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**

ΕΥΡΕΤΗΡΙΑΣΗ ΔΕΔΟΜΕΝΩΝ ΚΙΝΗΣΗΣ ΜΕ ΧΡΗΣΗ ΜΟΝΤΕΛΩΝ

Αρτέμης Ν. Ηλιάκης

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
Ιούλιος 2016

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:

- Ν. Πελέκης, Επίκουρος Καθηγητής (Επιβλέπων)
- Μ. Κούτρας, Καθηγητής
- Α. Πικράκης, Επίκουρος 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

MODEL-BASED INDEXING OF MOBILITY DATA

By

Artemis N. Iliakis

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
July 2016

Περίληψη

Την τελευταία δεκαετία, ο τομέας της εξόρυξης γνώσης και διαχείρισης δεδομένων κίνησης έχει αναδυθεί παρέχοντας πολλές αποτελεσματικές μεθόδους για την εξόρυξη διαισθητικών προτύπων τα οποία αντιπροσωπεύουν ομαδικές συμπεριφορές κινούμενων αντικειμένων. Μία ενδιαφέρουσα ερευνητική τάση είναι ότι αντί να διαχειρίζονται τα ακατέργαστα δεδομένα κίνησης που συλλέγονται από διάφορους αισθητήρες, οι ερευνητές κάνουν χρήση των σημασιολογικά εμπλουτισμένων δεδομένων, τα οποία είτε δηλώνονται (από τους χρήστες) είτε προκύπτουν-εξάγονται (με κάποια μέθοδο σχολιασμού). Με αυτόν τον τρόπο τα δεδομένα των ακατέργαστων τροχιών των κινούμενων αντικειμένων μετατρέπονται σε χωρο-χρονο-κειμενικές ακολουθίες, όπου η επιπλέον κειμενική πληροφορία, που προστίθεται στις διαστάσεις του χώρου και του χρόνου, αντιπροσωπεύει τη σημασιολογία της κίνησης. Τέτοιες χωρο-χρονο-κειμενικές ακολουθίες σχηματίζουν ένα πιο ρεαλιστικό μοντέλο αναπαράστασης της πολυσύνθετης καθημερινότητας (και ως εκ τούτου της κινητικότητας) των ατόμων. Τα τελευταία χρόνια, μία κατηγορία στοχαστικών μοντέλων, τα Μαρκοβιανά μοντέλα, χρησιμοποιούνται ευρέως για την επίλυση ζητημάτων ανάκλησης πληροφορίας σε συστήματα ακολουθιακών δεδομένων. Συγκεκριμένα, μία ιδιαίτερη κατηγορία Μαρκοβιανών μοντέλων, τα Κρυμμένα Μαρκοβιανά Μοντέλα – Hidden Markov Models (HMMs), έχουν εφαρμοστεί με μεγάλη επιτυχία σε συστήματα αναγνώρισης λόγου, αναγνώρισης μουσικών προτύπων, αναγνώρισης της καταναλωτικής συμπεριφοράς και σε πολλά άλλα είδη ακολουθιακών δεδομένων. Με στόχο την επιτυχή ευρετηρίαση δεδομένων κίνησης θα εφαρμόσουμε βασισμένη σε μοντέλο κατηγοριοποίηση αναπαραστώντας κάθε κλάση μίας βάσης δεδομένων κίνησης με ένα HMM.

Abstract

During the last decade, the domain of mobility data management and mining has emerged, providing many effective methods for mining intuitive patterns that represent collective behavior of trajectories of moving objects. An interesting research line is that instead of operating on the raw data collected from various sensors, researchers make use of semantically enriched data, which are either declared (by the users) or inferred (by some annotation method). This way raw trajectory data is transformed into spatio-temporal-textual sequences, where the extra textual information that is added to the dimensions of space and time, represent the movement semantics. Such spatio-temporal-textual sequences form a more realistic representation model of the complex every-day life (and as such the mobility) of individuals. The last years, there is a mainstream using stochastic models in information retrieval systems of sequential data. Specifically, a particular type of Markov models, named Hidden Markov Models (HMMs) have been successfully applied in speech recognition, music pattern recognition, consumer pattern recognition and many other domains of sequential data. Aiming to achieve high accuracy in indexing mobility data we will apply a model-based classification by representing each class of a mobility database via a HMM.

Περιεχόμενα

Κ Ε Φ Α Λ Α Ι Ο 1

Εισαγωγή

Σκοπός της παρούσας διπλωματικής εργασίας είναι μέσω μίας στοχαστικής μοντελοποίησης να επιτύχουμε αποτελεσματική ευρετηρίαση σημασιολογικών δεδομένων κίνησης (*Model-Based Indexing of Mobility data*). Η επίλυση του συγκεκριμένου ερευνητικού προβλήματος παρουσιάζει εξαιρετικό ενδιαφέρον τόσο ως προς τον τελικό του στόχο, που ειναι η επιτυχής ευρετηρίαση όσο και ως προς την μεθοδολογία που θα ακολουθηθεί. Η ομαδοποίηση, κατηγοριοποίηση και ευρετηρίαση δεδομένων κίνησης που προέρχονται από συσκευές GPS αλλά και άλλες φορητές συσκευές αποτελεί ενα σύγχρονο ερευνητικό πεδίο το οποίο ελκύει ολοένα και περισσότερους ερευνητές καθώς έχει πολλές εφαρμογές σε πολλούς διαφορετικούς τομείς στις μέρες μας. Η ανάλυση δεδομένων κίνησης τουριστών για τη δημιουργία τουριστικών καταλόγων, η ανάλυση δεδομένων κίνησης οχημάτων για την παροχή άμεσης και κατάλληλης οδικής βοήθειας, η συνεχής παρακολούθηση της κίνησης σε συνδυασμό με την καταγραφή βασικών μέτρων υγείας ατομών ευπαθών ομάδων (ηλικιωμένων ή ατόμων με χρόνιες παθήσεις) για άμεση προσφορά ιατρικής περίθαλψης αποτελούν ελάχιστα παραδείγματα εφαρμογής του συγκεκριμένου ερευνητικού πεδίου.

Ωστόσο, πέρα από την σπουδαιότητα του καθεαυτού σκοπού της εργασίας, εξαιρετικό ενδιαφέρον παρουσιάζει η στοχαστική μοντελοποίηση με την οποία θα προσεγγίσουμε το πρόβλημα. Συγκεκριμένα, θα προσπαθήσουμε να εφαρμόσουμε μία στοχαστική διαδικασία Markov αναζητώντας το πλέον κατάλληλο μοντέλο από την ευρύτερη οικογένεια των Μαρκοβιανών μοντέλων. Η χρήση στοχαστικών μοντέλων Markov έχει ήδη εφαρμοστεί με επιτυχία σε πολλά προβλήματα ανάλυσης ακολουθιακών δεδομένων όπως της αναγνώρισης της αγοραστικής συμπεριφοράς των καταναλωτών, της αναγνώρισης μουσικών προτύπων, στην αναγνώριση λόγου, στην αναγνώριση βίντεο και σε πολλά άλλα ερευνητικά πεδία. Η αποτελεσματικότητα των Μαρκοβιανών μοντέλων στην μοντελοποίηση των ιδιαίτερων δομών ποικιλόμορφων συστημάτων ακολουθιακών δεδομένων έχει ανάγει τα ζητήματα της εφαρμογής τους σε συγκεκριμένα συστήματα αλλά και της βελτίωσης των ήδη υπαρχόντων μοντέλων σε καίρια ζητήματα ενασχόλησης πολλών ερευνητικών εργασιών.

Στην επόμενη ενότητα θα παρουσιάσουμε ορισμένα στοχαστικά μοντέλα της οικογένειας των Μαρκοβιανών μοντέλων που έχουν ήδη εφαρμοστεί με πολύ ικανοποιητικά αποτελέσματα σε συγκεκριμένα ερευνητικά πεδία. Αρχικά, θα δοθεί το θεωρητικό υπόβαθρο και τα βασικά αξιώματα κάθε διαφορετικού τύπου Μαρκοβιανού μοντέλου. Στην τρίτη ενότητα θα δοθεί αναλυτική περιγραφή των δεδομένων της ανάλυσης και του τρόπου με τον οποίο αυτά ομαδοποιούνται και επίσης θα παρουσιάσουμε την προσέγγιση του [8] στην οποία πραγματοποιήθηκε εφαρμογή Μαρκοβιανής μοντελοποίησης για την κατηγοριοποίηση χρονοσειρών και αποτελεί τον οδηγό της εκπόνησης της παρούσας εργασίας. Συγκρίνοντας

τα πλεονεκτήματα και μειονεκτήματα κάθε μοντέλου και λαμβάνοντας υπόψη τα ιδιαίτερα χαρακτηριστικά του συστήματος των κινούμενων χρηστών, θα αναζητήσουμε το πλέον κατάλληλο μοντέλο που θα μπορούσε να εφαρμοστεί στο πρόβλημά μας και θα παρουσιάσουμε το σύνολο των θεωρήσεων και ζητημάτων που αφορούν σε αυτή την μοντελοποίηση. Τέλος, θα δώσουμε τα πειραματικά αποτελέσματα της βασισμένης σε μοντέλο κατηγοριοποίησης και κάποια γενικά συμπεράσματα για την εφαρμογή της συγκεκριμένης κατηγορίας στοχαστικών μοντέλων στην ανάλυση της κινητικότητας των ατόμων.

Κ Ε Φ Α Λ Α Ι Ο 2

Μοντέλα Markov

2.1 Εισαγωγή

Ένα σύστημα αποτελεί μία συνάθροιση, συλλογή οντοτήτων/αντικειμένων, υλικών ή αφηρημένων, τα οποία αποτελούν σύνολο και το κάθε στοιχείο αλληλεπιδρά ή συσχετίζεται με τουλάχιστον ένα ακόμη στοιχείο του συνόλου. Το σώμα μας δουλεύει σαν σύστημα ενώ όλοι είμαστε μέλη πολλών κοινωνικών συστημάτων όπως π.χ. της οικογένειάς μας, της Ακαδημαϊκής κοινότητας κ.λπ. Η ανάλυση και κατανόηση της δομής και λειτουργίας τέτοιων συστημάτων αποτελεί αντικείμενο έρευνας για πολλές επιστήμες. Από την πλευρά των μαθηματικών η μελέτη των συστημάτων γίνεται με τη χρήση των μαθηματικών μοντέλων. Ένα μαθηματικό μοντέλο αποτελεί μία προσομοίωση του πραγματικού κόσμου στο οποίο οι σχέσεις μεταξύ των οντοτήτων του συστήματος εκφράζονται με παρόμοιες σχέσεις μεταξύ των μαθηματικών οντοτήτων [1].

Τα μαθηματικά μοντέλα διακρίνονται σε *προσδιοριστικά (deterministic)* και *στοχαστικά (stochastic)*. Αν οι συνέπειες οποιασδήποτε αλλαγής στο σύστημα μπορούν να προβλεφθούν με βεβαιότητα τότε το μοντέλο ονομάζεται προσδιοριστικό. Αν οι αλλαγές στο σύστημα μπορούν να εκφραστούν μαθηματικά μόνο με τυχαίες μεταβλητές τότε το μοντέλο ονομάζεται στοχαστικό. Τα σημαντικότερα ίσως στοχαστικά μοντέλα στις σύγχρονες εφαρμογές είναι αυτά που μπορούν να τεθούν κάτω από το γενικό τίτλο *Μαρκοβιανές Αλυσίδες*. Κοινό χαρακτηριστικό αυτών των μοντέλων είναι ότι έχουν την Μαρκοβιανή ιδιότητα, σύμφωνα με την οποία η μελλοντική εξέλιξη του συστήματος εξαρτάται από την παρούσα του κατάσταση και δεν εξαρτάται από το παρελθόν του.

2.2 Μαρκοβιανές Αλυσίδες

Μία στοχαστική διαδικασία με την ιδιότητα, ότι δεδομένης της τιμής X_t, οι τιμές των X_s, για $s > t$, δεν εξαρτώνται από τις τιμές των X_r, για $r < t$, καλείται μία διαδικασία Markov [1]. Η ιδιότητα αυτή σημαίνει ότι η πιθανότητα οποιασδήποτε μελλοντικής κατάστασης της διαδικασίας, όταν η παρούσα κατάσταση είναι γνωστή, δεν αλλοιώνεται από τις παρελθοντικές καταστάσεις της διαδικασίας. Η ιδιότητα αυτή είναι γνωστή σαν ιδιότητα του Markov (1907). Η κατηγορία των Μαρκοβιανών διαδικασιών όταν ο χρόνος και ο χώρος των καταστάσεων είναι διακριτός ονομάζονται *Μαρκοβιανές αλυσίδες*. Έτσι, μπορούμε να ορίσουμε μία Μαρκοβιανή αλυσίδα σαν μία ακολουθία $X_0, X_1, X_2, \ldots$ διακεκριμένων τυχαίων μεταβλητών με την ιδιότητα ότι η υπό συνθήκη κατανομή της X_{n+1} όταν δίνονται οι $X_0, X_1, \ldots, X_n$ εξαρτάται μόνο από την τιμή της X_n δηλ.:

$$P(X_{n+1} = k|X_n = r, X_{n-1} = z, \dots, X_1 = c, X_0 = b) = P(X_{n+1} = k|X_n = r)$$

Οι Μαρκοβιανές αλυσίδες οι οποίες παρουσιάζουν εξάρτηση μόνο από την αμέσως προηγούμενη χρονική στιγμή, καλούνται *πρώτης τάξης*.

Έστω, τώρα, $S_i, i = 1, 2, \dots, M$ ο (πεπερασμένος) χώρος των καταστάσεων μίας Μαρκοβιανής αλυσίδας και έστω $X_t, t = 1, 2, \dots$ η Μαρκοβιανή αλυσίδα που παίρνει τιμές στο χώρο των καταστάσεων S. Ορίζουμε τις υπό συνθήκη πιθανότητες:

$$a_{ij}(t) = P(X_t = j|X_{t-1} = i), \gamma\iota\alpha\ i, j = 1, 2, \dots, M\ \kappa\alpha\iota\ t = 1, 2, \dots$$

Οι πιθανότητες $a_{ij}(t)$ είναι οι πιθανότητες η Μαρκοβιανή αλυσίδα X_t να μεταβεί στην κατάσταση j δεδομένου ότι την προηγούμενη χρονική στιγμή ήταν στην κατάσταση i και ονομάζονται *πιθανότητες μετάβασης* της Μαρκοβιανής αλυσίδας. Ο πλέον συνηθισμένος τρόπος παρουσίασής τους είναι με χρήση ενός πίνακα $A(t)$. Ο πίνακας $A(t)$ ονομάζεται *πίνακας μετάβασης* της Μαρκοβιανής αλυσίδας για το χρονικό διάστημα $[t-1, t)$. Αν οι πιθανότητες $a_{ij}(t)$ είναι συναρτήσεις της χρονικής στιγμής, δηλαδή μεταβάλλονται συναρτήσει του χρόνου t κατά τη διάρκεια της Μαρκοβιανής διαδικασίας, τότε χρειαζόμαστε τις τιμές τους για κάθε χρονική στιγμή t. Με βάση αυτό το γνώρισμα, διακρίνουμε τις εξής δύο περιπτώσεις:

Μια Μαρκοβιανή αλυσίδα καλείται *στατική (stationary) ή ομογενής (homogeneous)* αν η πιθανότητα μετάβασης από τη μία κατάσταση στην άλλη είναι ανεξάρτητη του χρόνου που πραγματοποιείται η μετάβαση. Έτσι, για όλες τις καταστάσεις i και j μίας στατικής Μαρκοβιανής αλυσίδας έχουμε:

$$P(X_t = j|X_{t-1} = i) = a_{ij}, \gamma\iota\alpha\ \kappa\acute{\alpha}\theta\varepsilon\ t = 1, 2, \dots$$

Όταν μία Μαρκοβιανή αλυσίδα δεν είναι στατική/ομογενής τότε ονομάζεται *μή-στατική (non-stationary) ή μή-ομογενής (non-homogeneous).* Από τα παραπάνω γίνεται εύκολα αντιληπτό ότι σε μία ομογενή Μαρκοβιανή αλυσίδα έχουμε μόνο έναν πίνακα μετάβασης A ενώ σε μία μή ομογενή έχουμε μία ακολουθία πινάκων μετάβασης $\{A(t)\}_{t=0}^{\infty}$ (ή $\{A(t)\}_{t=0}^{T}$ για πεπερασμένου χρόνου Μαρκοβιανή αλυσίδα). Οι χαρακτηριστικές ιδιότητες του πίνακα μετάβασης A είναι ότι όλα τα στοιχεία του είναι θετικά ή μηδέν και το άθροισμα κάθε γραμμής του είναι ίσο με τη μονάδα. Κάθε πίνακας που έχει αυτές τις ιδιότητες καλείται *στοχαστικός*.

