνάφεια του υποψήφιου movelet, λαμβάνοντας υπόψιν όλες τις διαστάσεις. Η μέθοδος αυτή ονομάζεται MasterRelevance. Έστω οι τροχιές $T_1, T_2, \ldots, T_8$ με δύο κλάσεις L_1, L_2 που αντιπροσωπεύονται από τις διαστάσεις χρόνος (Time) και τόπος (Venue) όπου οι τροχιές



ΣΧΗΜΑ 4 ΓΡΑΜΜΗ ΠΑΡΑΓΓΕΛΙΑΣ ΓΙΑ ΤΙΣ ΔΙΑΣΤΑΣΕΙΣ ΤΟΥ ΧΡΟΝΟΥ ΚΑΙ ΤΟΥ ΤΟΠΟΥ, ΠΗΓΗ: FERRERO ET AL (2020)

T_1, T_2, T_3, T_4 ανήκουν στην τάξη L_1 ενώ οι T_5, T_6, T_7, T_8 στην τάξη L_2 ένα υποψήφιο Movelet M που προήλθε από την T_1. Το παραπάνω σχήμα (Σχήμα 4) δείχνει τις γραμμές παραγγελίας όπου το κάθε σημείο ορίζει την τιμή της απόστασης για την ι-οστή τροχιά για κάθε διάσταση ενώ τα σύμβολα X και O είναι οι τάξεις.



ΣΧΗΜΑ 5 ΠΑΡΑΔΕΙΓΜΑ ΕΥΡΕΣΗΣ ΣΗΜΕΙΟΥ ΔΙΑΧΩΡΙΣΜΟΥ ΣΕ ΜΙΑ ΠΟΛΥΔΙΑΣΤΑΤΗ ΓΡΑΜΜΗ ΠΑΡΑΓΓΕΛΙΑΣ, ΠΗΓΗ: FERRERO ET AL., 2020

Το Σχήμα $5a$ δείχνει τις τιμές αποστάσεων του Σχήματος 4 με τη χρήση διαγραμμάτων διασποράς όπου κάθε σημείο αναπαριστά τις αποστάσεις και στις δύο διαστάσεις. Το πρώτο βήμα είναι να επιλεγούν μόνο τα σημεία που ανήκουν στην τάξη L_2 και στη συνέχεια να διαγραφτούν τα σημεία με μεγαλύτερες τιμές αποστάσεων από άλλα σημεία και στις δύο διαστάσεις. Στο Σχήμα $5b$ το σημείο T_8 διαγράφτηκε επειδή έχει υψηλότερη τιμή από από το σημείο T_7 . Το Σχήμα $5c$ δείχνει ότι χρησιμοποιώντας τα σημεία T_5, T_6 ως σημεία διαχωρισμού θα επικαλυφθεί μόνο το T_1 ενώ το σημείο T_7 μπορεί να επικαλύψει τα T_1, T_2, T_3, T_4. Στο τελευταίο βήμα πραγματοποιείται η επιλογή των σημείων διαχωρισμού που έχουν το υψηλότερο σκορ συνάφειας (relevance score) το οποίο υπολογίζεται με την μετρική F-measure που αποτελεί τον αρμονικό μέσο της ακρίβειας και της ανάκλησης. Εδώ, η ακρίβεια είναι η αναλογία των σημείων που καλύπτονται από τα σημεία διαχωρισμού που ανήκουν στην τάξη-στόχο ενώ η ανάκληση είναι η αναλογία των σημείων σημείων που καλύπτονται από τα σημεία διαχωρισμού και ανήκουν στην τάξη-στόχο σε σχέση με ολα τα σημεία της ταξης αυτής.

Η μέθοδος MasterMovelets παράγει υποτροχιές όλων των μεγεθών (ένα , δύο , τρία σημεία) και συνδυασμούς διαστάσεων, αναζητώντας τους καλύτερους υποψήφιους $movelets$. Αυτό σημαίνει ότι για κάθε υποτροχιά υπάρχει μεγάλο πλήθος συνδυασμών διαστάσεων. Σε μια τροχιά, δηλαδή, με 8 σημεία και 4 διαστάσεις, θα παραχθούν 36 υποτροχιές, ενώ με 15 πιθανούς συνδυασμούς διαστάσεων το αποτέλεσμα θα είναι 540 υποψήφια $movelets$. Το πρόβλημα έγγειται στο ότι η αύξηση των διαστάσεων οδηγεί στην εκθετική αύξηση των υποψηφίων. Ωστόσο, το MasterMovelets-Log μπορεί να περιορίσει τις υποτροχιές στον φυσικό λογάριθμο του μεγέθους του χωρίς σημαντική απώλεια ακρίβειας καθώς τα $movelets$ που χαρακτηρίζουν καλύτερα μια κλαση είναι αυτά με τα λιγότερα

στοιχεία στις υποτροχιές τους. Ο περιορισμός του μεγέθους της τροχιάς μειώνει αισθητά τον χρόνο που απαιτείται για την εύρεση των $movelets$.

Η μέθοδος εφαρμόστηκε σε δεδομένα προερχόμενα από τα LSBN Gowalla, Brightkite, Foursquare ενώ τα μοντέλα ταξινόμησης δημιουργήθηκαν με τη χρήση Νευρωνικών Δικτύων και Random Forest. Σε όλες τις περιπτώσεις, συγκριτικά με άλλες υπάρχουσες προσεγγίσεις, η μέθοδος MasterMovelets παρουσίασε την υψηλότερη ακρίβεια ενώ σε σχέση με τα RF μοντέλα, τα NN έχουν καλύτερη απόδοση καθώς διαχειρίζονται καλύτερα πολυδιάστατους χώρους.

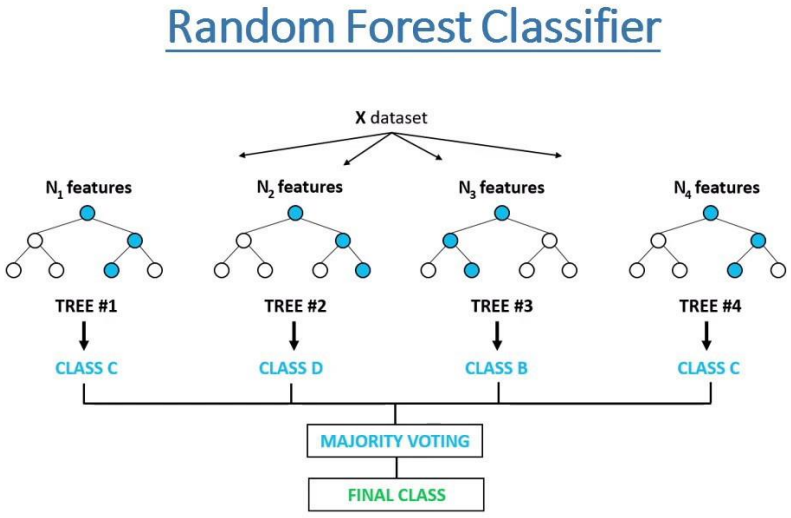
Κεφάλαιο 3

Μηχανική Μάθηση

Στο παρόν κεφάλαιο θα γίνει μια επισκόπηση των αλγορίθμων της μηχανικής μάθησης που έχουν χρησιμοποιηθεί στο πειραματικό μέρος.

