ΑΝΑΛΥΣΗ ΑΚΟΛΟΥΘΙΑΚΩΝ ΔΕΔΟΜΕΝΩΝ ΜΕ ΤΗ ΧΡΗΣΗ ΜΟΝΤΕΛΩΝ MARKOV

Η
ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ ΕΞΕΙΔΙΚΕΥΣΗΣ

Υποβάλλεται στην

ορισθείσα από την Γενική Συνέλευση Ειδικής Σύνθεσης
του Τμήματος Πληροφορικής
Εξεταστική Επιτροπή

από τον

Ανδρέα Κακολύρη

ως μέρος των Υποχρεώσεων

για τη λήψη

του

ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΔΙΠΛΩΜΑΤΟΣ ΣΤΗΝ ΠΛΗΡΟΦΟΡΙΚΗ

ΜΕ ΕΞΕΙΔΙΚΕΥΣΗ ΣΤΙΣ ΤΕΧΝΟΛΟΓΙΕΣ-ΕΦΑΡΜΟΓΕΣ

Ιανουάριος 2006

ΕΥΧΑΡΙΣΤΙΕΣ

Νιώθω την ανάγκη να ευχαριστήσω θερμά τον επιβλέποντά μου καθηγητή κ. Κωνσταντίνο Μπλέκα. Ο λόγος δεν είναι μόνο η ουσιαστική και συνεχής καθοδήγηση που με βοήθησαν σε όλη τη διάρκεια της εκπόνησης της παρούσας εργασίας και υπήρξαν καθοριστικοί παράγοντες στην επιτυχία της. Εξίσου σημαντικός παράγοντας – αν όχι σημαντικότερος – υπήρξε για μένα η ηθική και ψυχολογική στήριξη που μου παρείχε σε όλη τη διάρκεια της συνεργασίας μας. Ιδιαίτερα τους τελευταίους μήνες, όταν και βρέθηκα αντιμέτωπος με διάφορες καταστάσεις, η ενθάρρυνσή του με βοήθησε να παραμείνω συγκεντρωμένος και να μπορέσω να φέρω σε πέρας με επιτυχία την εργασία και γι' αυτό το λόγο είμαι ευγνώμων.

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΥΡΕΤΗΡΙΟ ΠΙΝΑΚΩΝ

ΕΥΡΕΤΗΡΙΟ ΣΧΗΜΑΤΩΝ

ΠΕΡΙΛΗΨΗ

Ανδρέας Κακολύρης του Σταματίου και της Αγγελικής. MSc, Τμήμα Πληροφορικής, Πανεπιστήμιο Ιωαννίνων, Ιανουάριος, 2006, Ανάλυση Ακολουθιακών Δεδομένων με τη χρήση μοντέλων Markov. Επιβλέποντας: Κωνσταντίνος Μπλέκας.

Στην παρούσα εργασία προσεγγίζουμε το πρόβλημα της ανάλυσης ακολουθιακών δεδομένων. Τα δεδομένα θεωρούμε ότι αποτελούνται από ακολουθίες συμβολοσειρών και η ανάλυσή τους έγκειται στον προσδιορισμό ομάδων ακολουθιών με κοινά χαρακτηριστικά. Το πρόβλημα της ομαδοποίησης μπορεί να επιλυθεί με τη χρήση μικτών μοντέλων Markov, τα οποία εκπαιδεύουμε με τον γνωστό αλγόριθμο ΕΜ. Προτείνουμε μια νέα προσέγγιση στην εκπαίδευση μικτών μοντέλων Markov. Κεντρική ιδέα είναι η αυξητική κατασκευή του μικτού μοντέλου. Αποτέλεσμα της διαδικασίας είναι οι παράμετροι των μοντέλων να αρχικοποιούνται με κριτήριο τη μεγιστοποίηση της πιθανοφάνειας του μοντέλου, ενώ ταυτόχρονα το πλήθος των συνιστωσών του μικτού μοντέλου καθορίζεται δυναμικά κατά την εκπαίδευση. Συνεπώς, αντιμετωπίζονται τα δυο βασικότερα προβλήματα του μικτού μοντέλου: Ο καθορισμός του πλήθους των συνιστωσών, το οποίο αντιστοιχεί στο πλήθος των ομάδων στα δεδομένα και η αρχικοποίηση των παραμέτρων του μοντέλου.

EXTENDED ABSTRACT IN ENGLISH

Kakoliris, Andreas, A.K. MSc, Computer Science Department, University of Ioannina, Greece. January, 2006. Markov Model Approach to Sequential Data Analysis. Thesis Supervisor: Konstantinos Blekas.

The purpose of this thesis is to study the problem of sequential data analysis. The term sequential data refers to sequences of discrete symbols. Such data are found in many areas of interest such as text recognition, DNA analysis or user pattern recognition. The analysis we consider is the separation of sequential data sets into clusters with common patterns. The clustering is achieved by learning a mixture of first order Markov models using the Expectation – Maximization algorithm.

Markov models can be used to represent the underlying stochastic process, responsible for generating the observation sequences. A sequence is produced by choosing the initial state (symbol) according to the initial state probabilities of the model and then by choosing the next states (symbols) according to the transition probabilities of the model.

Hidden Markov models are more powerful and consequently more complex than simple Markov models. The difference is that the states in a hidden Markov Model no longer represent the observation symbols but the observation is a probabilistic function of the state.

A simple or hidden Markov model can be created to represent the process responsible for creating sequences of a single cluster. Mixtures of Markov models can be created

to model the process of creating a set of sequences from different clusters. The EM algorithm can be used to train such mixture models so they can be used for clustering. However, this technique faces two major problems. Firstly, the number of mixtures used is not usually known in advance and secondly, initial parameter values are required for the training of the mixture models.

We propose a novel methodology for creating and training mixture models of simple Markov models. The key idea is to incrementally add components to the mixture: We begin by constructing the one – component global model of the dataset. At each step we choose a new model from a pool of pre – constructed models and add it to the m mixture. The added model is trained using the EM algorithm (Partial EM) and the resulting m+1 mixture model is then trained using EM (General EM). The algorithm stops when the desired mixture has been created.

The number of components is determined dynamically, during the incremental process. We use a fast and easy algorithm to obtain a rough clustering of the dataset. A pool of Markov models is then created by these clusters. The component added is chosen, so that the likelihood of the mixture is maximized at each step.

Experiments to synthetic as well as to real data demonstrate the effectiveness of the incremental method. Incrementally created mixture models are equal or better to the best models created by other methods with predetermined number of components in most cases. Only when the data display very noisy patterns the incremental method fails to perform well, which is considered natural, since the clusters can no longer be considered homogeneous and the other mixture models also fail to perform well.

Further study includes the use of better methods for obtaining better pre – constructed models to use in the algorithm, further analysis of the stopping criterion for the incremental algorithm and application of a similar method to hidden Markov models.

ΚΕΦΑΛΑΙΟ 1. ΕΙΣΑΓΩΓΗ

1.1. Στόχοι

Ο στόχος της διατριβής είναι η μελέτη του προβλήματος της ανάλυσης ακολουθιακών δεδομένων. Με τον όρο ακολουθιακά δεδομένα εννοούμε δεδομένα που αποτελούνται από ακολουθίες διακριτών συμβόλων. Σε πολλές περιοχές ερευνητικού – και όχι μόνο – ενδιαφέροντος συναντώνται ακολουθιακά δεδομένα: Στην ανάλυση και αναγνώριση κειμένου τα δεδομένα είναι ακολουθίες συμβόλων. Η ανάλυση γεννετικών χαρακτηριστικών χρησιμοποιεί ακολουθίες DNA. Η ανάλυση της συμπεριφοράς χρηστών ενός δικτυακού τόπου απαιτεί την ύπαρξη ακολουθιών που περιγράφουν τις ενέργειες που πραγματοποιεί. Η εξέλιξη της τιμής μετοχών κατά τη διάρκεια κάποιας χρονικής περιόδου μπορεί να αναπαρασταθεί με ακολουθίες τιμών της μετοχής για περαιτέρω ανάλυση και πρόβλεψη της πορείας της τιμής των μετοχών. Οι επιλογές προϊόντων των πελατών μιας εταιρείας μπορούν να καταγραφούν, προκειμένου να γίνει πρόβλεψη των προτιμήσεών τους και να προσαρμοστούν κατάλληλα οι προσφερόμενες υπηρεσίες.