2.3 Απλά Μοντέλα Markov

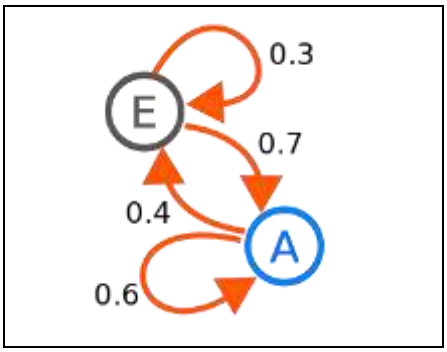
Το σύνολο των καταστάσεων και των μεταξύ τους συσχετίσεων μίας Μαρκοβιανής αλυσίδας αποτελούν ένα απλό μοντέλο Markov. Ένα μοντέλο Markov λ ορίζεται από [1]:

- Το σύνολο Μ των καταστάσεων της διαδικασίας $S_i, i = 1, 2, \dots, M$. Η κατάσταση του μοντέλου τη χρονική στιγμή t συμβολίζεται s_t.
- Τον πίνακα $A = \{a_{ij}\}$ των πιθανοτήτων μετάβασης του μοντέλου, όπου a_{ij} είναι η πιθανότητα να μεταβούμε από την κατάσταση i στην κατάσταση j:
$$a_{ij} = P(s_{t+1} = j | s_t = i), για\ 1 \leq i, j \leq M$$
- Τον πίνακα $\pi = \{\pi_i\}$ των πιθανοτήτων έναρξης, δηλαδή την πιθανότητα η κατάσταση i να είναι η αρχική κατάσταση της αλυσίδας:
$$\pi_i = P(s_1 = i), για\ κάθε\ i = 1, 2, \dots, M$$

Ο πίνακας π είναι επίσης ένας στοχαστικός πίνακας καθώς αποτελείται από θετικά ή μή μηδενικά στοιχεία το άθροισμα των οποίων ισούται με την μονάδα.

Επομένως, το σύνολο των παραμέτρων του απλού μοντέλου Markov είναι οι πίνακες A και π με την βοήθεια των οποίων ορίζεται ως $\lambda = (A, \pi)$. Όπως φαίνεται από τον ορισμό του (ανεξάρτητου του χρόνου) μοναδικού πίνακα μετάβασης μεταξύ των καταστάσεών του, το απλό μοντέλο Markov περιγράφει μία στατική Μαρκοβιανή αλυσίδα.

Στην Εικόνα 1, που ακολουθεί, δίνεται το γράφημα ενός απλού Μαρκοβιανού μοντέλου που αποτελείται από δύο καταστάσεις.



Εικόνα 1. Απλό μοντέλο Markov δύο καταστάσεων.

2.4 Κρυμμένα Μαρκοβιανά Μοντέλα – Hidden Markov Models (HMMs)

Στα απλά μοντέλα Markov το πλήθος των καταστάσεων του μοντέλου θεωρείται ότι είναι ίσο με το πλήθος των διακριτών παρατηρήσεων καθώς κάθε κατάσταση του μοντέλου αντιστοιχεί σε μία μόνο παρατήρηση. Το σύνολο, επομένως, των παρατηρήσεων είναι άμεσα ορατό και συνεπώς οι πιθανότητες μετάβασης είναι οι μόνοι παράμετροι του μοντέλου. Ωστόσο, σε πολλά προβλήματα συναντάμε πολύ πιο σύνθετες δομές τις οποίες τα απλά μοντέλα Markov δεν μπορούν να μοντελοποιήσουν ικανοποιητικά. Επέκταση των απλών μοντέλων Markov, που είναι ήδη εφαρμοσμένα σε πολλά ερευνητικά ζητήματα αναγνώρισης χρονικών

προτύπων όπως η αναγνώριση λόγου, χειρονομιών κ.ά., αποτελούν τα Κρυμμένα Μοντέλα Markov (Hidden Markov Models – HMMs) των οποίων τα κύρια χαρακτηριστικά θα παρουσιάσουμε στη συνέχεια.

Ένα HMM ορίζεται ως μία διπλή στοχαστική διαδικασία. Περιγράφει την από κοινού πιθανότητα ενός πεπερασμένου συνόλου από μή άμεσα ορατές (εξού και κρυμμένες ή hidden) καταστάσεις και ενός συνόλου από παρατηρούμενες διακριτές μεταβλητές. Όταν είμαστε σε μία κατάσταση, ένα σύμβολο/τιμή μπορεί να παρατηρηθεί σύμφωνα με μία, εξαρτώμενη της καταστάσης που βρισκόμαστε, κατανομή πιθανότητας [5].

Για να γίνει ευκολότερα κατανοητό το κρυμμένο μοντέλο παραθέτουμε το εξής παράδειγμα όπως αυτό δίνεται στο [6]. Είμαστε κλεισμένοι σε ένα δωμάτιο για αρκετές ημέρες και θέλουμε να μάθουμε τον καιρό έξω. Θεωρούμε ότι οι καιρικές συνθήκες περιγράφονται πλήρως από ένα σύνολο τριών συνθηκών {Ηλιόλουστος, Συννεφώδης, Ομιχλώδης}, το οποίο αποτελεί, ουσιαστικά, τον χώρο των μή παρατηρούμενων (hidden) καταστάσεων S. Καθώς δεν έχουμε καμία επαφή με τον εξωτερικό κόσμο το μόνο στοιχείο που μπορούμε να παρατηρήσουμε για να αντιληφθούμε τις καιρικές συνθήκες, είναι αν το άτομο που φέρνει καθημερινά το γεύμα μας έχει μαζί του ομπρέλα ή όχι. Επομένως, μόνη παρατηρούμενη μεταβλητή είναι το αν ο διανομέας του φαγητού έχει μαζί του ομπρέλα ή όχι, η οποία έχει μία εξαρτώμενη των καιρικών συνθηκών κατανομή πιθανότητας. Γνωρίζοντας τις πιθανότητες μετάβασης μεταξύ των καιρικών συνθηκών (απο Ηλιόλουστο σε Συννεφώδη κ.λπ.) αλλά και την πιθανότητα να έχει ο διανομέας μαζί του ομπρέλα, ανάλογα των καιρικών συνθηκών που επικρατούν, προσπαθούμε να αντιληφθούμε την συμπεριφορά του συστήματος ώστε να μπορέσουμε να απαντήσουμε πιθανοκρατικά για το ποιές καιρικές συνθήκες μπορεί να επικρατούν υπό συγκεκριμένες συνθήκες (όπως π.χ. να προβλέψουμε τί καιρό θα κάνει την επομένη μέρα δεδομένων: του καιρού που είχε την τελευταία μέρα πριν κλειστούμε, των πιθανοτήτων μετάβασης μεταξύ των καιρικών καταστάσεων και των πιθανοτήτων να έχει ο διανομέας ομπρέλα για κάθε μία από τις τρείς καταστάσεις του συστήματος).

Πιο φορμαλιστικά, ένα HMM λ' ορίζεται από [5]:

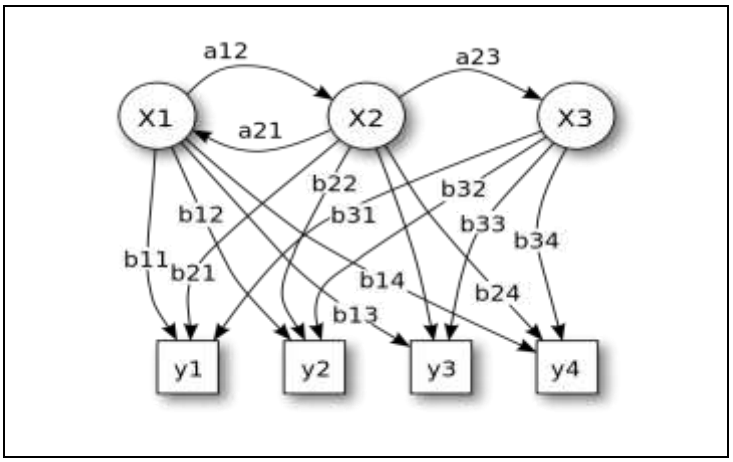
- Το σύνολο $S_i, i = 1, 2, \dots, M$ των M κρυφών καταστάσεων του μοντέλου.
- Τον πίνακα A των πιθανοτήτων μετάβασης μεταξύ των καταστάσεων (όπως ορίστηκε και στο απλό μοντέλο Markov):
$$a_{ij} = P(s_{t+1} = j | s_t = i), για\ 1 \leq i, j \leq M$$
- Το σύνολο O των K παρατηρήσιμων συμβόλων που παράγονται σε κάθε κατάσταση $O = \{o_1, o_2, \dots, o_k\}$.
- Τον πίνακα των πιθανοτήτων εμφάνισης των παρατηρούμενων συμβόλων σε κάθε κατάσταση j, $B = \{b_j(o_k)\}$, όπου:
$$b_j(o_k) = P(o_k | s_t = j), για\ 1 \leq k \leq K\ και\ 1 \leq j \leq M$$
- Τον πίνακα π των πιθανοτήτων έναρξης του μοντέλου (όμοια με το απλό μοντέλο):
$$\pi_i = P(s_1 = i), για\ κάθε\ i = 1, 2, \dots, M$$

Συνεπώς, το σύνολο των παραμέτρων ενός HMM είναι οι πίνακες A, B και π με την βοήθεια των οποίων και ορίζεται ως $\lambda' = (A, B, \pi)$.

Όπως παρατηρούμε, σε ένα HMM δεν μοντελοποιείται η διάρκεια του χρόνου παραμονής στις καταστάσεις του συστήματος. Σε κάθε Μαρκοβιανή αλυσίδα και επομένως και στα HMMs η διάρκεια του χρόνου παραμονής στις καταστάσεις, δηλαδή η πιθανότητα το σύστημα να παραμείνει u διαδοχικές χρονικές στιγμές στην κατάσταση i, δίνεται από την σχέση:

$$d_i(u) = P(S_{t+u+1} \neq i, S_{t+u} = i, S_{t+u-1} = i, \dots, S_{t+2} = i / S_{t+1} = i, S_t \neq i) = p_{ii}^{u-1}(1 - p_{ii})$$

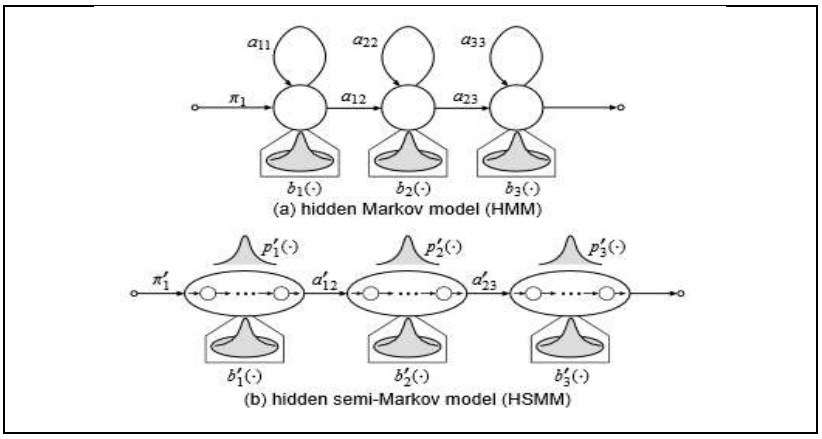
Η ισότητα στην οποία καταλήγουμε είναι ουσιαστικά η συνάρτηση πιθανότητας της Γεωμετρικής κατανομής. Έτσι, για όλες τις Μαρκοβιανές αλυσίδες η διάρκεια παραμονής στις καταστάσεις θεωρείται ότι ακολουθεί την Γεωμετρική κατανομή. Εφόσον μετράμε το πλήθος χρονικών στιγμών μέχρι να πραγματοποιηθεί μετάβαση σε άλλη κατάσταση, η πιθανότητα επιτυχίας δηλαδή μετάβασης θα είναι η συμπληρωματική της πιθανότητας παραμονής στην συγκεκριμένη κατάσταση δηλαδή η ποσότητα $(1 - p_{ii})$. Στην περίπτωση πρώτης τάξης μοντέλου η πιθανότητα αυτή ταυτίζεται με την πιθανότητα μετάβασης στην αμέσως επόμενη κατάσταση. Σύμφωνα με την μέση τιμή της Γεωμετικής κατανομής η μέση διάρκεια για την i κατάσταση θα δίνεται από την σχέση $E(u) = \frac{1}{(1-p_{ii})}$. Επομένως, στα HMMs όλη η πληροφορία για την χρονική διάρκεια των καταστάσεων προκύπτει από τον πίνακα μετάβασης *A*. Επισημαίνουμε ότι η στατική ιδιότητα εξακολουθεί να χαρακτηρίζει και τα HMMs καθώς ο πίνακας μετάβασης *A* είναι ανεξάρτητος του χρόνου. Στην Εικόνα 2, που ακολουθεί, δίνεται το γράφημα ενός HMM τριών κρυφών καταστάσεων και τεσσάρων παρατηρούμενων τιμών.



Εικόνα 2. Hidden Markov Model τριών κρυφών καταστάσεων και τεσσάρων παρατηρούμενων συμβόλων/τιμών.

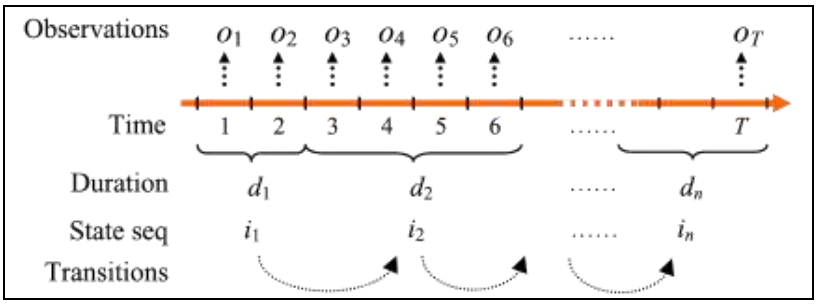
2.5 Κρυμμένα Ημι-Μαρκοβιανά Μοντέλα – Hidden Semi-Markov Models (HSMMs)

Σε πολλές εφαρμογές η θεώρηση της Γεωμετρικής κατανομής του χρόνου παραμονής στις καταστάσεις, που τα HMMs υιοθετούν, περιορίζει τις δυνατότητες και την απόδοση του μοντέλου καθώς δεν παρέχει επαρκή αναπαράσταση της χρονικής δομής του συστήματος. Ένα hidden semi-Markov model (HSMM) είναι μία επέκταση ενός HMM που επιτρέπει την υποκείμενη (κρυφή) διαδικασία να είναι μία ημί-Μαρκοβιανή (semi-Markov) διαδικασία, να έχει, δηλαδή, μεταβλητή διάρκεια χρόνου παραμονής σε κάθε κατάσταση και όχι Γεωμετρική όπως στο απλό μοντέλο και στο HMM. Η σημαντική διαφορά ανάμεσα σε ένα HMM και ένα HSMM είναι ότι σε ένα HMM μπορεί να παρατηρηθεί μία παρατήρηση σε κάθε κατάσταση αντίθετα με το HSMM όπου μπορούμε να παρατηρήσουμε μία ακολουθία παρατηρήσεων ανάλογα με το χρόνο που το σύστημα παρέμεινε στη συγκεκριμένη κατάσταση. Ο όρος ημι-Μαρκοβιανό έγκειται, επομένως, στο γεγονός ότι το σύστημα χάνει την Μαρκοβιανή ιδιότητα όσο παραμένει σε μία κατάσταση καθώς θα παραμείνει εκεί για διάρκεια εξαρτώμενη της παρούσας κατάστασης ενώ είναι Μαρκοβιανό τις χρονικές στιγμές που πραγματοποιεί μεταβάσεις σε επόμενες καταστάσεις. Ένα HSMM επιτρέπει την μοντελοποίηση του γεγονότος ότι ένας ταξιδιώτης αεροπλάνου π.χ. θα περάσει αρκετό χρόνο στην περιοχή ελέγχου ασφαλείας ενώ αρκετά λιγότερο χρόνο για να κινηθεί από αυτήν την περιοχή στην περιοχή επιβίβασης. Η διαφορά στην δομή ενός συμβατικού HMM και ενός HSMM απεικονίζεται στην Εικόνα 3 που ακολουθεί:



Εικόνα 3. Παράδειγμα ενός HMM και ενός HSMM τριών κρυφών καταστάσεων.