3.1. Τυχαίο Δάσος (Random Forest)

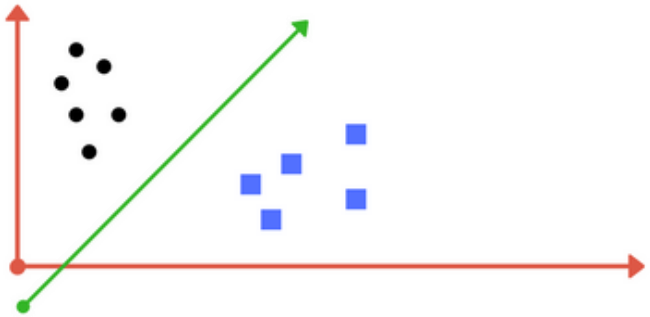
Το τυχαίο δάσος είναι μια μέθοδος μηχανικής μάθησης που λειτουργεί με την κατασκευή πολλαπλών δέντρων αποφάσεων κατά τη διάρκεια της φάσης εκπαίδευσης. Τα δέντρα αποφάσεων είναι μοντέλα που βασίζονται στον διαχωρισμό του χώρου των χαρακτηριστικών και στην αποθήκευση μιας κατανομής σε ετικέτες κλάσεων για κάθε περιοχή. Αυτό μπορεί να επαναληφθεί χρησιμοποιώντας ένα δέντρο, το οποίο με τη σειρά του υποδηλώνει ότι η συνάρτηση υπόθεσης ή η πρόβλεψη για αυτόν τον αλγόριθμο εκμάθησης έχει τη μορφή δέντρου. Κάθε φύλλο αυτού του δέντρου θα αντιστοιχεί στη συνέχεια σε κάθε περιοχή και του αποδίδεται η αντίστοιχη κατανομή πιθανοτήτων στις κλάσεις. Με βάση αυτήν την περιγραφή, μπορούμε να χρησιμοποιήσουμε δέντρα αποφάσεων για να δημιουργήσουμε προγνωστικούς παράγοντες για κάθε περίπτωση δοκιμής. Η δομή του δέντρου είναι πολύ ευαίσθητη στα παρεχόμενα δεδομένα. Αυτό σημαίνει ότι μικρές αλλαγές στα δεδομένα μπορεί να επιδράσουν δραστικά στην δομή του δέντρου που θα προκύψει. Για να μετριάσουμε αυτό το πρόβλημα, χρησιμοποιούμε τα τυχαία δάση. Τέτοια μοντέλα μπορούν να θεωρηθούν ως ένας συνδυασμός πολλαπλών δέντρων αποφάσεων. Δημιουργούνται εκπαιδεύοντας διαφορετικά δέντρα σε διαφορετικά υποσύνολα δεδομένων που επιλέγονται τυχαία με αντικατάσταση. Η προκείπτουσα πρόβλεψη από ένα τυχαίο δάσος παράγεται στη συνέχεια συνδυάζοντας αυτά τα δέντρα απόφασης.



ΣΧΗΜΑ 6 ΤΥΧΑΊΟ ΔΆΣΟΣ

3.2. Αλγόριθμος Μηχανών Διανυσμάτων Υποστήριξης (SVM)

Οι Μηχανές Διανυσμάτων Υποστήριξης είναι γραμμικά μοντέλα που χρησιμοποιούνται σε προβλήματα παλινδρόμησης και κατηγοριοποίησης. Μπορούν να επιλύσουν γραμμικά και μη προβλήματα. Στην πρώτη προσέγγιση οι ΜΔΥ βρίσκουν μια διαχωριστική γραμμή ή υπερεπίπεδο μεταξύ δεδομένων δύο κλάσεων. Ο SVM είναι ένας αλγόριθμος που λαμβάνει τα δεδομένα ως είσοδο και δίνει ως έξοδο μια γραμμή που διαχωρίζει αυτές τις κλάσεις. Σε αντίθεση με την λογιστική παλινδρόμηση, η οποία ορίζει την βέλτιστη από τη συνολική πιθανότητα, ο αλγόριθμος SVM θέλει την μικρότερη απόσταση μεταξύ των σημείων δεδομένων και του ορίου απόφασης να είναι όσο το δυνατόν μεγαλύτερη. Ο SVM επιλέγει τα ακραία σημεία ή διανύσματα που βοηθούν στην δημιουργία του υπερεπιπέδου. Αυτές οι ακραίες περιπτώσεις αποκαλούνται διανύσματα υποστήριξης. Στο παρακάτω σχήμα παρουσιάζονται δύο κλάσεις που ταξινομούνται κάνοντας χρήση ενός ορίου απόφασης (υπερεπίπεδο).



ΣΧΗΜΑ 7 ΠΑΡΑΔΕΙΓΜΑ SVM ΠΟΥ ΔΙΑΧΩΡΙΖΕΙ ΤΙΣ ΔΥΟ ΚΛΑΣΕΙΣ ΜΕ ΓΡΑΜΜΗ

Είναι πιθανό να υπάρχουν πολλά όρια απόφασης για τον διαχωρισμό των κλάσεων σε ν-διάστατο χώρο. Αυτό σημαίνει ότι πρέπει να βρεθεί το βέλτιστο όριο απόφασης (υπερεπίπεδο του SVM) που θα ταξινομήσει με μεγαλύτερη ακρίβεια τα δεδομένα. Οι διαστάσεις του υπερεπιπέδου εξαρτώνται από τα χαρακτηριστικά-μεταβλητές που θα χρησιμοποιήσουμε. Αν υπάρχουν μόλις δύο χαρακτηριστικά στο σύνολο των δεδομένων, τότε το υπερεπίπεδο θα είναι μία ευθεία γραμμή, όπως στο Σχήμα 6. Αν όμως υπάρχουν τρεις μεταβλητές τότε το υπερεπίπεδο θα είναι ένα επίπεδο δύο διαστάσεων.