Αυτές είναι μερικές μόνο από τις εφαρμογές στις οποίες εμφανίζονται διακριτά ακολουθιακά δεδομένα. Είναι φανερό, ότι η ανάλυσή τους παρουσιάζει σημαντικό ερευνητικό και πρακτικό ενδιαφέρον. Η ανάλυση στην οποία αναφερόμαστε στην παρούσα εργασία αναφέρεται στο διαχωρισμό των ακολουθιακών δεδομένων σε ομάδες με κοινά χαρακτηριστικά. Μια τέτοια ομαδοποίηση μπορεί να προσφέρει

πολύτιμες πληροφορίες, ανάλογα με τη φύση των δεδομένων. Για παράδειγμα, η ομαδοποίηση ακολουθιών DNA αναγνωρίζει ομάδες πληθυσμού με κοινά γεννετικά χαρακτηριστικά, ενώ η ομαδοποίηση ακολουθιών ενεργειών των χρηστών αποκαλύπτει κοινή συμπεριφορά και προτιμήσεις των χρηστών.

Για την πραγματοποίηση της ανάλυσης, χρησιμοποιούμε μοντέλα Markov, τα οποία προσωμοιώνουν το μηχανισμό παραγωγής των ακολουθιακών δεδομένων. Τα μοντέλα Markov μπορούν να εκπαιδευτούν κατάλληλα, ώστε να επιλύουν το πρόβλημα της ομαδοποίησης. Ο σκοπός της παρούσας διατριβής είναι να προτείνει μια νέα προσέγγιση στο πρόβλημα της εκπαίδευσης των μικτών μοντέλων Markov, η οποία αντιμετωπίζει ορισμένα από τα προβλήματα που παρουσιάζουν οι υπάρχοντες αλγόριθμοι.

1.2. Δομή της Διατριβής

Η διατριβή περιέχει 6 κεφάλαια: Στο Κεφάλαιο 1 περιγράφουμε το στόχο της εργασίας και διατυπώνουμε το πρόβλημα της ομαδοποίησης ακολουθιακών δεδομένων. Στο Κεφάλαιο 2 παρουσιάζουμε τη θεωρία για τα χαρακτηριστικά και τη λειτουργία του απλού και του κρυμμένου μοντέλου Markov για διακριτά δεδομένα. Στο Κεφάλαιο 3 περιγράφουμε τα μικτά μοντέλα Markov καθώς και τον αλγόριθμο ΕΜ που χρησιμοποιείται για την εκπαίδευσή τους. Στο κεφάλαιο 4 προτείνουμε μια μέθοδο για την αυξητική δημιουργία ενός μικτού μοντέλου Markov. Στο κεφάλαιο 5 εκθέτουμε τα αποτελέσματα των πειραμάτων και αξιολογούμε τις επιδόσεις του αυξητικού αλγορίθμου συγκρίνοντάς τον με τα παραδοσιακά μοντέλα Markov. Τέλος, στο Κεφάλαιο 6 συνοψίζουμε το περιεχόμενο της διατριβής και προτείνουμε μελλοντικές επεκτάσεις που μπορούν να γίνουν.

ΚΕΦΑΛΑΙΟ 2. ΜΟΝΤΕΛΑ MARKOV

2.1. Στοχαστική διαδικασία Markov

Στο πρόβλημά μας έχουμε ένα σύνολο ακολουθιών $X=\{X_i\}$ με i=1:N. Μια ακολουθία $X_i=\{Xi_l\}$ με l=1:T(i) αποτελείται από ένα πλήθος διακριτών συμβόλων $V=\{v_k\}$με k=1:K. Μια τέτοια ακολουθία συμβόλων μπορεί να θεωρηθεί ότι αναπαριστά την αλληλουχία μεταβάσεων μεταξύ ενός συνόλου καταστάσεων, όπου οι καταστάσεις αναπαρίστανται από τα σύμβολα της ακολουθίας. Επομένως, μια ακολουθία X_i μπορεί να θεωρηθεί ως το αποτέλεσμα μιας στοχαστικής διαδικασίας, σε διακριτό χρόνο t, με διακριτό χώρο καταστάσεων. Το μήκος της ακολουθίας εκφράζει το πλήθος των χρονικών στιγμών της ακολουθίας, ενώ το πλήθος των διακριτών συμβόλων το χώρο των καταστάσεων της στοχαστικής διαδικασίας. Επιπλέον, θεωρούμε ότι το αποτέλεσμα κάθε διαδικασίας δεν επηρεάζει τα αποτελέσματα των υπόλοιπων διαδικασιών, δηλαδή οι ακολουθίες είναι ανεξάρτητες μεταξύ τους.

Θεωρούμε ότι η παραπάνω στοχαστική διαδικασία αποτελεί μια *Μαρκοβιανή διαδικασία ή Μαρκοβιανή αλυσίδα*. Για τη διαδικασία Markov που περιγράφουμε ισχύουν οι παρακάτω ιδιότητες **[15]**:

Εάν μια χρονική στιγμή t βρισκόμαστε σε μια κατάσταση S_i η πιθανότητα να βρεθούμε σε μια κατάσταση S_j τη χρονική στιγμή t+1 εξαρτάται μόνο από την S_i και όχι από τις καταστάσεις που προηγήθηκαν της χρονικής στιγμής t, δηλαδή: $P(q_{t+1}=S_j \mid q_t=S_i, \ldots, q_1=S_k) = P(q_t=S_i \mid q_{t-1}=S_j)$.
Η ιδιότητα ομογένειας του χρόνου: Δεδομένου ότι τη χρονική στιγμή t η διαδικασία βρίσκεται στην κατάσταση S_i, η πιθανότητα τη χρονική στιγμή t+1 να βρισκόμαστε στην κατάσταση S_j είναι ανεξάρτητη από τη χρονική στιγμή t, δηλαδή οι πιθανότητες μετάβασης παραμένουν αμετάβλητες σε όλη τη διάρκεια της μαρκοβιανής διαδικασίας.

Το σύνολο των καταστάσεων και οι συσχετίσεις μεταξύ τους που αναπαριστούν την Μαρκοβιανή διαδικασία αποτελούν ένα μοντέλο Markov. Υπάρχουν τρία βασικά προβλήματα που μπορούμε να λύσουμε στα μοντέλα Markov **[11]**:

Η εύρεση της πιθανότητας μιας ακολουθίας X_i για ένα δεδομένο μοντέλο (πιθανοφάνεια).
Η εύρεση της πιθανότερης ακολουθίας καταστάσεων για την παραγωγή μιας ακολουθίας από ένα μοντέλο.
Η προσαρμογή των παραμέτρων του μοντέλου, ώστε να ταιριάζει καλύτερα στις παρατηρήσεις.

Ο απώτερος στόχος είναι να γίνει κατάλληλη ομαδοποίηση των ακολουθιών αυτών με βάση κοινά χαρακτηριστικά τους και οι ομάδες που θα σχηματιστούν να είναι όσο το δυνατόν περισσότερο ομογενείς.

Στη συνέχεια θα δούμε με ποιο τρόπο, επιλύνοντας τα προβλήματα 1 και 3, μπορούμε να επιτύχουμε την ομαδοποίηση των ακολουθιών. Πρώτα όμως παρουσιάζουμε περισσότερο αναλυτικά τη θεωρία για τα μοντέλα Markov. Παρουσιάζουμε δυο παραλλαγές, τα απλά μοντέλα Markov (Markov Models) και τα κρυμμένα μοντέλα Markov (Hidden Markov Models) **[11, 8]**.

2.2. Μοντέλα Markov (Markov Models)

Ορίζουμε το μοντέλο Markov που αναπαριστά μια μαρκοβιανή διαδικασία. Στην περίπτωση των μοντέλων Markov πρώτης τάξης που μας απασχολούν, ισχύουν οι ιδιότητες απώλειας μνήμης και ομογένειας των μαρκοβιανών αλυσίδων που αναφέραμε παραπάνω.

Το μοντέλο Markov περιέχει καταρχάς, το σύνολο των καταστάσεων της διαδικασίας. Την πιθανότητα, όντας κατάσταση S_i τη στιγμή t, στην επόμενη χρονική στιγμή t+1 να βρεθούμε στην κατάσταση S_j, $a_{ji} = P(q_{t+1}=S_i \mid q_t=S_i) \ \forall i,j$ την ονομάζουμε *πιθανότητα μετάβασης* του μοντέλου Markov. Προφανώς, από τη στιγμή που μιλάμε για πιθανότητες ισχύει ότι $a_{ij} \geq 0 \ \forall i,j$, καθώς επίσης και ότι $\sum_{j=1}^{K} a_{ij} = 1 \ \forall i$.