Το σύνολο των παραμέτρων ενός HSMM λ'' είναι οι πίνακες *A, B* και π ακριβώς όπως ορίστηκαν για το HMM και επιπλέον έχουμε την πληροφορία για την κατανομή της διάρκειας παραμονής στις καταστάσεις η οποία συμβολίζεται με *D*. Έτσι, με την βοήθεια των παραμέτρων του ένα HSMM συμβολίζεται ως $\lambda'' = (A, B, D, \pi)$. Η στατική ιδιότητα εξακολουθεί να χαρακτηρίζει και τα HSMMs μιας και ο πίνακας μετάβασης *A* παραμένει ανεξάρτητος του χρόνου. Στην Εικόνα 4, που δίνεται ακόλουθα, παρουσιάζεται ένα γενικό HSMM.



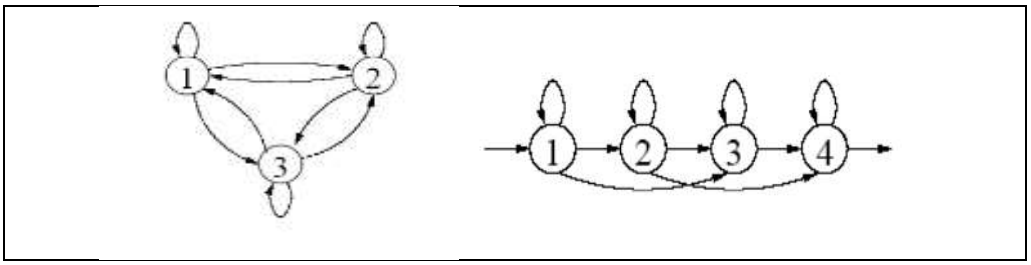
Εικόνα 4. Γενικό Hidden Semi-Markov Model.

Ένα γενικό HSMM διαιρείται σε συγκεκριμένα HSMM μοντέλα σύμφωνα με τις υποθέσεις που υιοθετεί το καθένα. Ένα από τα πιο απλά, σύμφωνα με τις υποθέσεις ανεξατησίας που θεωρεί, HSMM μοντέλα και παράλληλα πιο δημοφιλές στις εφαρμογές είναι το Explicit Duration HMM (EDHMM). Σε ένα EDHMM η μετάβαση σε μία κατάσταση είναι ανέξαρτητη της διάρκειας της προηγούμενης χωρίς να υπάρχουν πιθανότητες παραμονής στην ίδια κατάσταση. Η διάρκεια παραμονής θεωρείται εξαρτώμενη της τωρινής κατάστασης και ανεξάρτητη της προηγούμενης. Έτσι, η τυχαία κατάσταση j θα διαρκεί για μεταβλητή διάρκεια d (όπου d θα παίρνει τιμές σε ένα υποσύνολο των ακεραίων □) σύμφωνα με μία υπό συνθήκη πιθανότητα $p_j(d)$. Αν θεωρήσουμε την κατανομή της διάρκειας παραμονής στις καταστάσεις $p_j(d)$ να είναι Γεωμετρική τότε το EDHMM γίνεται ισοδύναμο με το συμβατικό HMM [5].

2.6 Τύποι HMMs

Ορίσαμε το σύνολο των παραμέτρων που περιγράφουν πλήρως ένα HMM. Ωστόσο, τα HMMs και κατ' επέκταση τα HSMMs διακρίνονται με βάση τις σχέσεις μεταξύ των κρυφών τους καταστάσεων [5].

Ένα HMM ονομάζεται *εργοδικό (ergodic)* ή *πλήρως διασυνδεδεμένο* όταν το σύστημα μπορεί να μεταβεί από κάθε κατάσταση σε οποιαδήποτε άλλη σε έναν πεπερασμένο αριθμό βημάτων. Ένας δεύτερος τύπος HMM που χρησιμοποιείται συχνά είναι αυτός του *αριστερού-δεξιού (left-right)* ή *Bakis* μοντέλου. Σε ένα μοντέλο Bakis το σύστημα μπορεί είτε να παραμείνει στην ίδια κατάσταση είτε να μεταβεί σε κάποια από τις επόμενες, μή έχοντας την δυνατότητα να επιστρέψει σε κάποια προηγούμενη κατάσταση, κινούμενο με αυτόν τον τρόπο από τα αριστερά πρός τα δεξία. Πρόσθετα, σε ένα μοντέλο Bakis οι πιθανότητες έναρξης έχουν την ιδιότητα ότι το σύστημα ξεκινάει υποχρεωτικά από την πρώτη κατάσταση, δηλαδή $\pi_i = \begin{Bmatrix} 0, i \neq 1 \\ 1, i = 1 \end{Bmatrix}$. Φυσικά, έχουν αναπτυχθεί και αρκετοί ακόμα τύποι HMM ανάλογα με τις ιδιαιτερότητες του συστήματος κάθε εφαρμογής. Στην Εικόνα 5 δίνονται τα γραφήματα ενός εργοδικού και ενός αριστερού-δεξιού μοντέλου.



Εικόνα 5. Γράφημα ενός εργοδικού ή πλήρως διασυνδεδεμένου μοντέλου (αριστερά) και ενός αριστερού-δεξιού ή Bakis μοντέλου (δεξιά).

2.7 Βασικά Ζητήματα για HMMs

Τρία είναι τα βασικά ζητήματα που έχουμε να αντιμετωπίσουμε στα HMMs, καθώς και στις επεκτάσεις αυτών [5]:

- *Εκτίμηση (Evaluation ή Classification)*: Δεδομένων της ακολουθίας παρατηρήσεων $O = o_1 o_2 \dots o_T$ και ενός HMM λ, αναζητούμε ποια είναι η πιθανότητα της ακολουθίας των παρατηρήσεων δεδομένου του μοντέλου, δηλαδή την πιθανότητα $P(O|\lambda)$.
- *Αναγνώριση (Decoding ή Recognition)*: Δεδομένων της ακολουθίας παρατηρήσεων $O = o_1 o_2 \dots o_T$ και ενός HMM λ, αναζητούμε ποια είναι η ακολουθία από κρυφές καταστάσεις $S = s_1 s_2 \dots s_M$ η οποία είναι πιο πιθανό να παράγει τη δεδομένη ακολουθία παρατηρήσεων.
- *Εκπαίδευση (Learning ή Training)*: Πώς προσαρμόζουμε τις παραμέτρους του μοντέλου ώστε να μεγιστοποιήσουμε την πιθανοφάνεια $P(O|\lambda)$.

Αρκετοί διαφορετικοί αλγόριθμοι έχουν αναπτυχθεί για την επίλυση των παραπάνω ζητημάτων. Στη συνέχεια παρουσιάζουμε τους σημαντικότερους και συχνότερα εφαρμοσμένους στην βιβλιογραφία [5].

Για να λύσουμε το ζήτημα της *εκτίμησης* κανονικά θα έπρεπε να απαριθμήσουμε κάθε πιθανή ακολουθία καταστάσεων μήκους Τ. Ωστόσο, αυτό είναι υπολογιστικά πολύ βαρύ ή και ακατόρθωτο. Για την επίλυση του συγκεκριμένου προβλήματος έχει αναπτυχθεί ο αλγόριθμος Forward, ο οποίος βασίζεται στο δυναμικό προγραμματισμό.

Ο στόχος του ζητήματος της *αναγνώρισης* είναι να βρεθεί η βέλτιστη ακολουθία κρυφών καταστάσεων που σχετίζονται με τη δεδομένη ακολουθία παρατηρήσεων. Το πιο ευρέως χρησιμοποιούμενο κριτήριο βελτιστοποιήσης είναι να βρούμε εκείνη την ακολουθία καταστάσεων για την οποία μεγιστοποιείται η πιθανοφάνεια $P(S|O,\lambda)$ το οποίο είναι ισοδύναμο με το να μεγιστοποιήσουμε την $P(S,O|\lambda)$. Ο αλγόριθμος *Viterbi* χρησιμοποιείται για να βρεθεί αυτή η βέλτιστη ακολουθία κρυφών καταστάσεων, ο οποίος βασίζεται, επίσης, σε μεθόδους του δυναμικού προγραμματισμού.

Για το ζήτημα της *εκπαίδευσης* δεν υπάρχει καμία γνωστή μέθοδος για να βρεθεί αναλυτική λύση. Ωστόσο, μπορούμε να προσαρμόσουμε τις παραμέτρους του μοντέλου ώστε η πιθανοφάνεια $P(O|\lambda)$ να μεγιστοποιείται τοπικά με χρήση μίας επαναληπτικής διαδικασίας όπως ο αλγόριθμος *Baum – Welch* (ή ισοδύναμα ο αλγόριθμος *Expectation – Maximization*).

Κ Ε Φ Α Λ Α Ι Ο 3

Σχετικές Εργασίες

3.1 Περιγραφή Δεδομένων

Αρχικά, θα παρουσιάσουμε τους βασικούς ορισμούς που θα βοηθήσουν στην κατανόηση των ακατέργαστων τροχιών (*raw trajectories*) και των σημασιολογικών χρονοδιαγραμμάτων κίνησης (*semantic mobility timelines*) όπως δίνονται στο [2]. Ακόλουθα θα δούμε τον τρόπο που πραγματοποιείται η ομαδοποίηση των σημασιολογικών τροχιών [4] και τέλος θα παρουσιάσουμε την βασισμένη σε μοντέλο κατηγοριοποίηση μεγάλων βάσεων δεδομένων χρονοσειρών που εφαρμόστηκε στο [8] και αποτελεί τον οδηγό της δικής μας προσέγγισης.

Ακατέργαστη τροχιά: ως ακατέργαστη τροχιά (*raw trajectory*) τ ενός κινούμενου αντικειμένου ορίζεται μία τριάδα της μορφής *(o-id, traj-id, T)*, όπου: *o-id (traj-id)* είναι το αναγνωριστικό του κινούμενου αντικειμένου (της συγκεκριμένης τροχιάς του κινούμενου αντικειμένου, αντίστοιχα) και T είναι μια τρισδιάστατη πολυγραμμή αποτελούμενη από $N+1$ ζεύγη της μορφής (p_i, t_i), $0 \leq i \leq N$, θεωρώντας γραμμική παρεμβολή ανάμεσα σε δύο διαδοχικά ζεύγη (p_i, t_i) και (p_{i+1}, t_{i+1}), όπου p_i είναι ένα δισδιάστατο σημείο (x_i, y_i) στο επίπεδο και t_i είναι ο αντίστοιχος χρόνος που το κινουμένο αντικείμενο βρέθηκε στο σημείο p_i.

Ακατέργαστη υποτροχιά: ως ακατέργαστη υποτροχιά (*raw sub-trajectory*) τ' μιας ακατέργαστης τροχιάς τ ορίζεται μία τετράδα της μορφής *(o-id, traj-id, subtraj-id, T')*, όπου *o-id (traj-id, subtraj-id)* είναι το αναγνωριστικό του κινούμενου αντικειμένου (της συγκεκριμένης τροχιάς και υποτροχιάς του κινούμενου αντικειμένου, αντίστοιχα) και T' είναι το τμήμα του T ανάμεσα σε δύο χρονικές στιγμές, t_i και t_j, $t_i < t_j$.

Stop: ένα *Stop* επεισόδιο κίνησης αντιστοιχεί σε μία υποτροχιά τ' με συγκεκριμένες χωρο-χρονικές ιδιότητες. Συγκεκριμένα, μία υποτροχία τ' επισημαίνεται ως *Stop* αν και μόνον αν η χωρική της (χρονική) προβολή υπακούει έναν προκαθορισμένο χωρικό (χρονικό, αντίστοιχα) περιορισμό C_{space} (C_{time}, αντίστοιχα).

ActivityStop: ένα *ActivityStop* αντιστοιχεί σε μία υποτροχιά τ' με συγκεκριμένες χωρο-χρονικές και θεματικές ιδιότητες. Συγκεκριμένα, μία υποτροχιά τ' επισημαίνεται ως *ActivityStop* αν και μόνον αν η χωρική (χρονική) της προβολή υπακούει έναν προκαθορισμένο χωρικό (χρονικό, αντίστοιχα) περιορισμό C_{space} (C_{time}, αντίστοιχα) και υπακούει έναν προκαθορισμένο θεματικό περιορισμό $C_{thematic}$. Οι θεματικές ιδιότητες υπονοούν τον σκοπό για τον οποίο το κινούμενο αντικείμενο κινήθηκε στην αντίστοιχη χωρική περιοχή.

MeteorStep: ένα MeteorStep αντιστοιχεί σε μία υποτροχιά τ'της οποίας η συνιστώσα T' (δηλαδή η αντίστοιχη τρισδιάστατη πολυγραμμή) είναι άγνωστη ή κενή.

Move: μία υποτροχιά *τ΄* επισημαίνεται ως *Move* αν και μόνον αν δεν είναι *Stop* ή *ActivityStop* ή *MeteorStep*.

Επεισόδιο Κίνησης: ένα επεισόδιο κίνησης (*LifeStep- ls*) αντιστοιχεί σε μία υποτροχιά *τ΄* και ορίζεται ως μία πλειάδα *(LifeStepID, LifeStepFlag, MBB, tags, T-link, Z-In, Z-Out)*, όπου: *LifeStepID* είναι το αναγνωριστικό του *LifeStep, LifeStepFlag* είναι μία flag που παίρνει τιμές στο σύνολο {'Move', 'Stop', 'ActivityStop', 'StopGap', 'MoveGap', 'Stop&Go'}, *MBB* είναι μία πλειάδα *(MBR,* t_{start}*,* t_{end}*)* που αντιστοιχεί στην τρισδιάστατη προσέγγιση της *τ΄*, με *MBR* να είναι το δισδιάστατο περικλείον ορθογώνιο της χωρικής προβολής της *τ΄* στο επίπεδο και *[*t_{start}*,* t_{end}*]* να είναι η μονοδιάστατη χρονική προβολή της τ΄, *tags* είναι ένα σύνολο από λέξεις-κλειδιά που περιγράφουν τις αντίστοιχες δραστηριότητες και τα σημασιολογικά σχόλια που σχετίζονται με αυτό το τμήμα της κίνησης, *T-link* είναι ένας σύνδεσμος στην *τ΄*, *Z-In* και *Z-Out* είναι δείκτες στις *Semantic Mobility Timelines* οι οποίοι επιτρέπουν αναπαράσταση ενός *LifeStep* σε διαφορετικά επίπεδα λεπτομέρειας (π.χ. ένα *ActivityStop* σε ένα εμπορικό κέντρο μπορεί να αναλυθεί σε ένα πιο ιδιαίτερο επίπεδο κοιτώντας τα καταστήματα που επισκέφθηκε ο χρήστης).

Σημασιολογικά Εμπλουτισμένη Τροχιά: η σημασιολογικά εμπλουτισμένη τροχιά (*semantic mobility timeline*) τ_{sem} ενός κινούμενου αντικειμένου ορίζεται ως μία τριάδα *(o-id, timeline-id,* T_{LS}*)*, όπου: *o-id (timeline-id)* είναι το αναγνωριστικό του κινούμενου αντικειμένου (της semantic mobility timeline του κινούμενου αντικειμένου, αντίστοιχα) και T_{LS} είναι μία ακολουθία από *Lifesteps* που ανήκουν στην ίδια τροχιά *τ* και είναι διαδοχικά στο χρόνο, δηλαδή, $s_i[t_{end}] = s_{i+1}[t_{start}]$.