3.3. Μετρικές Επίδοσης

- **Ευστοχία (Accuracy)**

Η ευστοχία ενός μοντέλου είναι ουσιαστικά ο λόγος του πλήθους των ορθών προβλέψεων προς το σύνολο των προβλέψεων.

$$\boldsymbol{Accuracy = \frac{Number\ of\ correct\ predictions}{Number\ \ of all\ predictions}}$$

Εάν ένα μοντέλο προβλέψει σωστά τις 90 από τις συνολικά 100 παρατηρήσεις ή στην περίπτωση μας τις κατηγοριοποιήσει σωστά, τότε η ακρίβεια του είναι 90%. Σε μερικές περιπτώσεις η συγκεκριμένη

μετρική επίδοσης μπορεί να είναι παραπλανητική. Για παράδειγμα αν έχουμε ένα μοντέλο με δύο κλάσεις(δυαδικό πρόβλημα κατηγοριοποίησης) με μη ισορροπημένο σύνολο δεδομένων, δηλαδή να έχουμε πολλές παραπάνω περιπτώσεις της μια κλάσης σε σχέση με την άλλη, π.χ. 91% κλάση 0 και 9% κλάση 1, το μοντέλο μπορεί να επιτύχει ευστοχία 91% ταξινομώντας ορθά τα σημεία δεδομένων της μηδενικής κλάσης. Υπάρχουν περιπτώσεις που το ζητούμενο είναι να γίνει σωστή πρόβλεψη/ταξινόμηση της κλάσης που έχει τις λιγότερες εγγραφές (π.χ. έρευνες σχετικές με καρκίνο). Σε τέτοιες περιπτώσεις αν θέλουμε να χρησιμοποιήσουμε την συγκεκριμένη μετρική, θα πρέπει με κάποιον τρόπο να αντιμετωπίσουμε το πρόβλημα της ανισορροπίας των δεδομένων.

- **Επανάκληση (Recall)**

Η επανάκληση επίσης χρησιμοποιείται στην περίπτωση της δυαδικής κατηγοριοποίησης και επικεντρώνεται στις θετικές προβλέψεις.

Είναι ο λόγος των ορθά θετικών προς το σύνολο των ορθά θετικών και λανθασμένα αρνητικών (δηλ. όλων των εγγραφών που είναι στην θετική κλάση). Αξιολογεί την ικανότητα του μοντέλου να προβλέψει την θετική κλάση.'

$$\boldsymbol{Recall = \frac{TP}{TP + FN}}$$

Για μια εργασία ανίχνευσης όγκου, πρέπει να μεγιστοποιήσουμε την ανάκληση επειδή θέλουμε να ανιχνεύσουμε θετικές κλάσεις όσο το δυνατόν περισσότερο. Δεν έχουμε την πολυτέλεια να ταξινομήσουμε λάθος καμία θετική κλάση (δηλαδή μια περίπτωση με όγκο).

- **Βαθμολογία F1**

Η βαθμολογία F1 είναι ο σταθμισμένος μέσος όρος ακρίβειας και ανάκλησης.

$$F1\ Score = 2\frac{Precision*Recall}{Precision+Recall}$$

Η μετρική αξιολόγησης F1 είναι από τις πιο χρήσιμες για προβλήματα ταξινόμησης με άνιση κατανομή των κλάσεων/ετικετών καθώς λαμβάνει υπόψιν και τα λανθασμένα θετικά και τα λανθασμένα αρνητικά.

Κεφάλαιο 4

Υλοποίηση

4.1. Σκοπός

Σε αυτό το κεφάλαιο θα γίνει η περιγραφή και παρουσίαση της ανάλυσης που εφαρμόστηκε. Σκοπός της ανάλυσης ήταν η ταξινόμηση των τροχιών που δημιουργήσαν χρήστες του Twitter μέσα από τα check-in που πραγματοποίησαν. Οι κλάσεις στις οποίες θα ταξινομηθούν αυτές οι τροχιές είναι οι ίδιοι οι χρήστες. Η προεπεξεργασία των δεδομένων βασίστηκε στην μέθοδο MasterMovelets των Ferrero et al.(2020). Στη συνέχεια γίνεται προσπάθεια κατηγοριοποίησης των χρηστών του Twitter χρησιμοποιώντας όχι τροχιές αλλά σημεία των τροχιών, καθώς όπως θα φανεί στη συνέχεια η κατηγοριοποίηση των χρηστών με τα συγκεκριμένα δεδομένα και την συγκεκριμένη μέθοδο, δεν απέδωσε τόσο καλά. Πρέπει να σημειωθεί ότι δε γίνεται σύγκριση των δύο μεθόδων, καθώς στην πρώτη γίνεται χρήση τροχιών ενώ στην δεύτερη σημείων των τροχιών. Αλλά στην περίπτωση που ο σκοπός μας είναι να γίνει κατηγοριοποίηση των χρηστών χωρίς να υπάρχει περιορισμός στην μορφή των δεδομένων (τροχιές ή σημειακά δεδομένα) τότε η δεύτερη μέθοδος στα συγκεκριμένα δεδομένα παρουσιάζει καλύτερα αποτελέσματα.