Εκτός από τις καταστάσεις και τις πιθανότητες μετάβασης, ένα ακόμα ζήτημα είναι η επιλογή της πρώτης κατάστασης της διαδικασίας. Ορίζουμε την πιθανότητα π_i η κατάσταση Si να είναι η αρχική κατάσταση της αλυσίδας, δηλαδή $\pi_i = P(q_1 = S_i) \ \forall i$.

Ισχύει κι εδώ $\pi_i \geq 0 \ \forall i$, καθώς και $\sum_{i=1}^{K} \pi_i = 1 \ \forall i$. Οι πιθανότητες π ονομάζονται *πιθανότητες έναρξης*.

Συνεπώς, ένα μοντέλο Markov περιγράφεται από το σύνολο παραμέτρων των πιθανοτήτων έναρξης και μεταβάσεων. Θεωρώντας ότι υπάρχουν K καταστάσεις στο μοντέλο, έχουμε το σύνολο των πιθανοτήτων έναρξης: $P = [\pi_1, \pi_2, \ldots, \pi_K]$ και των πιθανοτήτων μετάβασης:

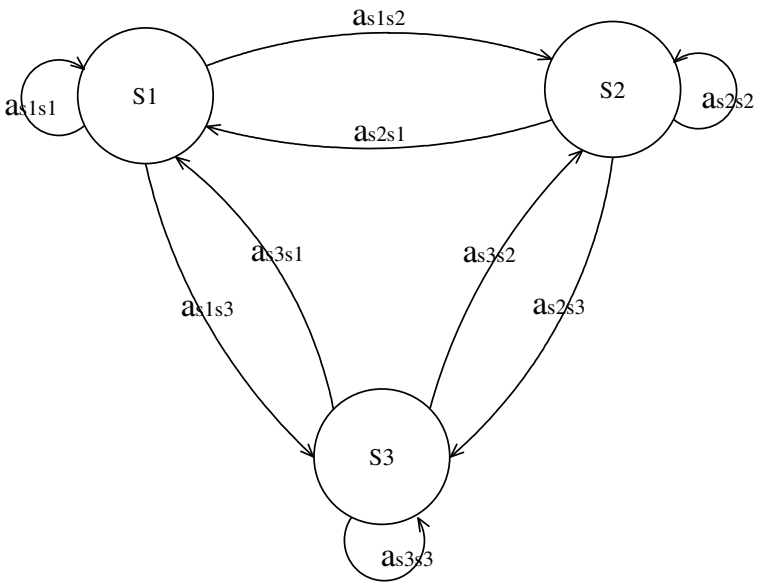
A=

a_{11}	a_{12}	…	a_{1K}
a_{21}	a_{22}	…	a_{2K}
⋮	⋮		⋮
a_{K1}	a_{K2}		a_{KK}

Επομένως, το σύνολο των παραμέτρων του μοντέλου Θ είναι οι δυο παραπάνω πίνακες. Ο πίνακας αρχικών καταστάσεων P και ο πίνακας μεταβάσεων A και το μοντέλο ορίζεται με τη βοήθεια των παραμέτρων του ως Θ={P, A}.

Επίσης, ένα μοντέλο Markov μπορεί να οπτικοποιηθεί ως ένας γράφος καταστάσεων και κατευθυνόμενων μεταβάσεων. Οι κόμβοι του γράφου αντιστοιχούν στις καταστάσεις του μοντέλου και οι ακμές στις πιθανότητες μετάβασης μεταξύ των καταστάσεων. Για κάθε κατάσταση του μοντέλου S_i υπάρχει ένα σύνολο πιθανοτήτων μετάβασης a_{ij} για όλες τις καταστάσεις S_j του μοντέλου (συμπεριλαμβανομένης τις ίδιας της S_i).

Ένα παράδειγμα μοντέλου Markov με τρεις καταστάσεις δίνεται στο Σχήμα 2.1.



Σχήμα 2.1 Κατευθυνόμενος Γράφος Μοντέλου Markov

Έχοντας κατασκευάσει το μοντέλο, η διαδικασία Markov είναι η ακόλουθη:

Επιλέγουμε την πρώτη κατάσταση σύμφωνα με τις πιθανότητες έναρξης π_i του μοντέλου.

Επιλέγουμε στοχαστικά τις επόμενες καταστάσεις με βάση τις πιθανότητες μετάβασης, μέχρι το τέλος της ακολουθίας.

Έχοντας ορίσει πλήρως το μοντέλο Markov μπορούμε να βρούμε την πιθανότητα μια ακολουθία X_i από σύμβολα του V να παραχθεί από το μοντέλο Θ. Η ζητούμενη πιθανότητα $P(X_i \mid \Theta)$ υπολογίζεται ως εξής:

$$P(Xi \mid \Theta) = P(X_{i1}, X_{i2}, \ldots, X_{iT(i)} \mid \Theta) = P(X_{i1}) * P(X_{i2} \mid X_{i1}) * \ldots * P(X_{iT(i)} \mid X_{iT(i)-1})$$

Υπενθυμίζουμε ότι τα X_{ik} παίρνουν τιμές στο σύνολο V, ότι οι καταστάσεις του μοντέλου αντιστοιχούν στα στοιχεία του V και ότι θεωρούμε πως η παρατήρηση τη χρονική στιγμή t εξαρτάται μόνο από την παρατήρηση τη στιγμή t-1.

Επομένως, η πιθανότητα $P(X_{i1})$ αντιστοιχεί στην πιθανότητα έναρξης της αντίστοιχης κατάστασης του μοντέλου, δηλαδή $P(X_{i1} = v_k) = \pi_k$.

Αντίστοιχα, οι πιθανότητες $P(X_{it} \mid X_{it-1})$ αντιστοιχούν στις πιθανότητες μεταβάσεων του μοντέλου, δηλαδή $P(X_{it} = v_f \mid X_{it-1} = v_g) = a_{fg}$.

Άρα, $P(Xi \mid \Theta) = \pi_{Xi1} * a_{Xi1X2} * \ldots * a_{XiT(i)-1XiT(i)}$ ➔ $P(Xi \mid \Theta) = \pi_{Xi1} \bullet \prod_{t=1}^{Ti-1} a_{XitXit+1}$

Απλοποιούμε την παραπάνω σχέση. Ορίζουμε τη συνάρτηση δ: $\delta(X_{it}, S_k) = \begin{cases} 1, & \alpha\nu \ Xit=Sk \\ 0, & \delta\iota\alpha\phi o\rho\varepsilon\tau\iota\kappa\acute{\alpha} \end{cases}$. Με τη βοήθεια της δ μπορούμε να γράψουμε καλύτερα την πιθανότητα π, ως εξής: $\pi_{Xi1}^{m} = \prod_{k=1}^{K} \pi_k^{m\,\delta(Xi1, Sk)}$ **(Εξίσωση 2.1)**, όπου K είναι το πλήθος των συμβόλων του αλφαβήτου.

Παρόμοια με τη $\delta(X_{it}, S_f, X_{it+1}, S_g) = \begin{cases} 1, & \alpha\nu \ Xit=Sf \ \kappa\alpha\iota \ Xit+1=Sg \\ 0, & \delta\iota\alpha\phi o\rho\varepsilon\tau\iota\kappa\acute{\alpha} \end{cases}$ μπορούμε να γράψουμε την a ως: $a^{m}_{Xit\,Xit+1} = \prod_{f=1}^{K}\prod_{g=1}^{K} a_{fg}^{m\,\delta(Xit, Sf, Xit+1, Sg)}$ **(Εξίσωση 2.2)**.

Επομένως, με τη βοήθεια των εξισώσεων 2.1 και 2.2 γράφουμε: $P(X_i \mid \Theta_m) = \prod_{k=1}^{K} \pi_k^{m\,\delta(Xi1, Sk)} \prod_{t=1}^{Ti-1}\prod_{f=1}^{K}\prod_{g=1}^{K} a_{fg}^{m\,\delta(Xit, Sf, Xit+1, Sg)}$ **(Εξίσωση 2.3)**.

Η ποσότητα $P(X_i \mid \Theta_m)$ ονομάζεται πιθανοφάνεια της X_i για το μοντέλο Θm και αποτελεί το μέτρο με το οποίο αξιολογούμε την ικανότητα των μοντέλων να ταιριάζουν στα δεδομένα.