3.2 Ομαδοποίηση Σημασιολογικά Εμπλουτισμένων Τροχιών

Η εκ των προτέρων γνώση της κλάσης στην οποία ανήκουν τα δεδομένα είναι απαραίτητη ώστε να μπορέσουμε να εφαρμόσουμε την βασισμένη σε μοντέλο κατηγοριοποίηση με την αναπαράσταση κάθε κλάσης των δεδομένων εκπαίδευσης μέσω ενός HMM. Στη συνέχεια, θα παρουσιάσουμε τον τρόπο με τον οποίο πραγματοποιείται η ομαδοποίηση των σημασιολογικών χρονοδιαγραμμάτων κίνησης όπως αυτή παρουσιάζεται στο [4]. Το βήμα της ομαδοποίησης στην παρούσα εφαρμογή θα είναι δεδομένο, δηλαδή τα δεδομένα εκπαίδευσης θα συνοδεύονται από την κλάση στην οποία ανήκει ο κάθε χρήστης.

Όπως είδαμε στην περιγραφή των δεδομένων, η σημασιολογικά εμπλουτισμένη τροχιά ενός χρήστη είναι μία ακολουθία από επεισόδια κίνησης (*LifeSteps*) ενώ κάθε επεισόδιο χαρακτηρίζεται από ένα ζεύγος τιμών, το οποίο συμβολίζουμε ως *(θ, k)*, όπου *θ* είναι η χωροχρονική τιμή που μας παρέχει μία προσέγγιση του τμήματος της κίνησης του χρήστη (δηλαδή το *MBB* που ορίσαμε πιο πάνω) και *k* είναι η αντίστοιχη κειμενική περιγραφή (δηλαδή το σύνολο λέξεων-κλειδιών των μεταβλητών *tags*). Το ζήτημα της ομαδοποίησης των σημασιολογικά εμπλουτισμένων τροχιών αποτελεί την διαμέριση μίας βάσης δεδομένων *SMD* σε συστάδες (*clusters*) έτσι ώστε κάθε συστάδα να περιέχει όμοιες τροχιές σύμφωνα με κάποια συνάρτηση απόστασης. Όπως σημειώνεται, δύο τροχιές μπορεί να είναι όμοιες με αρκετούς διαφορετικούς τρόπους ανάλογα με την εκάστοτε εφαρμογή και τον στόχο της ανάλυσης. Έτσι, δύο τροχιές μπορεί να συμπίπτουν πλήρως ή μερικώς στο χώρο, να έχουν

κοινά χρονικά σημεία έναρξης και/ή λήξης, να είναι πλήρως ή μερικώς συγχρονισμένες, να είναι ασύγχρονες αλλά να παρουσιάζουν όμοια συμπεριφορά (δηλαδή ως προς τις *tags* μεταβλητές) κ.λπ.

Ακολουθώντας την προσέγγιση του βασισμένου στην πυκνότητα αλγορίθμου ομαδοποίησης T-OPTICS, στο [4] προτείνεται μία νέα μετρική συνάρτηση απόστασης η οποία μοντελοποιεί την ανομοιότητα δύο σημασιολογικών τροχιών και βασίζεται στην απόσταση μεταξύ δύο επεισοδίων (*LifeSteps*). Έτσι, ο συγκεκριμένος αλγόριθμος ομαδοποίησης, ο οποίος καλείται SemT-OPTICS, μπορεί να εφαρμοστεί τόσο σε ολόκληρες τις σημασιολογικές τροχιές όσο και μεταξύ των επεισοδίων επιλέγοντας την κατάλληλη μετρική.

Η συνάρτηση απόστασης D_{LS} μεταξύ δύο επεισοδίων ls_i *και* ls_j ορίζεται ως

$$D_{LS}(ls_i, ls_j) = \lambda dist_\theta(ls_i, ls_j) + (1-\lambda) dist_k(ls_i, ls_j)$$

Όπου η συνάρτηση απόστασης $dist_\theta$ εκφράζει την χωροχρονική απόσταση μεταξύ των δύο επεισοδίων και η συνάρτηση απόστασης $dist_k$ την κειμενική απόσταση των δύο επεισοδίων. Η παράμετρος $\lambda \in [0, 1]$ χρησιμοποιείται για να εκφράσει την σχετισμένη σημαντικότητα μεταξύ των δύο στοιχείων και στην ομαδοποίηση των δεδομένων της ανάλυσης μας τίθεται ίση με 0.5, δηλαδή ίση για το χωροχρονικό και κειμενικό στοιχείο.

Με βάση τον ορισμό της μετρικής συνάρτησης απόστασης D_{LS} μεταξύ δύο επεισοδίων , η απόσταση D_{MT} μεταξύ δύο σημασιολογικά εμπλουτισμένων τροχιών mt_i *και* mt_j ορίζεται ως

$$D_{MT}(mt_i, mt_j) = min \begin{Bmatrix} D_{MT}\left(R(mt_i), R(mt_j)\right) + D_{LS}(ls_{i,1} - ls_{j,1}) \\ D_{MT}\left(R(mt_i), R(mt_j)\right) + D_{LS}(ls_{i,1} - gap) \\ D_{MT}\left(R(mt_i), R(mt_j)\right) + D_{LS}(gap - ls_{j,1}) \end{Bmatrix}$$

Όπου $R(mt_i)$ δηλώνει τα υπόλοιπα επεισόδια (*LifeSteps)* της τροχιάς mt_i αφότου αφαιρεθεί το πρώτο της επεισόδιο $ls_{i,1}$ και *gap* είναι ένα εικονικό επεισόδιο του οποίου το *MBB* έχει μία ελάχιστη έκταση γύρω από την περιοχή του *MBB* του συνολικού dataset και το κειμενικό του στοιχείο αντιστοιχεί σε ένα μηδενικό διάνυσμα. Αποδεικνύεται ότι η συνάρτηση απόστασης D_{MT} είναι μία μετρική καθώς ικανοποιεί τα αξιώματα της ταύτισης, της συμμετρίας και της τριγωνικής ανισότητας [4].

3.3 Βασισμένη σε Μοντέλο Κατηγοριοποίηση Χρονοσειρών

Η οικογένεια των Μαρκοβιανών μοντέλων μπορεί να χρησιμοποιηθεί στην ανάλυση πολλών διαφορετικών τύπων ακολουθιακών δεδομένων καθώς έχουν αποδειχθεί πολύ ευέλικτα στο σύνολο των θεωρήσεων που υιοθετούν. Η χρήση της συγκεκριμένης κατηγορίας στοχαστικών

μοντέλων στην ανάλυση δεδομένων κίνησης και επομένως στην αναγνώριση της κινητικότητας των ανθρώπων έχει κεντρήσει το ενδιαφέρον πολλών ερευνητών.

Στο [8] εφαρμόστηκε βασισμένη σε μοντέλο (*model-based*) και βασισμένη σε παράδειγμα (*exemplar-based*) αναζήτηση σε μεγάλες βάσεις δεδομένων χρονοσειρών. Εστιάζουμε την προσοχή μας στην μεθοδολογία που ακολουθήθηκε στην βασισμένη σε μοντέλο προσέγγιση καθώς η μοντελοποίηση έγινε με χρήση των HMMs. Δεδομένης μίας βάσης δεδομένων από χρονοσειρές όπου έχει πραγματοποιηθεί ήδη το βήμα της ομαδοποίησης (*clustering*), κάθε κλάση του συνόλου αναπαριστάται από ένα hidden Markov model. Με αυτόν τον τρόπο αναπαράστασης των ομάδων, δεδομένου ενός query *Q* και του συνόλου των μοντέλων αναζητείται το μοντέλο εκείνο που μεγιστοποιεί την πιθανότητα να έχει παραχθεί το *Q* από αυτό. Το συγκεκριμένο ερευνητικό ζήτημα δείχνει να είναι ανάλογο με το δικό μας καθώς οι *Semantic Mobility Timelines (SMT)* αποτελούν, ουσιαστικά, χρονοσειρές από επεισόδια κίνησης. Στη συνέχεια θα παρουσιάσουμε τα κύρια χαρακτηριστικά της μεθοδολογίας που εφαρμόστηκε και θα μπορούσε να αποτελέσει οδηγό στην επίλυση του δικού μας προβλήματος. Επισημαίνουμε ότι ακόμη και αν αποφασίσουμε, τελικά, υπέρ της χρήσης κάποιου άλλου Μαρκοβιανού μοντέλου (π.χ. ενός hidden semi-Markov model) η κατανόηση της μεθοδολογίας που θα περιγράψουμε ακόλουθα είναι απαραίτητη καθώς τα βασικά ζητήματα που έχουμε να αντιμετωπίσουμε είναι κοινά σε κάθε εφαρμογή Μαρκοβιανού συστήματος.

Η φάση εκπαίδευσης των HMMs (δηλαδή το ζήτημα του *learning* όπως ορίστηκε πιο πάνω) χωρίστηκε σε δύο βήματα: πρώτα σε αυτό της αρχικοποίησης (*initialization*) και έπειτα στο βήμα του επαναληπτικού ραφιναρίσματος (*iterative refinement*). Θεωρώντας μία βάση δεδομένων *U*, αποτελούμενη από *C* κλάσεις, με το πλήθος των χρονοσειρών σε κάθε κλάση C_i να συμβολίζεται με $|C_i|$, η υλοποίηση της φάσης εκπαίδευσης, όπως περιγράφεται ακόλουθα, αναπαριστά μία αυθαίρετη κλάση C_i της *U*, με $|C_i| = n$, μέσω ενός *πρώτης τάξης*, τύπου *Bakis* HMM.

Στο πρώτο βήμα της φάσης εκπαίδευσης του HMM, για την κλάση C_i υπολογίστηκε, αρχικά, το medoid της κλάσης, δηλαδή η χρονοσειρά που απέχει την μικρότερη μέση απόσταση από όλες τις υπόλοιπες χρονοσειρές της κλάσης. Για τον υπολογισμό των αποστάσεων μεταξύ των χρονοσειρών, ώστε να βρεθεί το medoid της κλάσης, χρησιμοποιήθηκε μία μετρική η οποία ονομάζεται *Dynamic Time Warping* και βασίζεται στον δυναμικό προγραμματισμό. Αφότου υπολογιστεί, το medoid χωρίζεται σε M ίσα τμήματα τα οποία αποτελούν τις M κρυφές καταστάσεις του μοντέλου (αναφέρεται ότι το τελευταίο τμήμα μπορεί να αποτελείται από λιγότερα στοιχεία από τα υπόλοιπα). Στη συνέχεια, βρίσκονται και αποθηκεύονται τα στοιχεία κάθε μίας από τις υπόλοιπες χρονοσειρές της κλάσης που εμπίπτουν σε κάθε τμήμα-κρυφή κατάσταση όπως επίσης και τα στοιχεία του medoid που ανήκουν σε κάθε κατάσταση. Θεωρώντας την κατανομή της πιθανότητας εμφάνισης των παρατηρούμενων συμβόλων σε κάθε κατάσταση να είναι Κανονική, το οποίο είναι σύνηθες στα HMMs, αυτή θα περιγράφεται από την μέση τιμή και την τυπική απόκλιση των στοιχείων που την αποτελούν, τα οποία έχουν προηγουμένως υπολογιστεί και αποθηκευτεί.

Όπως ορίστηκε ένα HMM, πέρα από τον υπολογισμό του πίνακα *B*, των πιθανοτήτων εμφάνισης των συμβόλων, για τον προσδιορισμό του πρέπει να υπολογιστεί ο πίνακας μετάβασης μεταξύ των καταστάσεων (*A*). Οι πιθανότητες μετάβασης υπολογίστηκαν ως εξής: αν $|S_t|$ είναι ο συνολικός αριθμός των στοιχείων που ανήκουν σε μία κατάσταση S_t τη χρονική στιγμή *t*, τότε η πιθανότητα να μεταβούμε στην επόμενη κατάσταση είναι $P = \frac{n}{|St|}$, όπου n το πλήθος των χρονοσειρών της κλάσης. Συνεπώς, η πιθανότητα παραμονής στην ίδια κατάσταση είναι $P = \frac{(|St|-n)}{|St|}$, ενώ όταν φτάσουμε στην τελευταία κατάσταση η πιθανότητα παραμονής είναι ίση με 1 καθώς δεν υπάρχει άλλη ακόλουθη κατάσταση. Ο πίνακας π των πιθανοτήτων έναρξης δεν χρειάστηκε υπολογισμό καθώς κάθε χρονοσειρά έπρεπε να ξεκινάει υποχρεωτικά από την πρώτη κατάσταση (λόγω τύπου *Bakis*).

Στο δεύτερο βήμα της εκπαίδευσης, ραφινάρονται τα HMMs που δημιουργήθηκαν για όλες τις κλάσεις της βάσης δεδομένων *U* κατά τη διάρκεια του πρώτου βήματος. Αρχικοποιείται ένας πίνακας ακολουθιών καταστάσεων όπου αποθηκεύονται για όλες τις χρονοσειρές οι καταστάσεις από τις οποίες περνάνε. Για να βρέθει η βέλτιστη ακολουθία κρυμμένων καταστάσεων για κάθε χρονοσειρά σε κάθε κλάση (δηλαδή το ζήτημα της *αναγνώρισης* όπως ορίστηκε πιο πάνω) χρησιμοποιήθηκε ο αλγόριθμος Viterbi ο οποίος είναι ο βιβλιογραφικά επικρατέστερος για την επίλυση του ζητήματος αυτού. Στο τέλος του βήματος του ραφιναρίσματος, η μέση τιμή και η τυπική απόκλιση των πιθανοτήτων εμφάνισης αλλά και οι πίνακες μετάβασης έχουν ενημερωθεί κατάλληλα, όπως έγινε και στο προηγούμενο βήμα όπου εκπαιδεύτηκαν τα μοντελα. Σημειώνεται ότι η διαδικασία του δεύτερου βήματος μπορεί να επαναληφθεί αρκετές φορές μέχρι να επιτευχθεί ικανοποιητικής ακρίβειας κατηγοριοποίηση στο σύνολο των δεδομένων εκπαίδευσης.

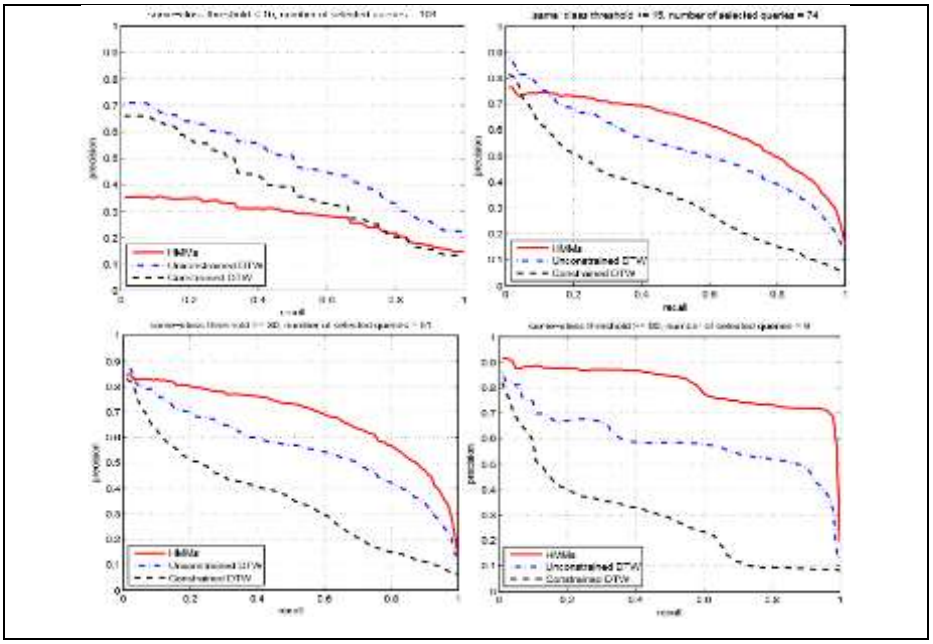
Αφότου έχουν εκπαιδευτεί τα HMMs για κάθε κλάση C_i, ακολουθεί η επίλυση του ζητήματος του evaluation. Για μία δεδομένη χρονοσειρά X, αναζητούμε την πιθανότητα της *X* δεδομένου ότι η *X* ανήκει στην κλαση C_i (η οποία πλέον περιγράφεται από το HMM λ που έχει εκπαιδευτεί) δηλαδή την πιθανότητα $P(X \mid \lambda)$. Η επίλυση του συγκεκριμένου ζητήματος πραγματοποιήθηκε με χρήση του αλγοριθμού Forward.