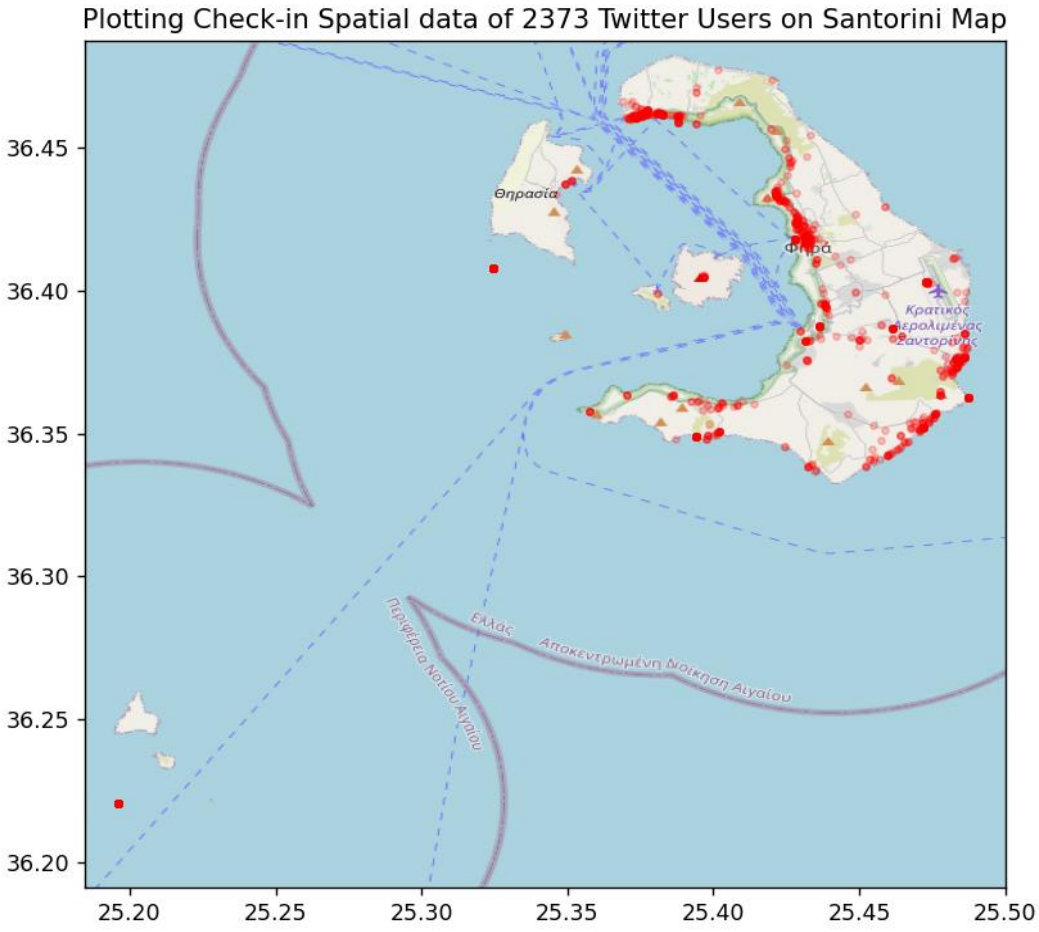
4.2. Περιγραφή δεδομένων

Πραγματοποιείται αξιολόγηση της μεθόδου MasterMovelets με ένα πραγματικό σύνολο δεδομένων του Twitter. Τα δεδομένα που χρησιμοποιήθηκαν προέρχονται από το Twitter API και αποτελούνται από tweets χρηστών που την περίοδο 2018-2019 βρέθηκαν στην Σαντορίνη. Παρακάτω παρουσιάζονται οι μεταβλητές του αρχικού συνόλου δεδομένων χωρίς να έχει υποστεί καμία επεξεργασία:

- Tweet ID: το κλειδί του tweet που είναι μοναδικό ανά tweet.
- Timestamp: υποδεικνύει την χρονική στιγμή (ημέρα και ώρα) που πραγματοποιήθηκε το tweet
- User ID: το κλειδί του χρήστη.
- Bounding Box: περίγραμμα γεωγραφικών συντεταγμένων με τέσσερα ζευγάρια γεωγραφικού μήκους και πλάτους.

- Hashtag: Λέξη – κλειδί που επιτρέπει στον χρήστη να ταξινομήσει την δημοσίευσή του με άλλες που φέρουν το ίδιο hashtag. Ανά tweet μπορούν να υπάρχουν πάνω από ένα hashtag ή και κανένα. Η συγκεκριμένη μεταβλητή δεν χρησιμοποιήθηκε στην ανάλυση.

Πριν γίνει η εφαρμογή της μεθοδολογίας Mastermovelets έπρεπε να φέρουμε τα αρχικά δεδομένα στην μορφή που απαιτείται. Το πρώτο βήμα ήταν να βρεθούν τα POI (Point Of Interest) του κάθε tweet χρησιμοποιώντας τις συντεταγμένες. Αυτό επιτεύχθηκε με την βοήθεια του Overpass API. Συγκεκριμένα, κάνοντας ερωτήματα στο API για τις συγκεκριμένες συντεταγμένες λαμβάνουμε ως αποτέλεσμα ένα σύνολο δεδομένων. Από αυτά τα δεδομένα μας ενδιαφέρει το μέρος που αντιστοιχεί στις συντεταγμένες πχ. εστιατόριο, κοσμηματοπωλείο, καφέ κλπ.



ΣΧΗΜΑ 8 ΧΑΡΤΗΣ ΣΑΝΤΟΡΙΝΗΣ ΜΕ ΤΙΣ ΓΕΩΓΡΑΦΙΚΕΣ ΘΕΣΕΙΣ (ΚΟΚΚΙΝΟ) ΤΩΝ 2373 ΧΡΗΣΤΩΝ TWITTER

Τα αρχικά δεδομένα προέρχονται από 2373 χρήστες του Twitter που βρέθηκαν στη Σαντορίνη την περίοδο 2018-2019. Στο παραπάνω σχήμα φαίνεται σε ποιες περιοχές του νησιού πραγματοποιούνται τα περισσότερα check-in. Η πλειοψηφία των tweets πραγματοποιήθηκε στις παραλιακές περιοχές του νησιού. Από αυτούς τους χρήστες επιλέχθηκαν όσοι είχαν κάνει πάνω από 7 tweets σε διάστημα μιας εβδομάδας και για τουλάχιστον 3 εβδομάδες, οι οποίοι ήταν 14. Αξίζει επίσης να σημειωθεί πως όσο αυξανόταν ο αριθμός των χρηστών τόσο μειωνόταν η ακρίβεια των αλγορίθμων κατηγοριοποίησης καθώς οι περισσότεροι χρήστες είχαν ελάχιστα tweets σε πολύ μικρό διάστημα. Σύμφωνα με τη μέθοδο MasterMovelets, γίνεται περιορισμός του μέγιστου μήκους των movelets στο μήκος της μικρότερης τροχιάς του συνόλου δεδομένων.

Αφού βρέθηκαν τα POI, τα δεδομένα επεξεργάστηκαν και έφτασαν στην τελική τους μορφή ώστε στη συνέχεια να περάσουμε στην μεθοδολογία MasterMovelets.