2.3. Κρυμμένα Μοντέλα Markov

Τα απλά μοντέλα Markov αποτελούν καλή προσέγγιση της διαδικασίας παραγωγής των παρατηρήσεων, αλλά θεωρούν ότι το πλήθος καταστάσεων του μοντέλου είναι ίσο με το πλήθος των διακριτών παρατηρήσεων και κάθε κατάσταση του μοντέλου αντιστοιχεί μόνο σε μια παρατήρηση. Αυτή η απλοποίηση καθιστά το απλό μοντέλο Markov μη αποδοτικό σε περιπτώσεις που η πραγματική διαδικασία παραγωγής των παρατηρήσεων είναι πιο σύνθετη. Τα κρυμμένα μοντέλα Markov μπορούν να ανταποκριθούν καλύτερα σε τέτοιες περιπτώσεις. Η διαφορά τους έγκειται στο ότι το πλήθος των καταστάσεων δεν είναι απαραίτητα ίσο με το πλήθος των διακριτών παρατηρήσεων και σε κάθε κατάσταση η παρατήρηση παράγεται από μια στοχαστική διαδικασία, η οποία **δεν** είναι άμεσα ορατή. Αυτή η εκ πρώτης όψεως μικρή διαφοροποίηση καθιστά τα κρυμμένα μοντέλα Markov πιο ισχυρά και τους επιτρέπει να προσεγγίσουν με μεγαλύτερη ακρίβεια τα προβλήματα **[11, 8]**.

Για να γίνει περισσότερο αντιληπτό το κρυμμένο μοντέλο παραθέτουμε το παρακάτω παράδειγμα. Έστω ένα πείραμα, στο οποίο υπάρχουν 3 καλάθια, σε κάθε ένα από τα οποία υπάρχουν χρωματιστές σφαίρες 4 διαφορετικών χρωμάτων (πράσινο – Π, κόκκινο – Κ, μπλε – Μ, λευκό – Λ). Σε κάθε καλάθι ο αριθμός των σφαιρών και η

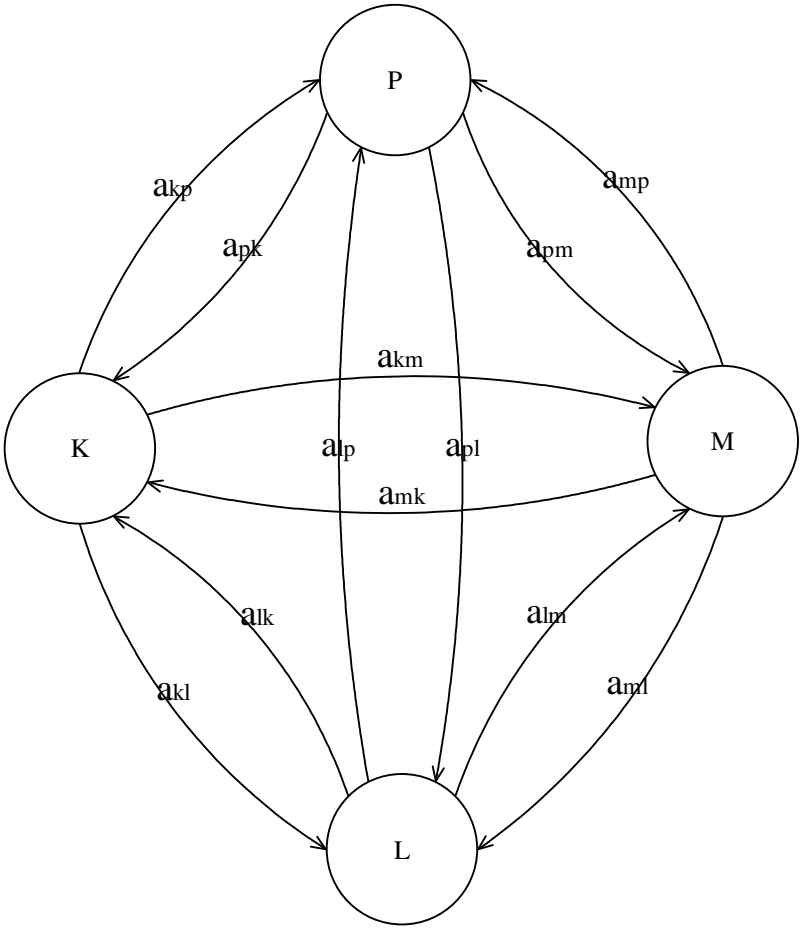
αναλογία τους δεν είναι ίδια. Επομένως, αν τραβήξουμε μια σφαίρα από ένα καλάθι η πιθανότητα να έχει κάποιο χρώμα είναι εν γένει διαφορετική για κάθε καλάθι. Έστω ότι οι πιθανότητες δίνονται από τον παρακάτω πίνακα Β:

Β=

	Π	Κ	Μ	Λ
Καλάθι 1	0.4	0.1	0.2	0.3
Καλάθι 2	0.6	0	0.2	0.2
Καλάθι 3	0.3	0.5	0.1	0.1

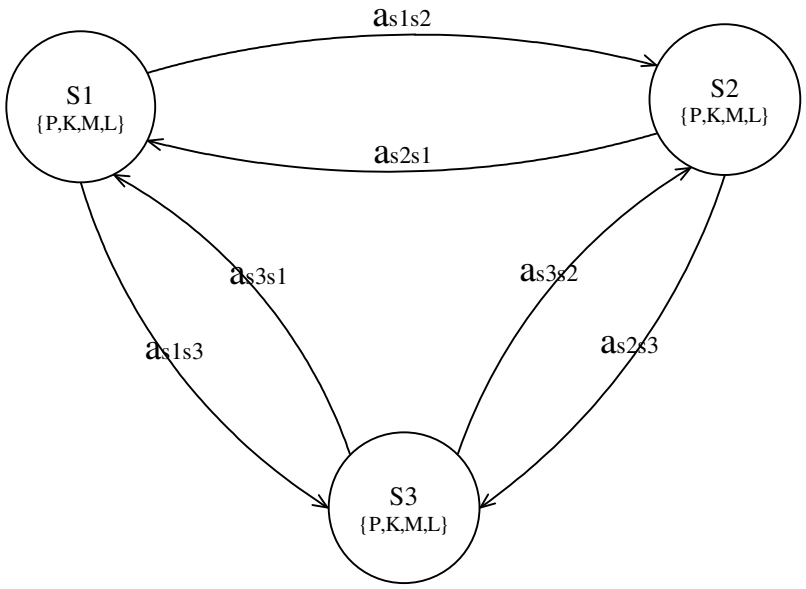
Έστω ότι επιλέγουμε Ν σφαίρες στη σειρά από τα καλάθια. Σε κάθε βήμα επιλέγουμε με κάποιο τυχαίο τρόπο ένα καλάθι και από αυτό τραβάμε επίσης στην τύχη μια σφαίρα. Τελικά παράγεται μια ακολουθία παρατηρήσεων πχ. Π Λ Λ Κ Μ Μ Π Π Κ Λ Λ Κ κ.ο.κ.

Το απλό μοντέλο Markov που μπορεί να κατασκευαστεί έχει τέσσερις καταστάσεις, μια για κάθε χρώμα και απεικονίζεται στο Σχήμα 2.2. Στο μοντέλο αυτό εκκινώντας από ένα χρώμα επιλέγουμε διαδοχικά τα χρώματα που εμφανίζονται κάθε στιγμή. Εντούτοις, στο πραγματικό πρόβλημα υπάρχει και η επιλογή του καλαθιού, από το οποίο προκύπτει η κάθε σφαίρα. Η επιλογή καλαθιού επηρεάζει το χρώμα των σφαιρών, αφού σε κάθε καλάθι οι πιθανότητες για επιλογή κάποιου χρώματος δεν είναι απαραίτητα οι ίδιες. Ακόμα, αν θεωρήσουμε ότι η επιλογή των καλαθιών δε γίνεται με την ίδια πιθανότητα για κάθε καλάθι η επιλογή καλαθιού γίνεται ακόμα πιο καθοριστική για το αποτέλεσμα του πειράματος. Συνεπώς, το απλό μοντέλο δεν μπορεί να αναπαραστήσει αρκετά καλά το συγκεκριμένο πρόβλημα.



Σχήμα 2.2 Παράδειγμα Απλού Μοντέλου Markov

Θεωρούμε τώρα ένα κρυμμένο μοντέλο Markov με τρεις καταστάσεις. Κάθε κατάσταση αντιστοιχεί σε ένα από τα τρία καλάθια του πειράματος. Σε κάθε κατάσταση μπορούν να επιλεγούν σφαίρες και των τεσσάρων χρωμάτων με κάποιες πιθανότητες, όπως φαίνεται στο Σχήμα 2.3.