Η αναζήτηση μιας χρονοσειράς γίνεται ως εξής (παραδειγματικά δίνεται ότι αναζητούνται στη βάση δεδομένων χρονοσειρές που να αντιστοιχούν στην δραστηριότητα «falling down»): Αρχικά, εκχωρείται ένα σκορ σε κάθε χρονοσειρά της βάσης δεδομένων το οποίο υποδεικνύει πόσο καλά ταιριάζει κάθε χρονοσειρά στην αναζήτηση που κάνουμε. Έπειτα, οι χρονοσειρές κατατάσσονται σύμφωνα με το σκορ που έλαβαν. Τέλος, επιστρέφεται στο χρήστη το σύνολο των K καλύτερων ταιριασμάτων (το K προσδιορίζεται από το χρήστη).

Στην περίπτωση της βασισμένης σε μοντέλο αναζήτησης, το ζητούμενο ερώτημα (query) προσδιορίζεται από ένα μοντέλο. Για το παράδειγμα της αναζήτησης της δραστηριότητας «falling down» το query που υποβάλλεται μπορεί να είναι ένα μοντέλο που εκπαιδεύτηκε από παραδείγματα «falling down» δραστηριότητας. Τότε, το σκορ που ανατίθεται σε κάθε χρονοσειρά της βάσης δεδομένων είναι η απόδοση του μοντέλου της συγκεκριμένης δραστηριότητας σε αυτήν την χρονοσειρά, δηλαδή η πιθανότητα $P(O \mid \lambda)$ η οποία υπολογίζεται με χρήση του αλγορίθμου Forward (όπου *O* το σύνολο των παρατηρήσεων της χρονοσειράς *X*).

Στα αποτελέσματα της ανάλυσης γίνεται σύγκριση των δύο μεθόδων, της βασισμένης σε μοντέλο και της βασισμένης σε παράδειγμα. Τα μέτρα ακρίβειας που χρησιμοποιήθηκαν για να εκτιμηθεί η απόδοση των δύο μεθόδων είναι το *precision* και το *recall*. Ως *precision* (γνωστό και ως *θετική προβλεπτική αξία*) είναι το κλάσμα των δεδομένων που κατηγοριοποιήθηκαν σωστά, δηλαδή ήταν σχετικά με το δεδομένο query, προς το σύνολο των δεδομένων που ανακλήθηκαν από την βάση για το συγκεκριμένο query (σωστά και λανθασμένα). Το *recall* (*γνωστό και ως ευαισθησία*) είναι το κλάσμα των δεδομένων που κατηγοριοποιήθηκαν σωστά προς το σύνολο των σχετικών, με το δεδομένο query, δεδομένων που υπάρχουν στη βάση.

Δεδομένου ότι αυτό που μας ενδιαφέρει είναι η αποτελεσματικότητα της χρήσης των HMMs, εστιάζουμε την προσοχή μας στο γεγονός ότι τα HMMs απέδοσαν καλύτερα από την βασισμένη σε παράδειγμα μέθοδο όταν το μέγεθος των δεδομένων εκπαίδευσης ήταν αρκετά μεγάλο (τουλάχιστον 15 παραδείγματα εκπαίδευσης). Το αποτέλεσμα αυτό μπορεί να αποτελέσει γνώμονα και για την δική μας προσέγγιση κατά το στάδιο της υλοποίησης. Ακόλουθα δίνονται οι πίνακες των πειραματικών αποτελεσμάτων που δείχνουν την βελτιωμένη απόδοση του HMM όσο αυξάνεται το πλήθος των δεδομένων εκπαίδευσης.



Εικόνα 6. Σύγκριση του γραφήματος Precision vs Recall της βασισμένης σε μοντέλο (κόκκινη γραμμή) με τις δύο μεθόδους βασισμένης σε παράδειγμα (μπλε και μαύρη διακεκομμένη γραμμή) αναζήτησης όταν το πλήθος των παραδειγμάτων εκπαίδευσης είναι <15, ≥15, ≥30 και ≥60 αντίστοιχα.

Κ Ε Φ Α Λ Α Ι Ο 4

Μοντέλα Markov για Σημασιολογικά Εμπλουτισμένες Τροχιές

4.1 Το Πρόβλημα

Δοθέντων, μίας βάσης δεδομένων (*semantic mobility database - SMD*) αποτελούμενη από τις σημασιολογικά εμπλουτισμένες τροχιές (*semantic mobility timelines*) *N* χρηστών, κάθε μία από τις οποίες ανήκει σε μία κλάση C_{mtli}, δηλαδή $SMD = \{(mtl_1, c_{mtl_1}), \dots, (mtl_N, c_{mtl_N})\}$, μίας συνάρτησης απόστασης *dist()* η οποία ποσοτικοποιεί την ομοιότητα ανάμεσα σε δύο τροχιές και μίας μή κατηγοριοποιημένης τροχιάς *Q*, να βρεθεί η κλάση της *Q* που ικανοποιεί την σχέση:

$$c_Q = \{c_{mtl_i} | argmin_i(dist(Q, mtl_i)), \forall\, i = 1, \dots, N\}$$

Στην προσέγγισή μας η αναζήτηση της κλάσης της μή κατηγοριοποιημένης τροχιάς θα πραγματοποιείται με πιθανοκρατικά κριτήρια. Απαραίτητη προϋπόθεση είναι να έχουμε στην διάθεσή μας την πληροφορία για την κλάση στην οποία ανήκει κάθε τροχιά των δεδομένων εκπαίδευσης, δηλαδή το αποτέλεσμα του αλγορίθμου της ομαδοποίησης των σημασιολογικών τροχιών που είδαμε νωρίτερα. Δοθέντος, επομένως, του συνόλου των κλάσεων της βάσης δεδομένων *SMD* των σημασιολογικά εμπλουτισμένων χρονοδιαγραμμάτων κίνησης, σκοπός μας είναι να επιτύχουμε την αναπαράσταση κάθε ομάδας μέσω ενός Μαρκοβιανού μοντέλου. Με αυτόν τον τρόπο, για κάθε μή κατηγοριοποιημένη τροχιά θα αναζητούμε το μοντέλο από το οποίο είναι πιθανότερο να έχει αυτή παραχθεί. Συγκεκριμένα, δοθείσας μίας μή κατηγοριοποιημένης σημασιολογικά εμπλουτισμένης τροχιάς *Q* και του συνόλου των μοντέλων λ^k, $k = 1, \dots, K$, όπου *K* είναι το πλήθος των μοντέλων (δηλαδή των κλάσεων) ψάχνουμε το μοντέλο εκείνο που μεγιστοποιεί την πιθανότητα $P(Q|\lambda^k)$, δηλαδή

$$c_Q = \{c_k | argmax_k[P(Q|\lambda^k)]\}.$$

4.2 Μοντελοποίηση

4.2.1 Εισαγωγή

Για την επίλυση του προβλήματος της κατηγοριοποίησης ενός συνόλου δεδομένων κίνησης με χρήση ενός Μαρκοβιανού μοντέλου πρέπει να υλοποιήσουμε τα επόμενα δύο βήματα [5]:

1) Για κάθε κλάση $C_k, k = (1, 2, \dots, K)$, της βάσης δεδομένων *SMD*, πρέπει να χτίσουμε ένα HMM λ^k, δηλαδή να εκτιμήσουμε τις παραμέτρους του μοντέλου (A, B, π) οι οποίες μεγιστοποιούν την πιθανοφάνεια των παρατηρήσεων των δεδομένων εκπαίδευσης για καθε κλάση C_k.

2) Για κάθε μή κατηγοριοποιημένη τρόχια Q υπολογίζουμε την πιθανοφάνεια των παρατηρήσεων της τροχιάς δοθέντος κάθε μοντέλου, δηλαδή την πιθανότητα $P(Q|\lambda^k), 1 \leq k \leq K$. Επιλέγουμε ως κλάση της μή κατηγοριοποιημένης τροχιάς Q την κλάση εκείνου του μοντέλου που μεγιστοποεί αυτήν την πιθανότητα.

Το πρώτο βήμα αποτελεί το ζήτημα της *εκπαίδευσης* του μοντέλου που ορίσαμε στα *Βασικά Ζητήματα των HMMs* και το οποίο θα επιλύσουμε με χρήση του αλγορίθμου *Expectation-Maximization.* Το δεύτερο βήμα είναι το ζήτημα της *αναγνώρισης* και το οποίο γενικά, αλλά και ειδικά στην εφαρμογή μας, εκτελείται με χρήση του αλγορίθμου *Viterbi*, δηλαδή το βέλτιστο μονοπάτι κρυφών καταστάσεων χρησιμοποιείται.

Η επιλογή του καταλληλότερου μοντέλου από την ευρύτερη οικογένεια των Μαρκοβιανών μοντέλων είναι το πρώτο που καλούμαστε να απαντήσουμε. Ακόλουθα της επιλογής του μοντέλου, πρέπει να καθοριστεί το σύνολο των υποθέσεων-θεωρήσεων της μοντελοποίησης που στοχεύουμε να εφαρμόσουμε. Συγκεκριμένα, θα ασχοληθούμε με τα ζητήματα που αφορούν στην τοπολογία του μοντέλου, δηλαδή να οριστεί το σύνολο των κρυφών του καταστάσεων καθώς και οι σχέσεις μεταξύ τους, δηλαδή να επιλεχθεί ο τύπος του μοντέλου (εργοδικό, αριστερό-δεξί ή και κάποιο άλλο). Επιπλέον, θα πρέπει να οριστεί ο χώρος των παρατηρούμενων ακολουθιών (ή ακολουθιών εξόδου). Η επιλογή εκείνου του μοντέλου, που θα είναι ικανό να αναπαραστήσει αποτελεσματικότερα τις ιδιαίτερες δομές και χαρακτηριστικά του συστήματος των κινούμενων χρηστών, είναι κομβικής σημασίας προκειμένου να καταφέρουμε να επιτύχουμε υψηλά ποσοστά κατηγοριοποίησης.

Στην ενότητα 2.3, που παρουσιάστηκαν τα Hidden Markov Models, είδαμε ότι υποθέτουν Γεωμετρική κατανομή του χρόνου παραμονής στις καταστάσεις. Στην ενότητα 2.4 είδαμε ότι τα κρυμμένα ημί-Μαρκοβιανά μοντέλα (hidden semi-Markov models) επεκτείνουν τα HMMs επιτρέποντας την διάρκεια του χρόνου παραμονής στις κρυφές καταστάσεις να είναι μεταβλητή και όχι Γεωμετρική με την επιπλέον διαφοροποίηση ότι σε ένα HMM μπορεί να παρατηρηθεί μία παρατήρηση σε κάθε κατάσταση αντίθετα με ένα HSMM όπου μπορούμε να παρατηρήσουμε μία ακολουθία παρατηρήσεων ανάλογα με το χρόνο που το σύστημα παρέμεινε στην κατάσταση. Στην περίπτωση θεώρησης της Γεωμετρικής κατανομής, για την μεταβλητή διάρκεια παραμονής του HSMM, τα δύο μοντέλα ταυτίζονται. Σε αυτό το σημείο, είναι απαραίτητο να παραθέσουμε μία πολύ σημαντική υποσημείωση από το [5] η οποία θα μας βοηθήσει στην ορθότερη επιλογή του μοντέλου: "Σε περιπτώσεις όπου χρησιμοποιούνται μοντέλα τύπου Bakis, δηλαδή αριστερά-δεξιά μοντέλα όπου το πλήθος των καταστάσεων είναι ανάλογο με την μέση διάρκεια, η ρητή συμπερίληψη της διάρκειας παραμονής στις καταστάσεις δεν είναι ούτε απαραίτητη αλλά ούτε χρήσιμη" ([5], σελ. 269). Δεδομένου ότι, όπως θα δούμε στην επόμενη ενότητα, ο *Bakis* τύπος μοντέλου δείχνει να προσαρμόζεται καταλληλότερα στη δομή των δεδομένων κίνησης της ανάλυσής μας καταλήγουμε στην

απόφαση να αναπαραστήσουμε κάθε κλάση C_k της βάσης δεδομένων *SMD* με χρήση ενός HMM.

4.2.2 Επιλογή των Παραμέτρων του Μοντέλου

Πρίν προχωρήσουμε στην επιλογή των παραμέτρων του μοντέλου είναι σημαντικό να εστιάσουμε, και πάλι, στα δεδομένα ώστε να κατανοήσουμε ορισμένα βασικά χαρακτηριστικά του συστήματος της ανάλυσής μας. Όπως είδαμε στην ενότητα 3.1, η ακατατέργαστη τροχία (raw trajectory) ενός χρήστη είναι ουσιαστικά μία χρονοσειρά από *timestamps* της μορφής *(t, (x, y), tags)* τα οποία παράγονται σε τυχαίες χρονικές στιγμές *t*. Είδαμε, ακόμη, ότι η εξαγωγή της σημασιολογικής τροχιάς ενός χρήστη αποτελεί ουσιαστικά ένα επόμενο βήμα επεξεργασίας των ακατέργαστων τροχιών. Πρακτικά, με την εξαγωγή της σημασιολογικής μορφής της κίνησης έχουμε το σύνολο των timestamps της ακατέργαστης τροχιάς να ομαδοποιούνται στα *stop* και *move* επεισόδια που πραγματοποίησε ο χρήστης.

Σε αυτό το σήμειο είναι απαραίτητο να τονίσουμε μία σημαντική διαφορά ανάμεσα στα *stop* και τα *move* επεισόδια κίνησης, η οποία θα επηρεάσει καταλυτικά την μοντελοποίηση της εφαρμογής μας. Τα *stop* επεισόδια εξελίσσονται εντός μίας περιορισμένης χωρικά περιοχής σε αντίθεση με τα *move* επεισόδια τα οποία εξελίσσονται σε πολύ ευρύτερες περιοχές. Δεδομένου ότι για τον παρατηρούμενο χώρο θα θεωρήσουμε, όπως συνηθίζεται στις εφαρμογές κρυφών Μαρκοβιανών μοντέλων, την Κανονική κατανομή [8] (ή την πολυδιάστατη Κανονική κατανομή [10]), η θεώρηση μιας διδιάστατης Κανονικής κατανομής για τις τιμές (x, y) των *move* επεισοδίων δεν θα είχε κάποιο νόημα. Τελικά, αποφασίζουμε να αποκόψουμε από την μοντελοποίηση τα *move* τμήματα της κίνησης θεωρώντας, επομένως, ότι η κίνηση ενός χρήστη περιγράφεται πλήρως από το σύνολο των *stop* επεισοδίων που πραγματοποίησε.

Όσον αφορά τον χρόνο, είδαμε ότι τα *stop* και *move* επεισόδια εναλλάσσονται διαδοχικά στον χρόνο ενώ κάθε επεισόδιο χαρακτηρίζεται από την ώρα έναρξης και λήξης του. Οι χρήστες της ίδιας κλάσης μπορούν να παρουσιάζουν μεταβλητότητα στους χρόνους έναρξης και λήξης των αντίστοιχων επεισοδίων σύμφωνα με τον τρόπο ομαδοποίησης των τροχιών τους.

Τέλος, αναφορικά με τις κειμενικές μεταβλητές στην γενική περίπτωση μπορούν να είναι αυθαίρετου μήκους και να διαφέρουν από χρήστη σε χρήστη. Είναι λογικό οι χρήστες της κλάσης να παρουσιάζουν κοινές τιμές σε ορισμένες απο τις κειμενικές τους τιμές ή να παρουσιάζουν μικρή μεταβλητότητα σε κάποιες άλλες κειμενικές μεταβλητές, μιας και η απόστασή τους καθορίστηκε, όπως είδαμε νωρίτερα, σύμφωνα και με κειμενικά κριτήρια. Ωστόσο, στα δεδομένα της ανάλυσής μας οι χρήστες παρουσιάζουν κοινές τιμές στις κειμενικές μεταβλητές τους.