Παρακάτω παρουσιάζονται τα δεδομένα με την τελική τους μορφή:

ΠΙΝΑΚΑΣ 1: ΜΕΤΑΒΛΗΤΕΣ ΓΙΑ ΜΕΘΟΔΟ MASTERMOVELETS

Μεταβλητή	Παράδειγμα	Τύπος Μεταβλητής
Space: Γεωγραφικό μήκος και πλάτος	36.2205594 25.1959315	Space2d
DateTime: Ημερομηνία και ώρα	2018-08-23 17:03:52	Ονομαστική
Time: Χρονική μεταβλητή που δημιουργήθηκε προσθέτοντας τα στοιχεία της ώρας σε μορφή λεπτών πχ 17:03:52 → 17*60 + 3 = 1023	**1023**	**Αριθμητική**
Day: Ημέρα Εβδομάδας	Friday	Ονομαστική
POI: Σημείο Ενδιαφέροντος	place_of_worship	Venue
POI number	22	Αριθμητική

Στη συνέχεια το σύνολο δεδομένων χωρίζεται σε συμπιεσμένα .r2 αρχεία το κάθε ένα από τα οποία περιέχει τις πληροφορίες κάθε tweet που πραγματοποιήθηκε μέσα σε μια εβδομάδα (Δευτέρα – Κυριακή) ανά χρήστη.

Ο διαχωρισμός των δεδομένων σε train και test set, καθώς και το cross-validation (5 folds) πραγματοποιήθηκαν στον κώδικα data_preparation.py. Αφού δημιουργηθούν οι φάκελοι train και test, αυτοί συμπιέζονται και δημιουργούνται τα αρχεία test.zip και train.zip. Σημειώνεται ότι η αναλογία ήταν 80% train και 20% test.

4.3. Υλοποίηση μεθοδολογίας MasterMovelets

Το επόμενο βήμα είναι η εκτέλεση του κώδικα MasterMovelets (Ferrero et al., 2020). Χρησιμοποιήθηκε ο κώδικας που βρίσκεται στο αποθετήριο Github : https://github.com/anfer86/dmkd_masterMovelets_results.

Έγινε χρήση της παρακάτω εντολής και εκτελέστηκε σε περιβάλλον Git bash:

```
JAVA_PATH=C:/Program\ Files/Java/jdk1.8.0_202/bin/java.exe

runs=(1 2 3 4 5)

for i in "${runs[@]}"; do

"$JAVA_PATH" -Xmx12g -Xms9g -jar programs/MasterMoveletsByClass.jar -
curpath data_5splits/run"$i" -respath results_5splits/results_run"$i" -
descfile data_5splits/descriptions/tweets.json -nt 2 -cache true -ms 1 -
Ms -1 -ed true -mnf -1 -samples 1 -sampleSize 0.5 -medium "none" -output
"discrete" -lowm "false"> results_5splits/run"$i"_results.txt

done
```

Όπου runs είναι τα 5 splits του συνόλου δεδομένων καθώς θέλουμε ο κώδικας Mastermovelets να τρέξει και στα 5. Το πρόγραμμα παράγει δύο αρχεία : train & test, για κάθε κλάση-χρήστη, τα οποία περιέχουν τα movelets. Στη συνέχεια, εκτελείται ο κώδικας MergeDatasets.R ο οποίος ενώνει όλα τα test όλων των κλάσεων και αντίστοιχα όλα τα train. Έτσι καταλήγουμε με δύο .csv αρχεία (train.csv και test.csv).

Αξίζει να σημειωθεί πως στην εργασία τους οι Ferrero et al (2020) επέλεξαν εκείνους τους χρήστες που είχαν πραγματοποιήσει τουλάχιστον 10 tweets την ημέρα επί 10 εβδομάδες. Στη δική μας περίπτωση αυτό δεν είναι εφικτό καθώς η Σαντορίνη είναι τουριστικός προορισμός και είναι ασυνήθιστο κάποιος να παραμείνει εκεί επί 10 εβδομάδες. Τα περισσότερα tweets προέρχονται από χρήστες που παρέμειναν εκεί για το διάστημα 1-2 εβδομάδες. Μόλις δύο χρήστες βρέθηκαν που είχαν τα ανωτέρω χαρακτηριστικά.

Εδώ τελειώνει το μέρος της επεξεργασίας των δεδομένων και μπορούμε να περάσουμε στο κομμάτι της κατηγοριοποίησης.

Για την κατηγοριοποίηση επιλέχθηκαν τρεις αλγόριθμοι: Random Forest και SVM. Από τους χρήστες που υπήρχαν στο αρχικό σύνολο δεδομένων επιλέχθηκαν εκείνοι που έκαναν πάνω από 7 tweets μέσα σε μία εβδομάδα και έκαναν tweets για τουλάχιστον 3 εβδομάδες. Εφαρμόζοντας αυτές τις προϋποθέσεις στο τέλος από τους 2373 μας έμειναν 14 χρήστες και συνολικά 3670 καταχωρίσεις.

Στο Νευρωνικό Δίκτυο έγινε χρήση ενός κρυφού στρώματος με 2000 units με παράμετρο dropout=0.02 και Adam optimizer με ρυθμό εκμάθησης (learning rate) 0.01.

Για την περίπτωση του Τυχαίου Δάσους, χρησιμοποιήθηκαν 500 εκτιμητές (δέντρα στο δάσος), το κριτήριο gini για την αξιολόγηση της ποιότητας διαχωρισμού (split), ο ελάχιστος αριθμός των δειγμάτων που απαιτούνταν για τον διαχωρισμό ενός εσωτερικού κόμβου είναι 5 και ο ελάχιστος αριθμός δειγμάτων σε ένα φύλλο είναι τα 4 δείγματα.