Σχήμα 2.3 Παράδειγμα Κρυφού Μοντέλου Markov

Επιλέγοντας κατάλληλα τις πιθανότητες αρχικών καταστάσεων και μεταβάσεων μπορούμε να προσομοιώσουμε την επιλογή των καλαθιών σε κάθε βήμα. Αν ορίσουμε τις πιθανότητες εμφάνισης των χρωμάτων των σφαιρών σε κάθε κατάσταση να είναι αυτές που είχαμε στον πίνακα Β, τότε έχουμε κατασκευάσει ένα απλό κρυμμένο μοντέλο Markov, το οποίο προσεγγίζει το πραγματικό πρόβλημα καλύτερα από το απλό μοντέλο.

Επομένως, για να ορίσουμε ένα κρυμμένο μοντέλο Markov χρειάζεται να ορίσουμε τα παρακάτω στοιχεία **[8, 11]**:

Το πλήθος των καταστάσεων Κ του μοντέλου. Ο καθορισμός του Κ είναι ιδιαίτερα σημαντικός για την επιλογή ενός καλού μοντέλου. Το πλήθος των καταστάσεων έχει άμεση σχέση με το πρόβλημα που επιλύουμε, όπως στο παραπάνω παράδειγμα κάθε κατάσταση αντιπροσώπευε ένα καλάθι. Συμβολίζουμε τις καταστάσεις $S = \{S_k\}$ με k=1:K.

Το πλήθος των παρατηρήσιμων συμβόλων L που παράγονται σε κάθε κατάσταση, όπως ήταν τα χρώματα σε κάθε μπάλα στο παράδειγμα. Συμβολίζουμε τα σύμβολα $L = \{v_l\}$ με l=1:L.

Τις πιθανότητες έναρξης των καταστάσεων, δηλαδή την πιθανότητα μια κατάσταση να επιλεγεί πρώτη σε ένα πείραμα. Στο παράδειγμα την πιθανότητα επιλογής του καλαθιού, από το οποίο θα τραβάγαμε την πρώτη μπάλα. Όπως στο απλό μοντέλο συμβολίζουμε $\pi_i = P(q_1 = S_i) \ \forall i$, ενώ ισχύουν ξανά οι περιορισμοί $\pi_i \geq 0 \ \forall i$ και $\sum_{i=1}^{K} \pi_i = 1$.

Τις πιθανότητες μεταβάσεων μεταξύ των καταστάσεων. Στο παράδειγμα η επιλογή του επόμενου καλαθιού για να τραβήξουμε νέα σφαίρα. Συμβολίζουμε την πιθανότητα μετάβασης $a_{ij} = P(q_{t+1}=S_i \mid q_t=S_i) \ \forall i,j$. Όπως στο απλό μοντέλο ισχύει ότι $a_{ij} \geq 0 \ \forall i,j$, καθώς επίσης και ότι $\sum_{j=1}^{K} a_{ij} = 1 \ \forall i$.

Τις πιθανότητες εμφάνισης των παρατηρήσιμων συμβόλων σε κάθε κατάσταση. Τις πιθανότητες δηλαδή ενώ βρισκόμαστε σε μια κατάσταση να εμφανιστεί κάποιο από τα σύμβολα του αλφαβήτου. Στο παράδειγμά μας την πιθανότητα ενώ έχουμε επιλέξει ένα καλάθι να τραβήξουμε σφαίρα ενός συγκεκριμένου χρώματος. Συμβολίζουμε $b_k(l)$ την πιθανότητα να εμφανιστεί το σύμβολο v_l στην κατάσταση s_k ,$P(X_{it}=v_l \mid q_t=S_k)$. Αφού σε κάθε κατάσταση μπορεί να εμφανιστούν όλα τα σύμβολα και μόνο αυτά θα ισχύουν οι περιορισμοί $b_k(l) \geq 0 \ \forall k,l$, όπως και $\sum_{l=1}^{L} b_k(l) = 1 \ \forall k$.

Έχοντας ορίσει τις παραμέτρους του μοντέλου η διαδικασία παραγωγής ακολουθιών συμβόλων είναι η ακόλουθη:

Επιλογή αρχικής κατάστασης k με βάση τις πιθανότητες έναρξης.
Παραγωγή ενός συμβόλου l σύμφωνα με πιθανότητα $b_k(l)$.
Μετάβαση σε μια νέα κατάσταση g σύμφωνα με τις πιθανότητες μεταβάσεων.
Παραγωγή νέου συμβόλου από την κατανομή της νέας κατάστασης
Επανάληψη των βημάτων 3 και 4 μέχρι το τέλος.

Όπως στο απλό μοντέλο Markov ορίζουμε την πιθανοφάνεια $P(X_i \mid \Theta)$ ως την πιθανότητα να παραχθεί η ακολουθία X_i από το μοντέλο Θ. Στο κρυμμένο μοντέλο όμως εκτός από την παρατήρηση του X_i υπάρχει η κρυμμένη πληροφορία των

καταστάσεων από τις οποίες περνάμε σε κάθε βήμα κατά την παραγωγή της ακολουθίας. Επομένως δεν είναι δυνατό να υπολογίσουμε απευθείας την $P(X_i \mid \Theta)$, αφού πρέπει να ληφθεί υπόψη και η ακολουθία των καταστάσεων. Για να βρούμε την πιθανότητα $P(X_i \mid \Theta)$ πρέπει να λάβουμε υπόψη όλα τα δυνατά μονοπάτια καταστάσεων για το μοντέλο μας. Η πιθανότητα να παραχθεί η X_i από το μοντέλο Θ είναι το άθροισμα των πιθανοτήτων για κάθε μονοπάτι. Δηλαδή **[11, 8]**:

$$P(X_i \mid \Theta) = \sum_Q P(X_i \mid Q, \Theta) \quad P(Q \mid \Theta)$$

.

Ας θεωρήσουμε τυχαίο μονοπάτι $Q = q_1, q_2, \ldots, q_T$. Η πιθανότητα $P(Q \mid \Theta)$ είναι η πιθανότητα να περάσουμε διαδοχικά από τις καταστάσεις του Q, δηλαδή $P(Q \mid \Theta) = \pi_{q1} * a_{q1q2} * \ldots * a_{qT-1qT}$.

Η πιθανότητα $P(X_i \mid Q, \Theta)$ είναι η πιθανότητα να παρατηρήσουμε τη συμβολοσειρά $X_i = X_{i1}, X_{i2}, \ldots, X_{iT}$, γνωρίζοντας εκ των πρότερων το μονοπάτι Q που ακολουθούμε, δηλαδή $P(X_i \mid Q, \Theta) = b_{q1}(X_{i1}) * b_{q2}(X_{i2}) * \ldots * b_{qT}(X_{iT})$.

Συνεπώς, $$P(X_i \mid \Theta) = \sum_Q \pi_{q1} \quad b_{q1}(X_{i1}) \quad a_{q1q2} \quad \ldots \quad a_{qT-1qT} \quad b_{qT}(X_T)$$. Η εξίσωση αυτή περιγράφει στην ουσία τη διαδικασία παραγωγής ακολουθιών που ορίσαμε προηγουμένως, για όλους τους δυνατούς συνδυασμούς καταστάσεων, δηλαδή την επιλογή αρχικής κατάστασης, την παραγωγή συμβόλων και τις μεταβάσεις μεταξύ καταστάσεων με στοχαστικό τρόπο.

Το πρόβλημα είναι ότι για να υπολογιστεί η πιθανοφάνεια με τον παραπάνω τρόπο, απαιτείται ο υπολογισμός για κάθε διαφορετική ακολουθία καταστάσεων Q. Κάτι τέτοιο είναι υπολογιστικά πολύ ακριβό και πρακτικά ανεφάρμοστο. Για τον υπολογισμό της πιθανοφάνειας στο κρυμμένο μοντέλο χρησιμοποιείται μια άλλη μέθοδος, γνωστή ως forward – backward διαδικασία **[11, 1]**.

Ορίζουμε τη forward μεταβλητή $\alpha_t(i) = P(X_{i1}, \ldots, X_{it}, q_t = S_i \mid \Theta)$. Δηλαδή η α εκφράζει την πιθανότητα να παρατηρήσουμε τα στοιχεία της ακολουθίας X_i έως τη

στιγμή t, και εκείνη τη στιγμή να βρεθούμε στην κατάσταση S_i του κρυμμένου μοντέλου. Ο υπολογισμός της α γίνεται ως εξής:

Υπολογίζουμε τις αρχικές πιθανότητες $\alpha_1(i) = \pi_i * b_i(X_{i1})$ για $1 \le i \le K$.