Το πρώτο στο οποίο καλούμαστε να απαντήσουμε για την εφαρμογή ενός HMM είναι το ζήτημα του ορισμού των κρυφών του καταστάσεων. Όπως είδαμε στην παρουσίαση των δεδομένων κίνησης, υπάρχει μία έμφυτη τμηματοποίηση μίας σημασιολογικής τροχιάς στα επεισόδια κίνησης, όπου το σύνολο των timestamps της ακατατέργαστης τροχιάς ομαδοποιούνται στα αντίστοιχα *stop* ή *move* επεισόδια. Επομένως, θα θέλαμε να εκμεταλλευτούμε αυτήν την έμφυτη τμηματοποίηση και κάθε κρυφή κατάσταση του μοντέλου να αναπαριστά ένα *stop* επεισόδιο κίνησης με συγκεκριμένα χωρικά, χρονικά και κειμενικά χαρακτηριστικά.

Για να καταλήξουμε στον καταλληλότερο τρόπο προσδιορισμού των κρυφών καταστάσεων του μοντέλου, θα παρατηρήσουμε, αρχικά, δύο πολύ σημαντικά στοιχεία τα οποία θα μας βοηθήσουν. Το πρώτο, που έχουμε να επισημάνουμε, είναι το γεγονός ότι οι χρήστες της ίδιας κλάσης μπορεί να πραγματοποιούν διαφορετικό πλήθος *stop* επεισοδίων κίνησης. Αυτό μπορεί να συμβαίνει είτε λόγω του ότι κάποιοι δεν πραγματοποίησαν ένα από τα *stop* επεισόδια της κλάσης είτε επειδή για κάποιους χρήστες οι καταγραφές σταματούν νωρίτερα (π.χ. λόγω προβλήματος του συστήματος που στέλνει τα δεδομένα ή οτιδήποτε άλλο). Το δεύτερο στοιχείο που θα σημειώσουμε είναι ότι εφόσον το αποτέλεσμα της κατηγοριοποίησης θα προκύπτει πιθανοκρατικά αυτό μας παρέχει τη δυνατότητα να αποφασίζουμε την πιθανότερη κλάση στην οποία μπορεί να ανήκει μία τροχιά ακόμα και αν αυτή είτε δεν "πέρασε" από κάποια κρυφή κατάσταση του μοντέλου είτε πραγματοποίησε κάποιο επεισόδιο κίνησης το οποίο δεν υπήρχε στα δεδομένα εκπαίδευσης και επομένως δεν είχε μοντελοποιηθεί. Δεδομένου ότι πρωταρχικός μας στόχος είναι η επιτυχής κατηγοριοποίηση μίας μή κατηγοριοποιημένης τροχιάς, θα θέλαμε να συμπεριλάβουμε στο μοντέλο ως κρυφές καταστάσεις όσο το δυνατόν περισσότερα *stop* επεισόδια.

Με βάση τα προαναφερόμενα, ορίζουμε ως σύνολο των κρυφών καταστάσεων το σύνολο των *stop* επεισοδίων μίας αντιπροσωπευτικής τροχιάς της κάθε κλάσης. Επιλέγουμε ως αντιπροσωπευτική εκείνη την τροχιά η οποία παρουσιάζει τα περισσότερα επεισόδια κίνησης. Αν υπάρχουν περισσότερες από μία που έχουν το ίδιο, μέγιστο πλήθος *stop* επεισοδίων μπορούμε να επιλέξουμε μία τυχαία. Με αυτόν τον τρόπο επιλογής της αντιπροσωπευτικής τροχιάς στοχεύουμε στη μοντελοποίηση όσο το δυνατόν περισσότερων *stop* επεισοδίων και τελικά στην καλύτερη δυνατή επίδοση του μοντέλου.

Έστω η αντιπροσωπευτική σημασιολογική τροχιά της κλάσης C_k, την οποία θα μπορούσαμε να συμβολίσουμε ως $\overline{mtl_{C_k}}$ και η οποία αποτελείται από ένα σύνολο *M stop* επεισοδίων κίνησης. Προκειμένου να ορίσουμε αυτά τα *M* επεισόδια ως τις *M* κρυφές καταστάσεις του μοντέλου δουλεύουμε ως εξής: αρχικά, βρίσκουμε το σύνολο των *stop* επεισοδίων όλων των υπόλοιπων χρηστών της κλάσης τα οποία αντιστοιχούν σε κάθε *stop* επεισόδιο της $\overline{mtl_{C_k}}$ τροχιάς. Υπολογίζουμε και πάλι το Minimum Bounding Box (*MBB*) κάθε *stop* επεισοδίου της $\overline{mtl_{C_k}}$ σύμφωνα με τις minimum και maximum τιμές των *(x, y)*, t_{start} και t_{end} του συνόλου των χρηστών της κλάσης. Επιπλέον, ορίζουμε έναν άξονα τιμών (μία ακόμη διάσταση) ο οποίος περιγράφει τις τιμές των κειμενικών μεταβλητών (*tags*). Όπως είδαμε πιο πάνω, το κειμενικό τμήμα κάθε επεισοδίου μπορεί να είναι εν μέρει κοινό για τους χρήστες της ίδιας κλάσης. Μπορούμε να αποδώσουμε στον προσδιορισμό των κρυφών καταστάσεων τις κειμενικές μεταβλητές που παραμένουν σταθερές για το σύνολο των χρηστών, οι οποίες μπορούμε να πούμε ότι δίνουν μία γενική περιγραφή του επεισοδίου (ή της δραστηριότητας που πραγματοποιείται στο συγκεκριμένο επεισόδιο κλπ). Για παράδειγμα, η τυχαία κατάσταση *i* θα είναι ένα *stop* επεισόδιο κίνησης το οποίο εξελίσσεται εντός συγκεκριμένου χωροχρονικού χωρίου και χαρακτηρίζεται από τις σταθερές κειμενικές τιμές των χρηστών όπως π.χ. ('stop', 'office').

Όσον αφορά τις συνδέσεις μεταξύ των κρυφών καταστάσεων θα ακολουθήσουμε την θεώρηση ενός *αριστερού-δεξιού* ή *Bakis* μοντέλου. Σε ένα αριστερό-δεξί μοντέλο το σύστημα δεν μπορεί να επιστρέψει σε προηγούμενη κατάσταση προχωρώντας με αυτόν τον τρόπο από

τα αριστερά προς τα δεξιά. Αυτή η θεώρηση δείχνει να προσαρμόζεται στην δική μας μοντελοποίηση όπου το σύστημα θα ξεκινά από ένα *stop* επεισόδιο-κρυφή κατάσταση και θα πραγματοποιεί τα ακολουθιακά επεισόδια κίνησης, καθένα από τα οποία θεωρείται μοναδικό, και επομένως δεν θα επανέρχεται ποτέ σε προηγούμενη κατάσταση. Επιπλέον, θα θεωρήσουμε ένα πρώτης τάξης HMM και επομένως το συστήμα θα μπορεί να πραγματοποιεί μεταβάσεις μόνο σε διαδοχικές κρυφές καταστάσεις. Όπως είναι αναμενόμενο, αυτές οι δύο θεωρήσεις θα επιφέρουν και συγκεκριμένους περιορισμούς στον πίνακα μετάβασης A του μοντέλου καθώς και στον πίνακα των πιθανοτήτων έναρξης π.

Ο πίνακας A, των πιθανοτήτων μεταβάσης μεταξύ των κρυφών καταστάσεων θα έχει τα εξής χαρακτηριστικά: ως Bakis μοντέλο, δεν θα έχει πιθανότητες μετάβασης σε προηγούμενες καταστάσεις ενώ το σύστημα θα ξεκινάει υποχρεωτικά από την πρώτη κατάσταση. Ακόμη, ως πρώτης τάξης μοντέλο θα δίνει πιθανότητες μετάβασης μόνο σε διαδοχικές καταστάσεις. Έτσι, σε κάθε χρονική στιγμή t όταν παράγεται μία παρατήρηση το σύστημα μπορεί είτε να παραμείνει στην ίδια κατάσταση είτε να μεταβεί στην αμέσως επόμενη. Επομένως, για να υπολογίσουμε τον πίνακα μετάβασης A, υπολογίζουμε τις πιθανότητες μετάβασης a_{ij}, όμοια με το [8], διαιρώντας το πλήθος των χρηστών στην i κατάσταση προς το σύνολο των παρατηρήσεων που ανήκουν σε αυτήν την κατάσταση, δηλαδή $a_{ij} = \frac{n_i}{|O_t|}$, όπου n_i είναι το πλήθος των χρηστών και $|O_t|$ το πλήθος των παρατηρήσεων όλων των χρηστών στο συγκεκριμένο *stop* επεισόδιο. Αντίστοιχα, η πιθανότητα παραμονής στην ίδια κατάσταση θα είναι $(1 - a_{ij})$ ενώ για την τελευταία κατάσταση θα είναι ίση με μονάδα.

Ο πίνακας π, των πιθανοτήτων έναρξης, θα είναι αυτός που χαρακτηρίζει τα μοντέλα Bakis και για τον οποίο ισχύει $\pi_i = \begin{Bmatrix} 1, S_i = 1 \\ 0, S_i \neq 1 \end{Bmatrix}$. Θα υιοθετήσουμε την απλοποιημένη θεώρηση των αρχικών συνθηκών σύμφωνα με την οποία η πρώτη κατάσταση ξεκινάει στον χρόνο 1 και η τελευταία τελειώνει στον χρόνο T [7]. Θα υιοθετήσουμε αυτήν την θεώρηση με το σύστημα να εισέρχεται στην πρώτη κατάσταση-*stop* επεισόδιο την χρονική στιγμή έναρξης των καταγραφών. Αντίστοιχα, ορίζουμε την τελευταία κατάσταση ως απορροφητική, δηλαδή το σύστημα παραμένει εκεί μέχρι την λήξη των καταγραφών χωρίς να πραγματοποιεί άλλες μεταβάσεις.

Τέλος, σχετικά με τον παρατηρούμενο χώρο (ή χώρο ακολουθιών εξόδου) θα ακολουθήσουμε την συνήθη θεώρηση της Κανονικής κατανομής για τις παρατηρούμενες τιμές ανά κρυφή κατάσταση και δεδομένου ότι οι σημασιολογικές τροχιές των χρηστών αποτελούν πολυδιάστατες ακολουθίες θα θεωρήσουμε ότι προσεγγίζονται από μία πολυδιάστατη Κανονική κατανομή σε κάθε κατάσταση. Σε κάθε κρυφή κατάσταση οι παρατηρούμενες τιμές των γεωγραφικών συντεταγμένων των χρηστών θα περιγράφονται από μία διδιάστατη Κανονική κατανομή. Σημειώνουμε ότι από τον παρατηρούμενο χώρο θα αποκόψουμε τις κειμενικές μεταβλητές καθώς στα δεδομένα της ανάλυσης μας δεν υπάρχουν κειμενικές μεταβλητές που να παρουσιάζουν μεταβλητότητα. Επιπλέον, εισάγουμε στον παρατηρούμενο χώρο τον χρόνο έναρξης t_{start} κάθε επεισοδίου, ο οποίος είδαμε ότι μπορεί να διαφέρει από χρήστη σε χρήστη. Συνεπώς, οι παρατηρούμενες ακολουθίες (x, y, t_{start}) θα περιγράφονται από μία πολυδιάστατη Κανονική κατανομή, δηλαδή θα περιγράφονται από τους μέσους των τριών χαρακτηριστικών καθώς και τον πίνακα διασπορών συνδιασπορών σε κάθε κρυφή κατάσταση.

Για την επιλογή της εισαγωγής του χρόνου έναρξης των επεισοδίων κάθε χρήστη στον παρατηρούμενο χώρο του μοντέλου έχουμε να σημειώσουμε το εξής: Με τον προσδιορισμό των κρυφών καταστάσεων με βάση το σύνολο των γνωρισμάτων των επεισοδίων αυτό που

τελικά στοχεύουμε είναι να μοντελοποιήσουμε τα χρονικά πρότυπα και τις ακολουθιακές δομές που υπάρχουν στα *stop* επεισόδια κίνησης των χρηστών. Για παράδειγμα, έστω ότι η τροχιά ενός χρήστη μίας τυχαίας κλάσης ξεκινάει με ένα συγκεκριμένο χωροχρονικά *stop* επεισόδιο το οποίο συνοδεύεται από μία ετικέτα που δηλώνει ότι ο χρήστης ξεκουράζεται σπίτι του. Στη συνέχεια πραγματοποιεί κάποια επεισοδία κίνησης όπως π.χ. στο γραφείο για εργασία και τελικά επιστρέφει σπίτι του παρουσιάζοντας ένα όμοιο τόσο κειμενικά όσο και χωρικά επεισόδιο κίνησης. Λαμβάνοντας υπόψη την χρονική διαφορά ανάμεσα σε αυτά τα δύο επεισόδια στοχεύουμε να τα διαχωρίσουμε ως δύο παράλληλα χρονικά επεισόδια. Οι δύο αυτές καταστάσεις μπορεί να είναι όμοιες ως προς την χωρική κατανομή και τις ετικέτες ωστόσο θα παρουσίαζουν διαφορετική χρονική κατανομή αλλά και θα ορίζουν πιθανότητες μετάβασης σε διαφορετικά, επόμενα, επεισόδια κίνησης. Η πρώτη κατάσταση του παραδείγματός μας θα μπορούσε να ακολουθείται από επεισόδια κίνησης όπως προς το γραφείο για εργασία ή το πανεπιστήμιο για μάθημα κ.λπ. ενώ η δεύτερη, που βρίσκει τον χρήστη σπίτι του τις απογευματινές ώρες, από καταστάσεις όπως για βόλτα με το ποδήλατο ή βόλτα στο σινεμά κ.λπ. Συνδυάζοντας, επομένως, την γεωγραφική τοποθεσία με την χρονική στιγμή που ο χρήστης έφθασε εκεί στοχεύουμε στο να καταφέρουμε να διακρίνουμε τις δύο αυτές καταστάσεις.

Συνοψίζοντας, μοντελοποιούμε κάθε κλάση C_k με ένα πρώτης τάξης, αριστερό-δεξί HMM $\lambda^k = (A, B, \pi)$, το σύνολο των παραμέτρων του οποίου ορίζεται ως εξής:

- Το σύνολο των κρυφών καταστάσεων S_i $(i = 1, 2, ..., M)$ είναι το σύνολο των *stop* επεισοδίων κίνησης της αντιπροσωπευτικής της κλάσης τροχιάς.
- Ο πίνακας A, των πιθανοτήτων μεταβάσης μεταξύ των κρυφών καταστάσεων θα χαρακτηρίζεται από τον τύπο (Bakis) και την τάξη (πρώτη) του μοντέλου με τις πιθανότητες μετάβασης και παραμονής να υπολογίζονται σύμφωνα με την μεθοδολογία που παρουσιάσαμε.
- Ο πίνακας B των πιθανοτήτων των παρατηρούμενων ακολουθιών σε κάθε κρυφή κατάσταση, θα ακολουθεί την πολυμεταβλητή Κανονική κατανομή. Συγκεκριμένα, η κατανομή των τριών χαρακτηριστικών σε κάθε κρυφή κατάσταση θα περιγράφεται από το διάνυσμα των τριών μέσων $\mu = (\mu_x, \mu_y, \mu_{t_{start}})$ και τον πίνακα διακύμανσης-συνδιακύμανσης Σ. Η συνάρτηση πυκνότητας πιθανότητας της πολυμεταβλητής Κανονικής κατανομής δίνεται:

$$f_X(X) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)}, x \in \mathbb{R}^p$$

- Ο πίνακας π, των πιθανοτήτων έναρξης, θα είναι αυτός που χαρακτηρίζει τα μοντέλα Bakis και για τον οποίο ισχύει: $\pi_i = \begin{cases} 1, S_i = 1 \\ 0, S_i \neq 1 \end{cases}$.