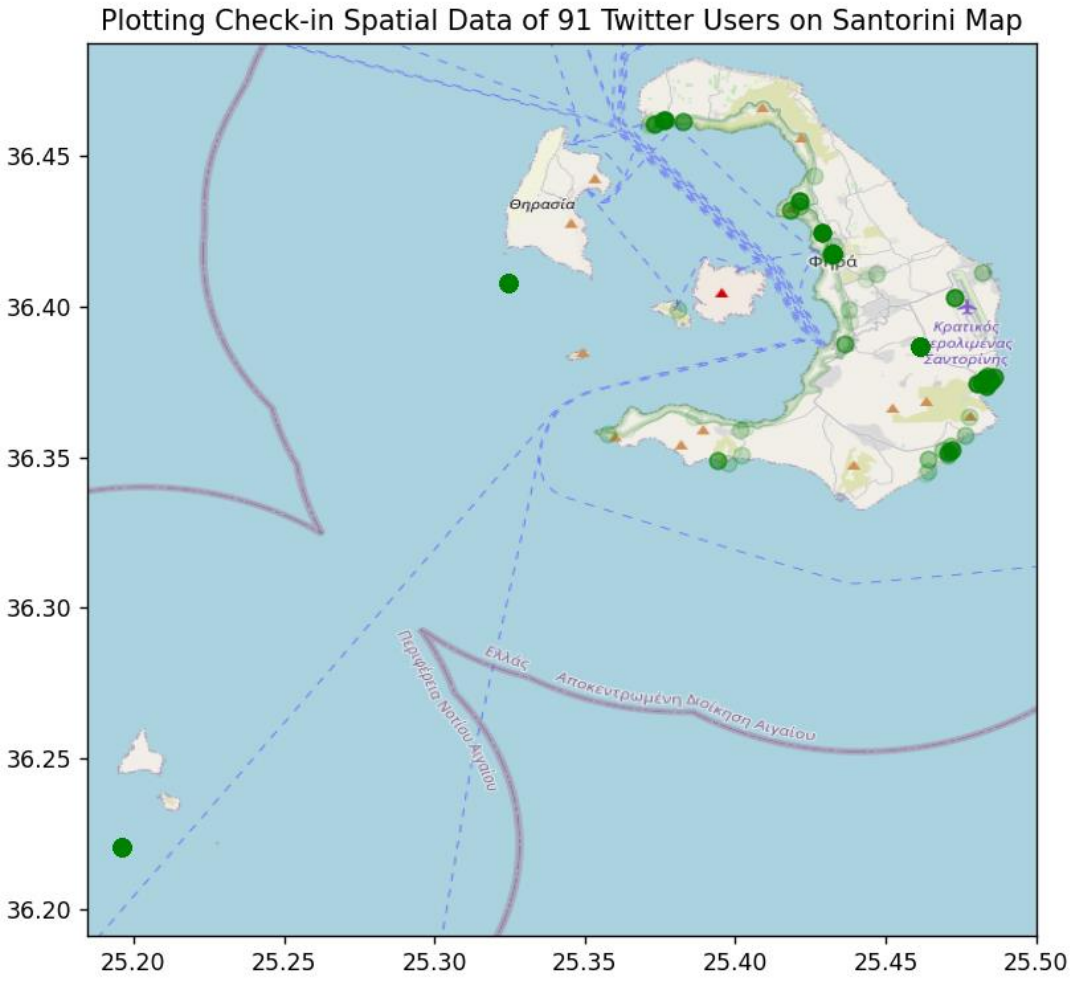
Τέλος, στην περίπτωση του SVC ταξινομητή, η μόνη παράμετρος που ορίστηκε ήταν της κανονικοποίησης και ήταν ίση με C=10000.

Δοκιμάσαμε επίσης να μην εφαρμόσουμε την μέθοδο Mastermovelets στα δεδομένα, περνώντας κατευθείαν στην κατηγοριοποίηση. Από τους 2373 χρήστες επιλέχθηκαν αυτοί που είχαν κάνει πάνω από 50 tweets σε όλο το διάστημα. Καταλήξαμε με 91 χρήστες και συνολικά 45.572 καταχωρίσεις. Τα αποτελέσματα βελτιώθηκαν αισθητά για όλους τους αλγορίθμους. Οι μεταβλητές δεν έχουν υποστεί κάποια περαιτέρω επεξεργασία και παρουσιάζονται στον παρακάτω πίνακα:

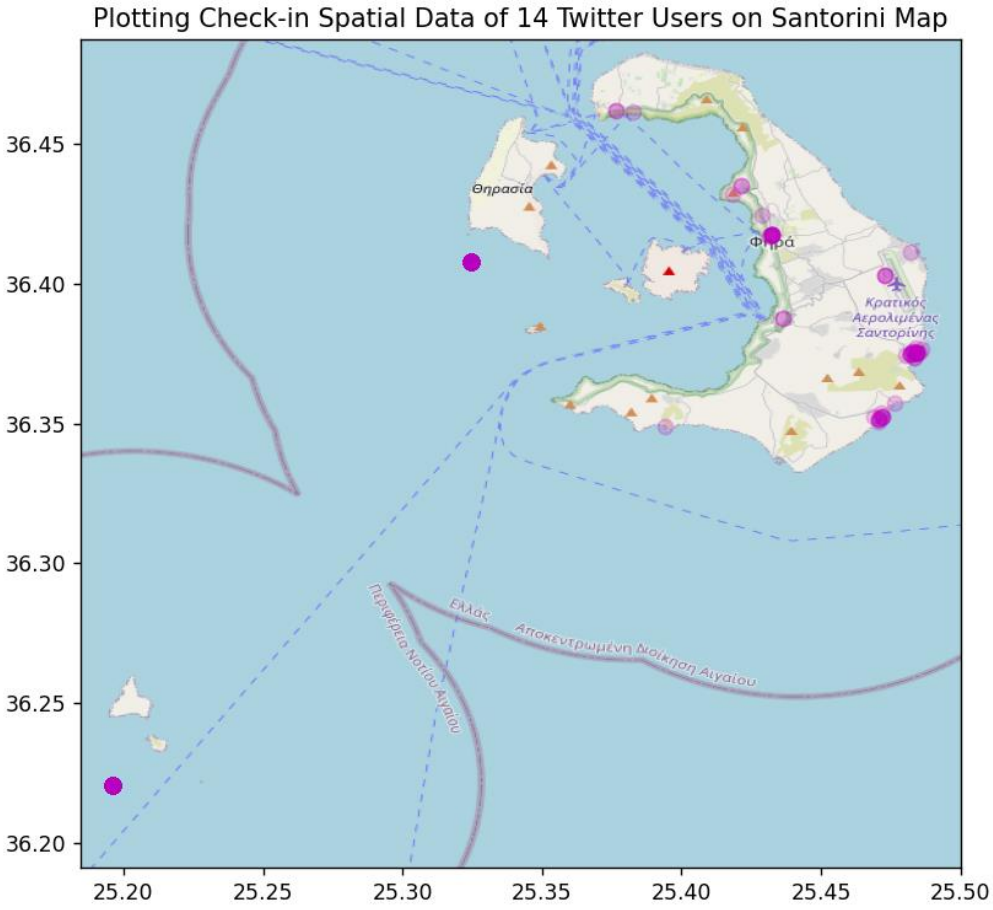
ΠΙΝΑΚΑΣ 2:ΜΕΤΑΒΛΗΤΕΣ ΜΕ ΑΠΛΗ ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ

ΜΕΤΑΒΛΗΤΗ	ΠΑΡΑΔΕΙΓΜΑ
LONGTITUDE	36.2205594
LATITUDE	25.1959315
DATE	2018-08-23
DAY	FRIDAY
MINUTES	22
POI NUMBER	3

Για τον ταξινομητή SVC επιλέξαμε τον πολυωνυμικό πυρήνα και 6000 μέγιστες επαναλήψεις. Στον αλγόριθμο του Τυχαίου Δάσους δεν μεταβλήθηκε κάποια παράμετρος από τις προκαθορισμένες.