Τη στιγμή t+1: $\alpha_{t+1}(i) = [\sum_{j=1}^{K} \alpha_t(j) * a_{ji}] * b_i(X_{it+1})$.

Τελικά η πιθανοφάνεια προκύπτει ως: $P(X_i | \Theta) = \sum_{j=1}^{K} \alpha_T(j)$.

Με παρόμοια λογική ορίζεται η backward μεταβλητή, ως η πιθανότητα να παρατηρήσουμε τα στοιχεία της ακολουθίας από τη στιγμή t+1 έως το τέλος, δεδομένου ότι τη στιγμή t βρισκόμαστε στην κατάσταση Si: $\beta_t(i) = P(X_{it+1}, ..., X_{iT} | q_t = S_i, \Theta)$. Ο υπολογισμός της β γίνεται ως εξής:

Θέτουμε $\beta_T(i) = 1$.

Τη στιγμή t: $\beta_t(i) = \sum_{j=1}^{K} a_{ij} * b_j(X_{it+1}) \beta_{t+1}(j)$.

Τελικά η πιθανοφάνεια προκύπτει: $P(X_i | \Theta) = \sum_{j=1}^{K} \beta_1(j)$

Την πιθανοφάνεια μπορούμε να την υπολογίσουμε με οποιονδήποτε από τους δυο προαναφερθείς τρόπους. Τις forward και backward μεταβλητές θα τις χρησιμοποιήσουμε στη συνέχεια και για να υπολογίσουμε δυο ακόμα ποσότητες που θα χρειαστούν κατά τον υπολογισμό των παραμέτρων των μοντέλων.

Ορίζουμε τη μεταβλητή $\gamma_t(i) = P(q_t = S_i | X_i, \Theta)$, δηλαδή την πιθανότητα τη χρονική στιγμή t για την ακολουθία X_i να βρισκόμαστε στην κατάσταση S_i. Με τη βοήθεια των forward και backward μεταβλητών η γ μπορεί να γραφεί ως:

$\gamma_t(i) = \frac{\alpha_t(i) * \beta_t(i)}{P(X_i \mid \Theta)} = \frac{\alpha_t(i) * \beta_t(i)}{\sum_{k=1}^{K} \alpha_t(k) * \beta_t(k)}$. Είναι προφανές ότι ισχύει $\sum_{i=1}^{K} \gamma_t(i) = 1$.

Ορίζουμε επίσης τη μεταβλητή $\xi_t(i,j) = P(q_t = S_i, q_{t+1} = S_j \mid X_i, \Theta)$, δηλαδή την πιθανότητα τη στιγμή t να βρισκόμαστε στην κατάσταση i και τη στιγμή t+1 στην κατάσταση j για την ακολουθία X_i στο μοντέλο Θ. Χρησιμοποιώντας ξανά τις forward και backward μεταβλητές μπορούμε να υπολογίσουμε την ξ:

$$\xi_t(i,j) = \frac{\alpha_t(i) * a_{ij} * b_j(X_{it+1}) * \beta_{t+1}(j)}{P(X_i \mid \Theta)} = \frac{\alpha_t(i) * a_{ij} * b_j(X_{it+1}) * \beta_{t+1}(j)}{\sum_{i=1}^{K} \sum_{j=1}^{K} \alpha_t(i) * a_{ij} * b_j(X_{it+1}) * \beta_{t+1}(j)}$$

Με τη βοήθεια των παραπάνω μεταβλητών μπορούμε να λάβουμε μια νέα εκτίμηση των παραμέτρων του Θ με βάση τις προηγούμενες τιμές.

Το άθροισμα $\sum_{t=1}^{T} \gamma_t(i)$ υπολογίζει πόσες φορές βρισκόμαστε στην κατάσταση i κατά το πέρασμα από όλη την ακολουθία.

Παραλείποντας το τελευταίο στοιχείο της ακολουθίας, δηλαδή παίρνοντας το άθροισμα $\sum_{t=1}^{T-1} \gamma_t(i)$ υπολογίζουμε πόσες φορές υπήρξε μετάβαση από την κατάσταση i σε κάποια άλλη (αφού δεν υπάρχει μετάβαση από το τελευταίο στοιχείο).

Επίσης, το άθροισμα $\sum_{t=1}^{T-1} \xi_t(i,j)$ υπολογίζει τον αριθμό των μεταβάσεων από την κατάσταση i στην κατάσταση j για την ακολουθία X_i.

Χρησιμοποιώντας τους παραπάνω τύπους μπορούμε να πάρουμε μια εκτίμηση των παραμέτρων του μοντέλου:

Η πιθανότητα αρχικής κατάστασης π_i συμβολίζει τη συχνότητα με την οποία βρισκόμαστε στην κατάσταση i τη χρονική στιγμή t=1, άρα μπορεί να υπολογιστεί ως $\pi_i = \gamma_1(i)$.

Η πιθανότητα μετάβασης a_{ij} αναπαριστά τη συχνότητα με την οποία πραγματοποιούνται μεταβάσεις από την i στην j. Μπορεί, επομένως, να εκφραστεί ως ο αριθμός των μεταβάσεων από την κατάσταση i στην j, προς τον συνολικό αριθμό των μεταβάσεων από την i. Δηλαδή:

$$a_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$$

.

Η πιθανότητα $b_i(l)$ είναι η συχνότητα παρατήρησης του συμβόλου l από την κατάσταση i, άρα μπορεί να υπολογιστεί ως το πηλίκο του πλήθους των παρατηρήσεων του l στην κατάσταση i προς το πλήθος των εμφανίσεων της κατάστασης i. Δηλαδή:

$$b_i(l) = \frac{\sum_{t=1}^{T-1} [\gamma_t(i) \wedge (X_t == l)]}{\sum_{t=1}^{T-1} \gamma_t(i)}$$

Η μέθοδος που χρησιμοποιεί τους παραπάνω τύπους για την εκτίμηση των τιμών των παραμέτρων του κρυμμένου μοντέλου ονομάζεται μέθοδος Baum – Welch **[11, 8, 1]**. Η μέθοδος είναι επαναληπτική. Ξεκινά με κάποιες αρχικές τιμές των παραμέτρων και χρησιμοποιώντας τις παραπάνω εξισώσεις σε κάθε βήμα πετυχαίνει τελικά την τοπική μεγιστοποίηση της πιθανοφάνειας $P(X \mid \Theta)$. Στην παρούσα εργασία χρησιμοποιούμε τον αλγόριθμο ΕΜ για τον υπολογισμό των παραμέτρων που είναι ισοδύναμος με τη μέθοδο Baum – Welch.

ΚΕΦΑΛΑΙΟ 3. ΟΜΑΔΟΠΟΙΗΣΗ ΑΚΟΛΟΥΘΙΑΚΩΝ ΔΕΔΟΜΕΝΩΝ ΜΕ ΤΗ ΧΡΗΣΗ ΜΙΚΤΩΝ ΜΟΝΤΕΛΩΝ MARKOV

3.1. Μικτά Μοντέλα

Αυτό που θέλουμε να πετύχουμε είναι η εύρεση ομάδων σε ένα σύνολο ακολουθιών $X = \{X_i\}$, i=1:N. Θεωρούμε ότι τα δεδομένα χωρίζονται σε M ομάδες, αλλά δεν γνωρίζουμε σε ποια κατηγορία ανήκει κάθε X_i. Θεωρούμε ότι κάθε ομάδα ακολουθιών κατασκευάζεται από μια διαδικασία Markov, οπότε ορίζουμε ένα μοντέλο Markov για κάθε ομάδα **[5]**.

Έχουμε, επομένως, ένα σύνολο $\Theta = \{\theta_m\} = \{\pi_m, A_m\}$, m=1:M μοντέλων Markov. Θεωρούμε το μοντέλο που προκύπτει από τη μίξη των παραπάνω μοντέλων Markov. Το μικτό μοντέλο έχει συνάρτηση πιθανότητας που προκύπτει από το γραμμικό συνδυασμό των επιμέρους μοντέλων Markov $P(X_i \mid \Theta) = \sum_{m=1}^{M} C_m * P(X_i \mid \theta_m)$ **[2, 13, 1]**.