4.3 Εκπαίδευση των HMMs – Ο αλγόριθμος EM

Όπως είδαμε το ζήτημα της εκπαίδευσης ενός HMM επιλύεται με την εφαρμογή του αλγορίθμου *Expectation-Maximization (EM)*. Η πιθανοφάνεια ενός HMM δίνεται:

$$P(S, X) = P(S_1) \prod_{t=2}^{T} P(S_t/S_{t-1}) \prod_{t=1}^{T} P(X_t/S_t)$$

Ένα τοπικό μέγιστο της πιθανοφάνειας ενός HMM μπορεί να βρεθεί μέσω του αλγορίθμου ΕΜ ο οποίος αποτελείται από το αρχικό βήμα (*initial step*) του *Expectation step* και το επαναληπτικό βήμα βελτίωσης (*iterative refinement step*) του *Maximization step*. Ακολουθώντας την μεθοδολογία που περιγράψαμε στην προηγούμενη ενότητα, υπολογίζουμε τις αρχικές εκτιμήσεις των παραμέτρων του μοντέλου (*A, B, π*). Προσδιορίζουμε ένα HMM σύμφωνα με αυτές τις εκτιμήσεις οι οποίες θα αποτελέσουν τις αρχικές τιμές για το *Expectation step*. Στο επαναληπτικό βήμα του *Maximization step* οι παράμετροι του μοντέλου (*A, B, π*) προσαρμόζονται ώστε να μεγιστοποιείται η πιθανοφάνεια του μοντέλου $P(O|\lambda)$. Αυτό το επαναληπτικό βήμα εκτελείται μέχρι ένα επιθυμητό σημείο σύγκλισης, δηλαδή η πιθανοφάνεια του μοντέλου να μην βελτιώνεται περαιτέρω. Στη συνέχεια θα παρουσιάσουμε τα δύο βήματα που εκτελούνται κατά την υλοποίηση του αλγορίθμου ΕΜ [10].

1) Το ***Expectation step (E-step)*** περιλαμβάνει την εκτίμηση δύο όρων, συγκεκριμένα την πιθανότητα το σύστημα να βρίσκεται την χρονική στιγμή *t* στην κατάσταση *i* δοθείσας της παρατηρούμενης ακολουθίας

$$\gamma_t(i) = P(S_t = i/X = x; \theta) \tag{3}$$

Καθώς και την πιθανότητα η διαδικασία να έφυγε από την κατάσταση *i* την χρονική στιγμή *t* και να εισήλθε στην κατάσταση *j* την χρονική στιγμή *t+1* δοθείσας της παρατηρούμενης ακολουθίας

$$\xi_t(i,j) = P(S_t = i, S_{t+1} = j/X = x; \theta) \tag{4}$$

Αυτές οι τιμές μπορούν να υπολογιστούν μέσω μίας μεθόδου δυναμικού προγραμματισμού γνωστή ως *προς τα εμπρός-προς τα πίσω μέθοδος (forward-backward algorithm).*

2) Στο ***Maximization step (M-step)***, σύμφωνα με τις τιμές των (3) και (4) οι αρχικές πιθανότητες μετάβασης υπολογίζονται ως

$$\hat{\pi}_i' = \gamma_0(i) \; και \; \hat{p}_{ij}' = \frac{\sum_{t=1}^{T} \xi_t(i,j)}{\sum_{t=1}^{T} \sum_{i \neq j} \xi_t(i,j)}$$

Στη συνήθη περίπτωση όπου θεωρούμε ότι οι παρατηρήσεις X_t ακολουθούν Κανονική κατανομή δοθείσας της κρυφής κατάστασης $S_t = i$, δηλαδή $X_t/S_t = i \sim N(\mu_i, \sigma_i^2)$, τότε οι παράμετροι μ_i and σ_i^2 σε κάθε κρυφή κατάσταση υπολογίζονται ως

$$\widehat{\mu_\iota} = \frac{\sum_{t=1}^{T} \gamma_t(i) x_t}{\sum_{t=1}^{T} \gamma_t(i)} \; and \; \widehat{\sigma_\iota} = \frac{\sum_{t=1}^{T} \gamma_t(i)(x_t - \widehat{\mu_\iota})^2}{\sum_{t=1}^{T} \gamma_t(i)}$$

ΚΕΦΑΛΑΙΟ 5

Πειραματική Μελέτη

5.1 Προεπεξεργασία Δεδομένων

Για την υλοποίηση της βασισμένης σε μοντέλο κατηγοριοποίησης δεδομένων κίνησης δουλέψαμε με συνθετικά δεδομένα τα οποία περιγράφουν το πώς κινήθηκαν τετρακόσιοι χρήστες, κατά την διάρκεια μίας ημέρας και οι οποίοι ανήκουν σε τέσσερα διαφορετικά κινητικά πρότυπα (κλάσεις). Για να καταφέρουμε να φέρουμε τα δεδομένα σε μορφή κατάλληλη σύμφωνα με τις υποθέσεις τόσο της μοντελοποίησης που αποφασίσαμε να εφαρμόσουμε όσο και του πακέτου "mhsmm" της R [10] με το οποίο θα υλοποιήσουμε την μοντελοποίηση, ήταν απαραίτητο να γίνουν μία σείρα διαδικασιών στα πλαίσια της προεπεξεργασίας των δεδομένων.

Αρχικά, όπως είδαμε οι παρατηρήσεις *(t, (x, y), tags)* κάθε χρήστη έρχονται σε τυχαίες χρονικές στιγμές t. Πρόσθετα, η μεταβλητή t παίρνει τιμές σε ακρίβεια δευτερολέπτου, δηλαδή είναι της μορφης $t = hh{:}\,mm{:}\,ss$. Αν για παράδειγμα έχουμε καταγραφές μίας ημέρας, όπως στο training dataset της εφαρμογής μας, τότε η μεταβλητή t παίρνει τυχαίες τιμές στο διάστημα [0, 23:59:59.999]. Ωστόσο, για την εφαρμογή ενός HMM που να αναπαριστά την κάθε κλάση είναι απαραίτητο να έχουμε τις παρατηρήσεις σε ισαπέχουσες χρονικές στιγμές [10]. Αποφασίσαμε ότι σύμφωνα με την φύση των δεδομένων και την συγκεκριμένη διάρκεια των καταγραφών (24 ώρες) είναι βολικό να θεωρήσουμε ότι λαμβάνουμε παρατηρήσεις ανά 10 λεπτά θεωρώντας ότι το σύστημα εξελίσσεται σε χρόνο διακριτό $t' = [1, \dots, T]$, όπου T είναι η συνολική διάρκεια σε δεκάλεπτα του συστήματος. Τότε, θέλουμε το σύστημα να παράγει παρατηρήσεις ανά δέκα λεπτά καθόλη την διάρκεια ενός επεισοδίου (σύμφωνα με τις τιμές των t_{start} και t_{end} του επεισοδίου). Για την αναπαραγωγή των παρατηρούμενων τιμών στις ενδιάμεσες χρονικές στιγμές χρησιμοποιούμε τις τιμές των (x, y) και των *tags* που έλαβε κάθε χρήστης στα αρχικά δεδομένα. Η θεώρηση της χρονικής κλίμακας σε επίπεδο δεκαλέπτου επιφέρει απώλεια της τάξης του ενός πενταλέπτου, σύμφωνα με την στρογγυλοποίηση που πραγματοποιούμε σε δεκάλεπτα, την οποία θεωρούμε αμελητέα.

Επιπλέον, όπως είδαμε στην παρουσίαση της μοντελοποίησης, η δομή των *move* επεισοδίων κίνησης μας οδήγησε στην αποκοπή τους από την μοντελοποίηση. Όπως είναι αναμενόμενο αυτό διαφοροποιεί τον χρόνο διάρκειας του συστήματος, ο οποίος ορίζεται πλέον να είναι ο χρόνος διάρκειας των *stop* επεισοδίων μόνο. Έτσι, το σύστημα θα εξελίσσεται στον διακριτό χρόνο $t' = [1, \dots, T]$ ο οποίος θα είναι κοινός για όλους τους χρήστες του dataset με τον τρόπο που ορίστηκε προήγουμενα και θα περιγράφει την εξέλιξη του συστήματος των *stop* επεισοδίων κίνησης των χρηστών. Άμεση συνέπεια αυτού είναι η

μετατόπιση των *stop* επεισοδίων ώστε να είναι διαδοχικά στο χρόνο. Συνεπώς, αν η χρονική στιγμή $t'_{end}(1)$ είναι η χρονική στιγμή λήξης του πρώτου επεισοδίου τότε η χρονική στιγμή έναρξης του δεύτερου επεισοδίου $t'_{start}(2)$ θα είναι ίση με $t'_{start}(2) = t'_{end}(1) + 1$, το οποίο δεν ίσχυε καθώς η επόμενη χρονική στιγμή της ληξης του πρώτου επεισοδίου ήταν η χρονική στιγμή έναρξης του *move* επεισοδίου που τελικά αποκόπηκε. Να διευκρινίσουμε ότι με την μετατοπίση των επεισοδίων δεν χάνουμε την πραγματική χρονική πληροφορία των επεισοδίων (τους πραγματικούς χρόνους έναρξης και λήξης) την οποία έχουμε αποδώσει στον ορισμό των κρυφών καταστάσεων.

Ακόμη, η χρονική στιγμή λήξης των παρατηρήσεων δεν είναι κοινή τόσο από κλάση σε κλάση όσο και για τους χρήστες της ίδιας κλάσης. Για να ορίσουμε την κοινή χρονική στιγμή λήξης του συστήματος επιλέγουμε τον χρήστη με την μέγιστη διάρκεια καταγραφών σε όλο το dataset και ορίζουμε αυτήν ως την διάρκεια T του συστήματος. Για τους υπόλοιπους χρήστες της κάθε κλάσης που έχουν μεν τερματίσει την διαδικασία, δηλαδή έχουν φτάσει στην τελευταία κρυφή κατάσταση (*stop*), αλλά εξέπεμψαν την τελευταία τους παρατήρηση νωρίτερα από την χρονική στιγμή T, προσομοιώνουμε παρατηρήσεις με τρόπο όμοιο με πρίν ώστε να επιτύχουμε το ζητούμενο κοινό T. Σε αυτό το σημείο είναι χρήσιμο να σχολιάσουμε το γεγονός ότι η τελευταία κατάσταση κάθε κλάσης είναι η απορροφητική κατάσταση κάθε μοντέλου, δηλαδή το σύστημα δεν πραγματοποιεί άλλες μεταβάσεις από εκεί και έπειτα. Επομένως, η προσθήκη παρατηρήσεων στην τελευταία κατάσταση δεν επηρεάζει καμία από τις παραμέτρους του μοντέλου αλλά και δεν αλλοιώνει την πληροφορία των δεδομένων.

Τέλος, για τις ανάγκες της μοντελοποίησης αποφασίσαμε να κράταμε τον χρόνο έναρξης t_{start} σε κάθε επεισόδιο κάθε χρήστη ως μία έξτρα μεταβλητή καθώς είναι το στοιχείο που χρησιμοποιούμε για να διακρίνουμε όμοια ως προς τα χωροκειμενικά τους χαρακτηριστικά *stop* επεισόδια τα οποία εξελίσσονται σε διαφορετικές ώρες της ημέρας. Διευκρινίζουμε ότι ο χρόνος έναρξης του επεισοδίου διαφοροποιείται πλέον από την χρονική στιγμή, ως προς t', εισόδου της Μαρκοβιανής διαδικασίας στην αντίστοιχη κρυφή κατάσταση λόγω της αποκοπής των *move* επεισοδίων. Επίσης, δεδομένου ότι θεωρούμε πολυδιάστατη Κανονική κατανομή στον παρατηρούμενο χώρο, προσθέσαμε τεχνητά αμελητέα διακύμανση στις τιμές t'_{start} του πρώτου επεισοδίου κάθε χρήστη, για κάθε κλάση οι οποίες κανονικά παίρνουν την τιμή $t'_{start} = 1$ (δεκάλεπτο) που είναι η ορισμένη κοινή αρχή των καταγραφών.

Στην Εικόνα 7, που ακολουθεί, δίνεται ένα παράδειγμα τριών κινητικών προφίλ όμοιας δομής με των δεδομένων της ανάλυσής μας.

Εικόνα 7. Παράδειγμα τριών κινητικών προτύπων.

5.2 Πειραματικά Αποτελέσματα

Όσον αφορά στα πειραματικά αποτελέσματα, το πρώτο που έχουμε να επισημάνουμε είναι η αποτυχία του συστήματος στην μοντελοποίηση των δύο εκ των τεσσάρων κλάσεων του training dataset. Συγκεκριμένα, στις κλάσεις 2 και 3 η εκπαίδευση του μοντέλου δεν ήταν εφικτή. Ο λόγος που αυτό συμβαίνει είναι οι μεγάλες αποκλίσεις στους χρόνους έναρξης ορισμένων *stop* επεισοδίων το οποίο είναι αποτέλεσμα της ομαδοποίησης που πραγματοποιήθηκε στα δεδομένα. Δεδομένου ότι θεωρήσαμε την ώρα έναρξης των επεισοδίων ως ένα εκ των τριών χαρακτηριστικών του παρατηρούμενου χώρου, σε αυτές τις δύο κλάσεις η πολυδιάστατη Κανονική κατανομή δεν μπορεί να αναπαραστήσει τα δεδομένα και τελικά η εκπαίδευση του μοντέλου αποτυχάνει. Το μοντέλο δεν καταφέρνει να συλλάβει την δομή του συστήματος καταλήγοντας να παράγει μόνο μία παρατήρηση στο τέταρτο επεισόδιο (κρυφή κατάσταση) των δύο αυτών κλάσεων αδυνατώντας, τελικά, να υπολογίσει την πολυδιάστατη Κανονική κατανομή της συγκεκριμένης κρυφής κατάστασης.

Λόγω του ότι ο πειραματισμός με την ομαδοποίηση των δεδομένων, ώστε να καταλήξουμε σε ομάδες οι οποίες να είναι όμοιες στους χρόνους έναρξης, είναι εκτός πλάνων της παρούσας εργασίας αλλά και λόγω της δυσκολίας να βρούμε άλλο κατάλληλο training dataset (δηλαδή dataset το οποίο να μας παρέχει την πληροφορία για την κλάση στην οποία ανήκει κάθε χρήστης) θα παρουσιάσουμε τα πειραματικά αποτελέσματα μόνο για τους χρήστες των δύο πρώτων κλάσεων καθεμία από τις οποίες μοντελοποίηθηκε με ένα HMM.

Θεωρούμε, επομένως, ότι έχουμε μία βάση δεδομένων αποτελούμενη από τις τροχιές 196 χρηστών οι οποίοι ανήκουν σε δύο κινητικά προφίλ. Εφαρμόζουμε την βασισμένη σε μοντέλο κατηγοριοποίηση σύμφωνα με την μεθοδολογία που περιγράψαμε. Αρχικά, εκπαιδεύουμε ένα HMM $\lambda^k, k = 0,1$ για κάθε κλάση $C_k, k = 0,1$ και στην συνέχεια εφαρμόζουμε τα δύο μοντέλα στο σύνολο των χρηστών του training dataset το οποίο θεωρούμε ως το testing dataset. Υπολογίζουμε την πιθανοφάνεια των παρατηρήσεων του κάθε χρήστη δοθέντος

καθενός από τα δύο μοντέλα $P(Q|\lambda^k), k = 0,1$. Αναθέτουμε ως κλάση σε κάθε χρήστη την κλάση εκείνου του μοντέλου το οποίο μεγιστοποιεί αυτήν την πιθανοφάνεια, δηλαδή $c_Q = \{c_k | argmax_k[P(Q|\lambda^k)], k = 0,1\}$.

Η μέτρηση της απόδοσης της κατηγοριοποίησης παρουσιάζεται ακόλουθα με χρήση μίας μήτρας σύγχυσης (*confusion matrix*), όπου συγκρίνουμε την πραγματική κλάση κάθε χρήστη με την κλάση η οποία του ανατέθηκε από το πρόγραμμα της κατηγοριοποίησης.