ΣΧΗΜΑ 9 ΧΑΡΤΗΣ ΣΑΝΤΟΡΙΝΗΣ ΜΕ ΤΙΣ ΓΕΩΓΡΑΦΙΚΕΣ ΘΕΣΕΙΣ (ΠΡΑΣΙΝΟ) ΤΩΝ 91 ΧΡΗΣΤΩΝ TWITTER



ΣΧΗΜΑ 10 ΧΑΡΤΗΣ ΣΑΝΤΟΡΙΝΗΣ ΜΕ ΤΙΣ ΓΕΩΓΡΑΦΙΚΕΣ ΘΕΣΕΙΣ (POZ) ΤΩΝ 14 ΧΡΗΣΤΩΝ TWITTER

Ίσως η μέθοδος MasterMovelets να μην ταιριάζει σε δεδομένα όπως είναι τα δικά μας, δηλαδή με λίγες εγγραφές ανά χρήστη συγκεντρωμένα όλα σε ένα περιορισμένο γεωγραφικό σημείο και με τροχιές παρόμοιες μεταξύ τους, δεδομένου του χαρακτήρα της συγκεκριμένης γεωγραφικής περιοχής (τουριστικός προορισμός, μικρή ποικιλία στα μέρη που μπορεί να επισκεφτεί κανείς). Όπως φαίνεται οι μόνιμοι κάτοικοι της Σαντορίνης δεν κάνουν χρήση του Twitter καθώς βρέθηκαν μόλις δύο χρήστες που μέσα στο διάστημα 2018-2019 είχαν πραγματοποιήσει πάνω από 10 tweets ανά εβδομάδα για διάστημα μεγαλύτερο των 10 εβδομάδων. Αντιθέτως τα δεδομένα που χρησιμοποίησαν οι Ferrero et al (2018) προέρχονταν από χρήστες μεγαλουπόλεων που έκαναν έντονη χρήση του Twitter ή άλλων LBSN.

Ακολουθουν πίνακες για την σύγκριση των αποτελεσμάτων με και χωρίς την μέθοδο MasterMovelets και για τους τρεις αλγορίθμους κατηγοριοποίησης που έχουν επιλεγεί.

ΠΙΝΑΚΑΣ 3 ΑΠΟΤΕΛΕΣΜΑΤΑ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΜΕ ΤΥΧΑΙΟ ΔΑΣΟΣ

Μέθοδος	Precision	Recall	Accuracy	F1 Score
MasterMovelets	0.48	0.45	0.45	0.437
Απλή Επεξεργασία	**0.978**	**0.977**	**0.978**	**0.976**

ΠΙΝΑΚΑΣ 4 ΑΠΟΤΕΛΕΣΜΑΤΑ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΜΕ SVM

Μέθοδος	Precision	Recall	Accuracy	F1 Score
MasterMovelets	0.48	0.39	0.3805	0.3658
Απλή Επεξεργασία	**0.575**	**0.566**	**0.566**	**0.494**

Για την αξιολόγηση των αποτελεσμάτων χρησιμοποιήθηκαν οι μετρικές precision, recall, accuracy, f1 score. Είναι εμφανές από τους παραπάνω πίνακες πως η μεθοδολογία MasterMovelets υστερεί σε όλες τις περιπτώσεις με αρκετά μεγάλη διαφορά και συγκεκριμένα η καλύτερη κατηγοριοποίηση επιτυγχάνεται με Τυχαίο Δάσος δίχως την χρήση της μεθόδου με όλες τις μετρικές να ξεπερνάνε το 95%.Χρησιμοποιήθηκε ένα κριτήριο για όλες τις κλάσεις καθώς και στις δύο περιπτώσεις αυτές ήταν πολλές, 14 στην περίπτωση με την μέθοδο MasterMovelets και 91 στην περίπτωση της απλής επεξεργασίας.

Κεφάλαιο 5

Συμπεράσματα και μελλοντικές επεκτάσεις

Η αύξηση των συσκευών εντοπισμού τοποθεσίας του χρήστη έχει δημιουργήσει πληθώρα διαθέσιμων δεδομένων κίνησης και κατ' επέκταση τροχιών που αντιπροσωπεύουν τα κινούμενα αντικείμενα που τις παράγουν. Έχουν αναπτυχθεί πολλές μέθοδοι εξόρυξης τροχιών, μερικές από τις οποίες παρουσιάστηκαν στο θεωρητικό κομμάτι της εργασίας, μία από τις οποίες ήταν η μέθοδος MasterMovelets (Ferrero et al.,) στην οποία και βασίστηκε η παρούσα εργασία. Παρόλα αυτά δεν έχει διερευνηθεί μια ολοκληρωμένη οπτική επίλυσης προβλημάτων στον τομέα της εξόρυξης τροχιών.

Στην παρούσα εργασία επιχειρήθηκε να εφαρμοστεί η μέθοδος MasterMovelets (Ferrero et al., 2020) σε ένα σύνολο δεδομένων που όπως αποδείχθηκε παρουσίασε κάποια προβλήματα. Τα δεδομένα που χρησιμοποιήθηκαν αν και με μια πρώτη ματιά έμοιαζαν με τα δεδομένα της εργασίας MasterMovelets, στην πράξη απείχαν κατά πολύ καθώς ήταν παρών ο γεωγραφικός περιορισμός που οδήγησε στην ομοιότητα των τροχιών και η φύση της περιοχής, δηλαδή ο λόγος επίσκεψης της Σαντορίνης είναι κατά βάση τουριστικός που σημαίνει ότι οι περισσότεροι χρήστες Twitter θα επισκεφτούν παρόμοια μέρη το οποίο συμβάλλει στο χαρακτηριστικό της ομοιότητας που έχουν οι συγκεκριμένες τροχιές. Τέλος, υπήρχαν ελάχιστοι χρήστες που ικανοποιούσαν τα αρχικά κριτήρια (τουλάχιστον 10 tweets ανά τροχιά σε σύνολο τουλάχιστον 10 τροχιών ανά χρήστη), οπότε αυτά έγιναν πιο ελαστικά με αποτέλεσμα να μειωθεί η ακρίβεια των ταξινομητών. Αυτός ήταν και ο λόγος που επιχειρήσαμε να κάνουμε ταξινόμηση χωρίς τροχιές αλλά με σημειακά δεδομένα. Αν σκοπός δεν ήταν η εξόρυξη τροχιών αλλά η κατηγοριοποίηση τότε ίσως θα μπορούσαμε να παραλείψουμε την εξόρυξη τροχιών όταν έχουμε τέτοια δεδομένα.