Οι πιθανότητες C_m είναι οι παράμετροι της μίξης των μοντέλων και συμβολίζουν την εκ των προτέρων πιθανότητα ένα X_i να ανήκει σε κάποια κατηγορία m, δηλαδή $C_m = P(c_i = m \mid \Theta)$. Η εκ των προτέρων πιθανότητα (prior) αναπαριστά τυχόν γνώση που μπορεί να έχουμε για τις κατηγορίες. Έστω, για παράδειγμα, ότι γνωρίζουμε ότι στην κατηγορία 1 αντιστοιχεί διπλάσιος αριθμός ακολουθιών X_i απ' ό,τι στην κατηγορία 2 χωρίς να γνωρίζουμε τα X_i. Μπορούμε να πούμε εκ των προτέρων (χωρίς να λάβουμε τίποτα άλλο υπόψη) ότι η πιθανότητα ένα X_i να ανήκει στην κατηγορία 1 είναι διπλάσια από το να ανήκει στην 2. Προφανώς ισχύουν οι περιορισμοί $C_m \geq 0$ και $\sum_{m=1}^{M} C_m = 1$. Το σύνολο των παραμέτρων για το μικτό μοντέλο είναι οι παράμετροι μίξης συν τις παραμέτρους των επιμέρους μοντέλων Markov: $\Theta = [C_m, P^m, A^m]_{m=1:M}$.

Ο τρόπος λειτουργίας του μικτού μοντέλου για την παραγωγή δεδομένων είναι ο εξής:

1. Επιλέγεται μια συνιστώσα του μικτού μοντέλου με βάση τις εκ των προτέρων πιθανότητες C_m.
2. Κατασκευάζουμε το δεδομένο από τη στοχαστική μαρκοβιανή διαδικασία (επιλογή αρχικής κατάστασης – μεταβάσεις μεταξύ καταστάσεων) χρησιμοποιώντας τις παραμέτρους του μοντέλου Markov που επιλέχθηκε.

Για το σύνολο των δεδομένων, το μικτό μοντέλο έχει συνάρτηση πιθανότητας την πιθανοφάνεια του συνόλου των δεδομένων X: $P(X \mid \Theta) = \prod_{i=1}^{N} P(X_i \mid \Theta)$, θεωρώντας ότι υπάρχει ανεξαρτησία μεταξύ των παρατηρήσεων.

3.2. Ομαδοποίηση με Μικτά Μοντέλα

Αυτό που επιδιώκουμε είναι να καθορίσουμε τις παραμέτρους του μικτού μοντέλου, ώστε η πιθανοφάνεια να μεγιστοποιείται. Δηλαδή να βρούμε Θ^*, ώστε $\Theta^* = \text{argmax}_\Theta P(X \mid \Theta)$. Η μεγιστοποίηση της πιθανοφάνειας σημαίνει ότι το μικτό μοντέλο

προσεγγίζει περισσότερο το μηχανισμό παραγωγής των δεδομένων, αντιπροσωπεύει καλύτερα τις αντίστοιχες ομάδες και επομένως επιτυγχάνει καλύτερη κατηγοριοποίηση που είναι ο πραγματικός στόχος.

Έστω ότι οι παράμετροι Θ του μικτού μοντέλου είναι γνωστοί. Υπολογίζοντας την πιθανοφάνεια για το μικτό μοντέλο, μπορούμε να υπολογίσουμε την εκ των υστέρων πιθανότητα $P(m \mid X_i)$ με τη χρήση του κανόνα του Bayes **[5]**:

$$P(m \mid X_i) = \frac{C_m * P(X_i \mid \theta^m)}{P(X_i \mid \Theta)}$$

Το άθροισμα της εκ των υστέρων πιθανότητας είναι 1 για όλες τις συνιστώσες, δηλαδή $\sum_{m=1}^{M} P(m \mid X_i) = 1$.

Η εκ των υστέρων πιθανότητα εκφράζει την πιθανότητα η συγκεκριμένη συνιστώσα m να ήταν υπεύθυνη για την παραγωγή της συγκεκριμένης παρατήρησης X_i. Επομένως, παρατηρώντας τις εκ των υστέρων πιθανότητες μπορούμε να υπολογίσουμε σε ποιο μοντέλο m είναι περισσότερο πιθανό να ανήκει κάθε δεδομένο X_i:

$Xi \in m^* = \text{argmax}_m\{P(m \mid X_i)\}, m=1, \ldots, M$

Το πρόβλημα επομένως είναι να υπολογίσουμε τις βέλτιστες τιμές των παραμέτρων, δηλαδή αυτές, για τις οποίες θα επιτυγχάνεται βέλτιστα η ομαδοποίηση. Το μέτρο για να το πετύχουμε αυτό είναι η πιθανοφάνεια, η οποία εκφράζει το βαθμό στον οποίο το μοντέλο προσεγγίζει την πραγματική κατανομή των δεδομένων. Επομένως, το πρόβλημα βελτιστοποίησης των παραμέτρων μπορεί να αναχθεί σε πρόβλημα μεγιστοποίησης της πιθανοφάνειας του μοντέλου.

Οι βέλτιστες παράμετροι είναι αυτές που μεγιστοποιούν την πιθανοφάνεια του μοντέλου στα δεδομένα:

$\Theta^* : \text{argmax}_\Theta \{ \log(P(X \mid \Theta)) \}$

Για να βρούμε τις τιμές των παραμέτρων του μικτού μοντέλου που μεγιστοποιούν την πιθανοφάνεια χρησιμοποιούμε τον αλγόριθμο EM (Expectation – Maximization) για απλά και κρυμμένα μοντέλα Markov.

3.3. Αλγόριθμος EM

Ο αλγόριθμος EM είναι ιδιαίτερα χρήσιμος στην επίλυση εύρεση λύσεων μέγιστης πιθανοφάνειας όταν δεν είναι δυνατό να υπολογίσουμε αναλυτικά τις τιμές των παραμέτρων ή όταν υπάρχει κάποια κρυμμένη πληροφορία. Στα μικτά μοντέλα που χρησιμοποιούμε στο πρόβλημά μας θα δούμε ότι υπάρχει κρυμμένη πληροφορία που δυσκολεύει την επίλυση του προβλήματος. Γι' αυτό το λόγο καταφεύγουμε στον αλγόριθμο EM **[6, 2]**.

Έστω ότι έχουμε την παρατήρηση X και υπάρχει κρυμμένη πληροφορία Z στο πρόβλημα. Το Y=(X, Z) είναι το πλήρες σύνολο δεδομένων του προβλήματος και αυτό πρέπει να λάβουμε υπόψη στην επίλυση του προβλήματος. Ορίζουμε επομένως την πλήρη πιθανοφάνεια $P(Z \mid \Theta) = P(X, Z \mid \Theta) = P(Z \mid X, \Theta) * P(X \mid \Theta)$. Ο αλγόριθμος EM μεγιστοποιεί την πλήρη πιθανοφάνεια ως προς τις παραμέτρους Θ **[1]**. Από την προηγούμενη εξίσωση είναι φανερό ότι αυτό γίνεται με την μεγιστοποίηση της $P(X \mid \Theta)$, ενώ πρέπει να λάβουμε υπόψη τον παράγοντα $P(Z \mid X, \Theta)$. Η $P(Z \mid X, \Theta)$ είναι η κατανομή των κρυμμένων μεταβλητών, η οποία είναι άγνωστη. Ο EM λειτουργεί επαναληπτικά σε δυο βήματα **[6, 1, 2]**:

Expectation Step: Εκτίμηση των κατανομών των κρυμμένων μεταβλητών $P(Z \mid X, \Theta)$ δοθέντων των παρατηρήσεων και των παραμέτρων.
Maximization Step: Μεγιστοποίηση, ως προς τις παραμέτρους Θ, της αναμενόμενης τιμής της πλήρους πιθανοφάνειας:
$Q(\Theta^t, \Theta^{t-1}) = E_{P(Z \mid X, \Theta)} [P(X, Z \mid \Theta)]$

Τα παραπάνω βήματα επαναλαμβάνονται μέχρι να θεωρήσουμε ότι έχει επιτευχθεί η μεγιστοποίηση της πιθανοφάνειας. Αποδεικνύεται ότι σε κάθε βήμα του ΕΜ η πιθανοφάνεια $P(X \mid \Theta)$ **δεν μειώνεται**, δηλαδή $P(X \mid \Theta^{t+1}) \geq P(X \mid \Theta^{t})$. Άρα σε κάθε βήμα οι τιμές των παραμέτρων που υπολογίζουμε μας δίνουν λύση μεγαλύτερης πιθανοφάνειας.

Ο αλγόριθμος ΕΜ είναι σχετικά εύκολος στην υλοποίηση και μπορούμε να τον χρησιμοποιήσουμε για τη βελτιστοποίηση της πιθανοφάνειας ικανοποιώντας τους περιορισμούς που έχουμε θέσει στις παραμέτρους C, π, a.