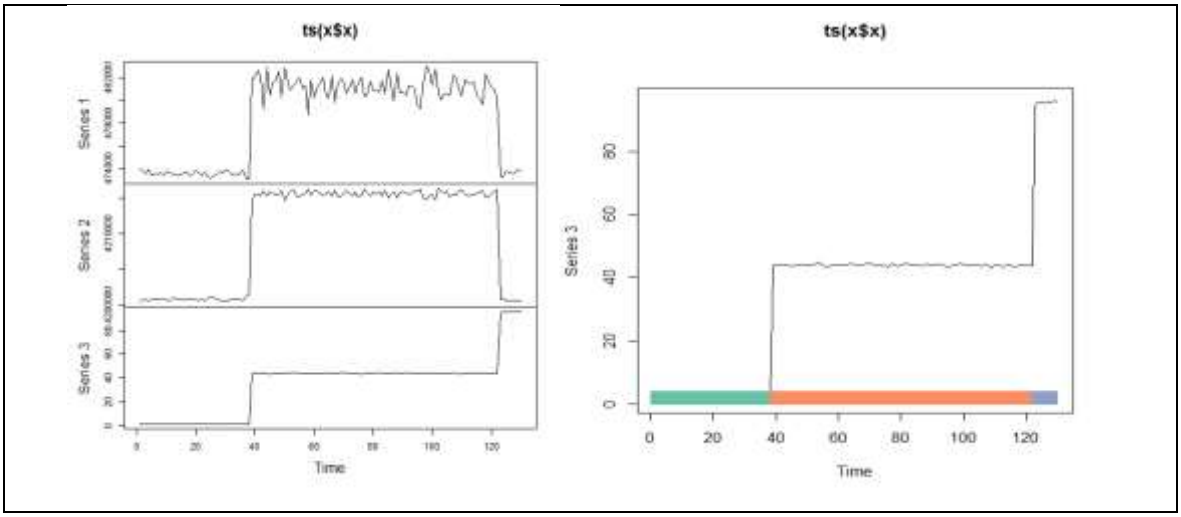
Πραγματική Κλάση	**Ανάθεση**	
	Κλάση 0	**Κλάση 1**
Κλάση 0	47	0
Κλάση 1	0	149

Πίνακας 1. Μήτρα Σύγχυσης (Confusion Matrix)

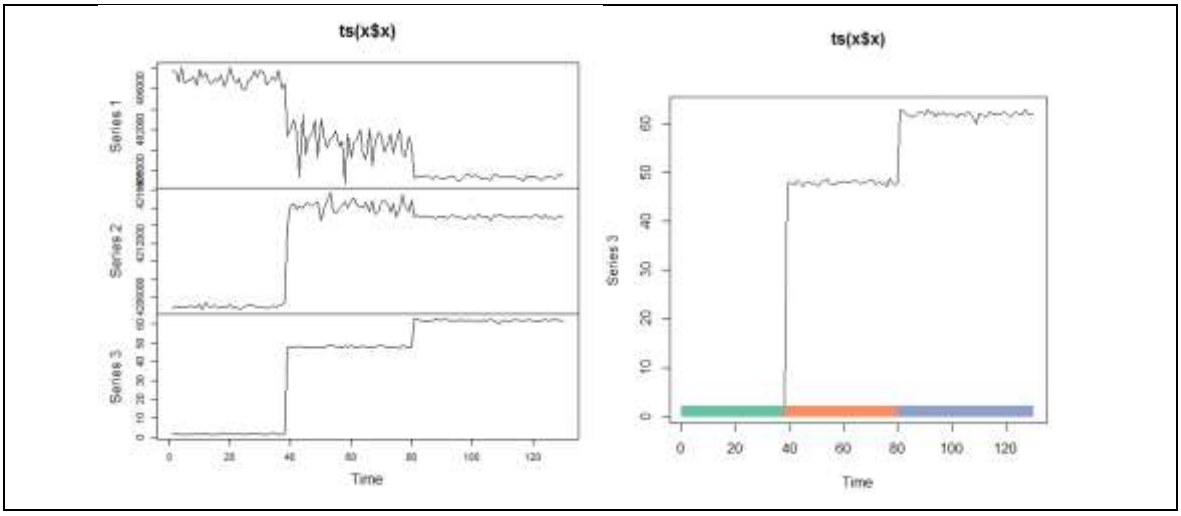
Όπως βλεπουμε στην μήτρα σύγχυσης η απόδοση του μοντέλου ήταν πολύ ικανοποιητική καθώς κατάφερε να προβλέψει απόλυτα σωστά την κλάση κάθε χρήστη του dataset. Δεδομένου ότι χρησιμοποιήσαμε ως testing dataset το ίδιο dataset που χρησιμοποιήθηκε στην εκπαίδευση των μοντέλων η απόλυτα σωστή επίδοση του μοντέλου είναι λογική.

Το δεύτερο σημαντικό στοιχείο των πειραματικών αποτελεσμάτων αφορά την πρόβλεψη του μοντέλου. Όπως είχαμε σχολιάσει στην ενότητα ορισμού των κρυφών καταστάσεων των μοντέλων, οι χρήστες της ίδιας κλάσης είναι δυνατόν να πραγματοποιούν διαφορετικό πλήθος *stop* επεισοδίων κίνησης. Έτσι, για παράδειγμα ο χρήστης με *ID* = 0 της κλάσης C_0 τερματίζει την κίνησή του στο δεύτερο επεισόδιο της κλάσης. Η πρόβλεψη για τις παρατηρούμενες τιμές της κίνησης του χρήστη πέρα από την ορθή αναγνώριση των δύο πρώτων κρυφών καταστάσεων, δηλαδή των επεισοδίων που πραγματοποίησε ο χρήστης, αποτελεί και πρόβλεψη για το πώς μπορεί αυτός να κινήθηκε συνολικά. Έστω, ότι οι καταγραφές του συγκεκριμένου χρήστη σταματούν κατά την διάρκεια του δεύτερου επεισοδίου λόγω βλάβης και εμείς αναζητούμε την θέση του χρήστη (για παράδειγμα σε ένα οποιοδήποτε σύστημα παροχής βοήθειας). Έχοντας αναγνωρίσει την κλάση του χρήστη, είμαστε σε θέση να απαντήσουμε για την πιθανότερη εξέλιξη της κίνησής του η οποία θα είναι η μετάβαση στην τελευταία κρυφή κατάσταση, δηλαδή η πραγματοποίηση του τρίτου *stop* επεισοδίου της κλάσης με τις συγκεκριμένες χωρο-χρονο-κειμενικές τιμές.

Στις Εικόνες 8 και 9 που ακολουθούν δίνουμε το γράφημα μίας τυχαίας τροχιάς της κάθε κλάσης. Η χρωματισμένη μπάρα στο κάτω μέρος του γραφήματος στο δεξιό μέρος της Εικόνας 8 και 9 απεικονίζει την κρυφή κατάσταση από την οποία προέρχονται οι αντίστοιχες παρατηρούμενες τιμές.



Εικόνα 8. Γράφημα των τριών παρατηρούμενων χαρακτηριστικών σε κάθε κρυφή κατάσταση ενός τυχαίου χρήστη της κλάσης C_0 των δεδομένων. Η πρώτη χρονοσειρά απεικονίζει τις τιμές του x, η δεύτερη του y και η τρίτη του χρόνου έναρξης t_{start} του αντίστοιχου επεισοδίου.



Εικόνα 9. Γράφημα των τριών παρατηρούμενων ακολουθιών σε κάθε κρυφή κατάσταση ενός τυχαίου χρήστη της κλάσης C_1 των δεδομένων.

Τέλος, στους πίνακες που ακολουθούν δίνουμε τις αρχικές τιμές των παραμέτρων (A, B, π) των μοντέλων όπως υπολογίστηκαν με την μεθοδολογία που περιγράψαμε και δόθηκαν ως αρχικές τιμές στο *E-step* του αλγορίθμου *EM* καθώς και το ραφινάρισμα αυτών κατά την υλοποίηση του *M-step* και επομένως την τελική εκτίμηση των παραμέτρων των δύο εκπαιδευμένων μοντέλων. Ο πίνακας π των πιθανοτήτων έναρξης που τίθεται ίσος με $\pi_i = \{1, 0, 0\}$ (δηλαδή το Bakis σύστημα ξεκινάει υποχρεωτικά από την κατάσταση S_1) δεν ραφινάρεται. Σημειώνουμε ότι, η διάρκεια παραμονής π.χ. στην πρώτη κρυφή κατάσταση του μοντέλου της κλάσης 0, η οποία υπολογίζεται σύμφωνα με την μέση τιμή της

Γεωμετρικής κατανομής, δηλαδή $E(u) = \frac{1}{(1-p_{12})} = \frac{1}{0.026} = 38.46$ (χρονικές στιγμές - δεκάλεπτα) είναι η πραγματική μέση διάρκεια του συγκεκριμένου *stop* επεισοδίου.

- **Κλάση 0**

p_{ij}	1	2	3
1	0.9762266	0.02377339	0
2	0	0.97970639	0.02029361
3	0	0	1

Πίνακας 2. Αρχική εκτίμηση των πιθανοτήτων μετάβασης του πίνακα Α.

p_{ij}	1	2	3
1	0.974	0.026	0
2	0	0.988	0.012
3	0	0	1

Πίνακας 3. Ο πίνακας Α των πιθανοτήτων μετάβασης του εκπαιδευμένου μοντέλου.

Κατάσταση	x	y	t_{start}
1	473552.0	4200811.0	1.530017
2	481274.1	4215541.1	26419.1
3	473570.0	4200776.8	57225.7

Πίνακας 4. Αρχικές τιμές των μέσων τιμών των τριών παρατηρούμενων χαρακτηριστικών σε κάθε κρυφή κατάσταση.

Κατάσταση	x	y	t_{start}
1	473559.8	4200822.0	1.538622
2	481189.23	4215585.17	26439.14
3	473645.40	4200702.59	57409.17

Πίνακας 5. Προσαρμοσμένες τιμές των μέσων τιμών των τριών παρατηρούμενων χαρακτηριστικών σε κάθε κρυφή κατάσταση μετά την εκπαίδευση του μοντέλου.

- **Κλάση 1**

p_{ij}	1	2	3
1	0.9782799	0.02172012	0
2	0	0.91694537	0.08305463
3	0	0	1

Πίνακας 6. Αρχική εκτίμηση των πιθανοτήτων μετάβασης του πίνακα Α.

p_{ij}	1	2	3
1	0.974	0.026	0
2	0	0.976	0.024
3	0	0	1

Πίνακας 7. Ο πίνακας Α των πιθανοτήτων μετάβασης του εκπαιδευμένου μοντέλου.

Κατάσταση	x	y	t_{start}
1	486906.8	4204826.0	1.530017
2	480813.41	4216094.99	28716.65
3	477427.65	4215001.52	37181.24

Πίνακας 8. Αρχικές τιμές των μέσων τιμών των τριών παρατηρούμενων χαρακτηριστικών σε κάθε κρυφή κατάσταση.

Κατάσταση	x	y	t_{start}
1	486985.2	4204867.0	1.521398
2	480799.80	4216169.28	28767.91
3	477404.82	4214993.56	37217.82

Πίνακας 9. Προσαρμοσμένες τιμές των μέσων τιμών των τριών παρατηρούμενων χαρακτηριστικών σε κάθε κρυφή κατάσταση μετά την εκπαίδευση του μοντέλου.

Συμπεράσματα

Στην παρούσα εργασία προσπαθήσαμε να λύσουμε το πρόβλημα της ευρετηρίασης δεδομένων κίνησης με μία πιθανοκρατική προσέγγιση, εφαρμόζοντας βασισμένη σε μοντέλο κατηγοριοποίηση. Χρησιμοποιήσαμε μία ιδιαίτερη κατηγορία στοχαστικών μοντέλων, τα HMMs, τα οποία έχουν ήδη εφαρμοστεί με επιτυχία σε ζητήματα ανάλυσης ποικιλόμορφων ακολουθιακών δεδομένων λόγω της ευελιξίας που παρουσιάζουν στις θεωρήσεις που υποθέτουν. Ωστόσο, όπως είναι λογικό και αναμενόμενο αποκλίσεις από τις βασικές θεωρήσεις του μοντέλου έχουν σαν αποτέλεσμα την κακή επίδοσή του. Στην περίπτωση των δύο κλάσεων των δεδομένων η εκπαίδευση των μοντέλων είδαμε ότι δεν ήταν εφικτή λόγω των μεγάλων αποκλίσεων στους χρόνους έναρξης των επεισοδίων. Ακόμη, η υλοποίηση των μοντέλων με χρήση έτοιμου πακέτου της R περιόρισε τις δυνατότητές μας για περαιτέρω πειραματισμούς καθώς έπρεπε να ικανοποιούνται όλες οι υποθέσεις του συγκεκριμένου πακέτου ώστε να μην καταλήξουμε σε επισφαλή αποτελέσματα.

Παρ' όλα αυτά, η αποτελεσματικότητα του μοντέλου στην αναπαράσταση των δύο άλλων κλάσεων και τελικά στην επιτυχή κατηγοριοποίηση των τροχιών των κινούμενων χρηστών αποτελεί μία ελπιδοφόρα βάση για την εφαρμογή των HMMs στην ανάλυση της κινητικότητας των ατόμων. Κλείνουμε αυτήν την προσπάθεια σε αυτό το σημείο βάζοντας, ουσιαστικά, μία άνω τελεία στους πειραματισμούς.

Ωστόσο, αφήνουμε ως ανοιχτά ζητήματα: Από την μία, τους περαιτέρω πειραματισμούς τόσο με το υπάρχον σύνολο δεδομένων (ίσως με εφαρμογή διαφορετικής ομαδοποίησης των δεδομένων) όσο και με δεδομένα πιο σύνθετης μορφής με μεγαλύτερο πλήθος επεισοδίων. Από την άλλη, την εφαρμογή πιο ευέλικτων μοντέλων από την ευρύτερη οικογένεια των Μαρκοβιανών μοντέλων όπως για παράδειγμα την θεώρηση μοντέλων μεγαλύτερης τάξης, την θεώρηση ημί-Μαρκοβιανών μοντέλων ή ακόμα και μή στατικών μοντέλων κ.ά.

Βιβλιογραφία

[1] Π.-Χ. Γ. Βασιλείου. Στοχαστικές Μέθοδοι στις Επιχειρησιακές Έρευνες. Εκδόσεις Ζήτη. Θεσσαλονίκη, 1999.

[2] Nikos Pelekis, Yannis Theodoridis and Davy Janssens. On the Management and Analysis of Our Lifesteps. ACM SIGKDD Explorations, Volume 15 Issue 1, pages 23-32, 2013.

[3] Christine Parent, Stefano Spaccapietra, Chiara Renso, Gennady Andrienko, Natalia Andrienko, Vania Bogorny, Maria Luisa Damiani, Aris Gkoulalas-Divanis, Jose Macedo, Nikos Pelekis, Yannis Theodoridis and Zhixian Yan. Semantic Trajectories Modeling and Analysis. ACM Computing Surveys (CSUR), Volume 45 Issue 4, Article No. 42, August 2013.

[4] Nikos Pelekis, Stylianos Sideridis, Panagiotis Tampakis and Yannis Theodoridis. Simulating our LifeSteps by Example. ACM Transactions on Spatial algorithms and Systems, to appear.

[5] Lawrence R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE, VOL. 77, NO. 2, pages 257-286, February 1989.

[6] Eric Fosler-Lussier. Markov Models and Hidden Markov Models: A Brief Tutorial, Technical Report (TR-98-041), International Computer Science Institute, Berkeley, California, December 1998.

[7] Shun-Zheng Yu. Hidden semi-Markov models. Artificial Intelligence, Vol. 174, Issue 2, pages 215-243, February 2010.

[8] Alexios Kotsifakos, Vassilis Athitsos, Panagiotis Papapetrou, Jaakko Hollmen and Dimitrios Gunopulos. Model-Based Search in Large Time Series Databases. Proceedings of the 4th International Conference on Pervasive Technologies Related to Assistive Environments, Article No. 36, 2011.

[9] J. Ferguson. Variable duration models for speech. Proceedings of the Symposium on the Application of Hidden Markov Models to Text and Speech, pages 143–179, October 1980.

[10] J. O'Connell and S. Hojsgaard. Hidden Semi Markov Models for Multiple Observation Sequences: The mhsmm Package for R. Journal of Statistical Software, Vol. 39, Issue 4, March 2011.