Η εργασία που παρουσιάζεται μπορεί να επεκταθεί με την δημιουργία αντίστοιχης μεθόδου εξόρυξης τροχιών που θα ταιριάζει σε δεδομένα που έχουν τέτοια χαρακτηριστικά, δηλαδή τροχιές με παρόμοια χωροχρονικά δεδομένα ή και ακόμα με την επέκταση της ήδη υπάρχουσας με τέτοιον τρόπο ώστε να διαφοροποιούνται οι τροχιές/υποτροχιές που προέρχονται από διαφορετικούς χρήστες. Άλλος τρόπος διαφοροποίησης των τροχιών ίσως θα μπορούσε να είναι προσθήκη κάποιας επιπλέον σημασιολογικής μεταβλητής πέρα από την μεταβλητή POI (Σημείο Ενδιαφέροντος).

Βιβλιογραφια

Almeida, V. & Güting, Ralf & Behr, T.. (2006). *Querying Moving Objects* in SECONDO. 47 - 47. 10.1109/MDM.2006.133.

Collier, Nigel & Nguyen, Son & Ngoc, Mai. (2010). *OMG U got flu? Analysis of shared health messages for bio-surveillance*.

Culotta, Aron. (2010). *Detecting influenza outbreaks by analyzing Twitter messages*.

Earle, Paul & Bowden, Daniel & Guy, Michelle. (2012). *Twitter earthquake detection: Earthquake monitoring in a social world*. Annals of geophysics = Annali di geofisica. 54. 10.4401/ag-5364.

Ferrero, Carlos & Alvares, Luis & Zalewski, Willian & Bogorny, Vania. (2018). *MOVELETS: Exploring Relevant Subtrajectories for Robust Trajectory Classification*. 10.1145/3167132.3167225.

Ferrero, Carlos & May Petry, Lucas & Alvares, Luis & Leite da Silva, Camila & Zalewski, Willian & Bogorny, Vania. (2020). *MasterMovelets: discovering heterogeneous movelets for multiple aspect trajectory classification.* Data Mining and Knowledge Discovery. 34. 10.1007/s10618-020-00676-x.

Freitas, Nicksson & Silva, Ticiana & Macêdo, José & Junior, Leopoldo & Cordeiro, Matheus. (2021). *Using Deep Learning for Trajectory Classification*. 664-671. 10.5220/0010227906640671.

Gao, Qiang & Zhou, Fan & Zhang, Kunpeng & Trajcevski, Goce & Luo, Xucheng & Zhang, Fengli. (2017). *Identifying Human Mobility via Trajectory Embeddings.*

Guo, Tao, and Lei Xie. 2022. *"Research on Ship Trajectory Classification Based on a Deep Convolutional Neural Network" Journal of Marine Science and Engineering* 10, no. 5: 568. https://doi.org/10.3390/jmse1005056

Güting, R.H., Bohlen, M.H., Erwig, M., Jensen, C.S., Lorentzos, N.A., Schneider, M., & Vazirgiannis, M. (2000). *A foundation for representing and querying moving objects. ACM Transactions on Database Systems*, 25(1), 1-42.

Khosroshahi, Aida & Ohn-Bar, Eshed & Trivedi, Mohan. (2016). *Surround vehicles trajectory analysis with recurrent neural networks*. 2267-2272. 10.1109/ITSC.2016.7795922.

Kumaran, Natarajan & Reddy, U. Srinivasulu. (2021). *Classification of human activity detection based on an intelligent regression model in video sequence*s. IET Image Processing. 15. 1-12. 10.1049/ipr2.12006.

Lee, Jae-Gil & Han, Jiawei & Li, Xiaolei & Gonzalez, Hector. (2008). *Traclass: Trajectory Classification Using Hierarchical Region-Based and Trajectory-Based Clustering*. Proceedings of The Vldb Endowment - PVLDB. 1. 10.14778/1453856.1453972.

Li, Jessica & Rao, Raghav. (2010). *Twitter as a Rapid Response News Service: An Exploration in the Context of the 2008 China Earthquake*. EJISDC: The Electronic Journal on Information Systems in Developing Countries, ISSN 1681-4835, Vol. 42, 2010. 42. 10.1002/j.1681-4835.2010.tb00300.

Li, Xiaolei & Han, Jiawei & Kim, Sangkyum & Gonzalez, Hector. (2007). *ROAM: Rule- and Motif-Based Anomaly Detection in Massive Moving Object Data Sets*. 10.1137/1.9781611972771.25.

Lucas May Petry, Camila Leite Da Silva, Andrea Esuli, Chiara Renso & Vania Bogorny (2020*) MARC: a robust method for multiple-aspect trajectory classification via space, time, and semantic embeddings*, International Journal of Geographical Information Science, 34:7, 1428-1450, DOI: 10.1080/13658816.2019.1707835

Pelekis, Nikos & Frentzos, Elias & Giatrakos, Nikos & Theodoridis, Yannis. (2015). *HERMES: A trajectory DB engine for mobility-centric applications*. Int. J. Knowl. Based Organ. 5. 19-41. 10.4018/ijkbo.2015040102.

Rossi,Luca&Musolesi,Mirco.(2014*).It's the Way You Check-in: Identifying Users in Location-based Social Networks*. COSN 2014-Proceedings of the 2014 ACM Conference on Online SocialNetworks.215-226. 10.1145/2660460.2660485

Rout, Dominic & Preotiuc-Pietro, Daniel & Bontcheva, Kalina & Cohn, Trevor. (2013*). Where's @wally: A classification approach to Geolocating users based on their social ties.*

Spaccapietra, Stefano & Parent, Christine & Macedo, Jose & Porto, Fabio & Vangenot, Christelle. (2008). *A Conceptual View on Trajectories. Data & Knowledge Engineering*. 65. 126-146. 10.1016/j.datak.2007.10.008.

Stöggl, Thomas & Holst, Anders & Jonasson, Arndt & Andersson, Erik & Wunsch, Tobias & Norström, Christer & Holmberg, Hans-Christer. (2014). *Automatic Classification of the Sub-Techniques (Gears) Used in Cross-Country Ski Skating Employing a Mobile Phone. Sensors*. 14. 20589-20601. 10.3390/s141120589.

Torres, Leigh & Orben, Rachael & Tolkova, Irina & Thompson, David. (2017). *Classification of Animal Movement Behavior through Residence in Space and Time*. PLOS ONE. 12. e0168513. 10.1371/journal.pone.0168513.