Το μειονέκτημα του ΕΜ είναι ότι χρειάζονται κάποιες αρχικές τιμές των παραμέτρων Θ, προκειμένου να πραγματοποιηθεί η πρώτη εκτίμηση και να ξεκινήσει η διαδικασία. Γενικά δεν είναι εύκολος ούτε και προφανής ο τρόπος επιλογής των αρχικών τιμών των παραμέτρων. Το αποτέλεσμα της εκπαίδευσης και ο αριθμός των επαναλήψεων που απαιτούνται για να επιτευχθεί η σύγκλιση εξαρτώνται άμεσα από τις αρχικές τιμές των παραμέτρων. Η μεγιστοποίηση της πιθανοφάνειας είναι τοπική που σημαίνει ότι μπορεί πάντα να υπάρχει κάποια καλύτερη λύση από αυτή που βρίσκουμε.

3.4. ΕΜ για Απλά Μοντέλα Markov

Τα δεδομένα που έχουμε είναι οι παρατηρήσεις $X = \{X_i\}$, i=1:N, με κάθε $X_i = [X_{it}]$, t=1:T(i). Από τη στιγμή που ορίζουμε ότι κάθε ακολουθία ανήκει σε μια από M κατηγορίες, υπάρχει κρυμμένη πληροφορία στο πρόβλημά μας, η κατηγορία στην οποία ανήκει κάθε ακολουθία. Εκφράζουμε αυτή την πληροφορία με ένα διάνυσμα Z_i = [0, …, 1, …, 0] μήκους M. Το Z_i έχει 1 μόνο στη θέση όπου αντιστοιχεί στην κατηγορία που ανήκει και 0 στις υπόλοιπες θέσεις. Εάν γνωρίζαμε εξαρχής την κατηγορία κάθε παρατήρησης η εύρεση των κατάλληλων τιμών των παραμέτρων θα ήταν προφανής όσο και απλή. Θα μπορούσαμε να βρούμε τις πιθανότητες έναρξης και μετάβασης με στατιστικό τρόπο για τα δεδομένα της κάθε κατηγορίας. Όμως η Z_i

είναι κρυμμένη και πρέπει να τη λάβουμε υπόψη, προκειμένου να λύσουμε το πρόβλημα.

Η πιθανοφάνεια που έχουμε ορίσει προηγουμένως $P(X_i \mid \Theta)$ δε λαμβάνει υπόψη την κρυμμένη πληροφορία Z_i. Η πιθανοφάνεια μπορεί να υπολογιστεί όμως με τη βοήθεια της Z_i ως: $P(X_i \mid \Theta) = \sum_{Zi} P(X_i, Z_i \mid \Theta)$. Η ποσότητα $P(X_i, Z_i \mid \Theta)$ αποτελεί την πλήρη πιθανοφάνεια για τις παρατηρήσεις και την κρυμμένη πληροφορία και υπολογίζεται: $P(X_i, Z_i \mid \Theta) = P(X_i \mid Z_i, \Theta)*P(Z_i \mid \Theta)$. Η $P(X_i \mid Z_i, \Theta)$ εκφράζει την πιθανότητα της X_i αν γνωρίζουμε την κατηγορία m στην οποία ανήκει. Άρα $P(X_i \mid Z_i, \Theta) = P(X_i \mid \theta_m) * P(Z_i = m) = C_m * P(Xi \mid \theta_m)$. Το Z_i όμως είναι δυαδικό διάνυσμα με 1 μόνο στη θέση που δείχνει την κατηγορία του X_i. Επομένως η πιο πάνω εξίσωση μπορεί να γραφεί ως: $P(X_i, Z_i \mid \Theta) = \prod_{m=1}^{M} (C_m \bullet P(X_i \mid \theta_m))^{Zim}$.

Η πλήρης πιθανοφάνεια για το σύνολο των παρατηρήσεων $P(X, Z \mid \Theta) = \prod_{i=1}^{N} \prod_{m=1}^{M} (C_m \bullet P(X_i \mid \theta_m))^{Zim}$ είναι η ποσότητα που πρέπει να μεγιστοποιήσουμε προκειμένου να βελτιστοποιήσουμε τις παραμέτρους του μικτού μοντέλου. Η μεγιστοποίηση της πλήρους πιθανοφάνειας για ένα X_i και Z_i σημαίνει ότι η πιθανότητα εμφάνισης μιας ακολουθίας X_i να ανήκει στην πραγματική κατηγορία της που δείχνει η Z_i μεγιστοποιείται για το μικτό μοντέλο Θ που θεωρούμε.

Στη βελτιστοποίηση χρησιμοποιούμε τη λογαριθμική πιθανοφάνεια

$$Q^L = \log P(X,Z \mid \Theta) = \log \prod_{i=1}^{N} \prod_{m=1}^{M} (C_m \bullet P(X_i \mid \theta_m))^{Zim} = \sum_{i=1}^{N} \sum_{m=1}^{M} Z_{im} \bullet \left[\log C_m + \log P(X_i \mid \theta_m) \right]$$

(Εξίσωση 3.1)

Επομένως, από την εξίσωση 2.3 έχουμε υπολογίσει την πιθανοφάνεια: $P(X_i \mid \theta_m) = \prod_{k=1}^{K} \pi_k^{m\,\delta(Xi1, Sk)} \prod_{t=1}^{Ti-1} \prod_{f=1}^{K} \prod_{g=1}^{K} a_{fg}^{m\,\delta(Xit, Sf, Xit+1, Sg)}$ και η εξίσωση 3.1 γίνεται με τη μετατροπή των γινομένων σε αθροίσματα εξαιτίας του λογαρίθμου: $Q^L = \sum_{i=1}^{N} \sum_{m=1}^{M} Z_{im} \bullet \{ \log C_m + \sum_{k=1}^{K} \delta(X_{i1}, S_k) * \log \pi_k^m + \sum_{t=1}^{Ti-1} \sum_{f=1}^{K} \sum_{g=1}^{K} \delta(X_{it}, S_f, X_{it+1}, S_g) * \log a_{fg}^m \}$.

Η συνάρτηση Q^L είναι η συνάρτηση της λογαριθμικής πιθανοφάνειας για το απλό μοντέλο Markov. Μεγιστοποιώντας την Q^L μεγιστοποιούμε ταυτόχρονα την πιθανοφάνεια του μικτού μοντέλου. Η μεγιστοποίηση γίνεται ως προς τις παραμέτρους του μικτού μοντέλου $\theta_m=\{ C_m, \pi_k^m, a_{fg}^m \}$. Βρίσκουμε δηλαδή τις τιμές των παραμέτρων που μεγιστοποιούν την Q^L χρησιμοποιώντας τον αλγόριθμο EM. Έχουμε αναφέρει τα δυο βήματα του EM:

E – Step: Γίνεται εκτίμηση της κρυμμένης πληροφορίας Z_i με βάση τις τρέχουσες τιμές των παραμέτρων.
M – Step: Στηριζόμενοι στην παραπάνω εκτίμηση υπολογίζουμε τις νέες τιμές των παραμέτρων που μεγιστοποιούν την συνάρτηση Q^L.

Ο αλγόριθμος ξεκινά με κάποιες αρχικές τιμές των παραμέτρων και λειτουργεί με επανάληψη των δυο παραπάνω βημάτων μέχρι να θεωρήσουμε ότι έχει επιτευχθεί η μεγιστοποίηση της πιθανοφάνειας.

3.4.1. E – Step

Σε αυτό το βήμα γίνεται υπολογισμός της αναμενόμενης τιμής του Z_{im}, δηλαδή βρίσκουμε το E[Zim] για τη χρονική στιγμή t που υπολογίζεται ως:

$$E[Z_{im}]^t = P(m \mid X_i, \Theta^{t-1}) = \frac{P(m)^{t-1} * P(X_i \mid m, \Theta^{t-1})}{P(X_i \mid \Theta^{t-1})} \Rightarrow E[Z_{im}]^t = \frac{C_m^{t-1} * P(X_i \mid \theta_m^{t-1})}{\sum_{m=}^{M} C_m^{t-1} * P(X_i \mid \theta_m^{t-1})}.$$

3.4.2. M- Step

Σε αυτό το βήμα γίνεται μεγιστοποίηση της τιμής της συνάρτησης Q^L, ως προς τις παραμέτρους του μικτού μοντέλου C_m, π_k^m, a_{fg}^m. Η μεγιστοποίηση γίνεται βρίσκοντας τις μερικές παραγώγους και εξισώνοντας με το μηδέν. Κατά την εύρεση των τιμών που μηδενίζουν τις παραγώγους δεν πρέπει να ξεχνάμε τους περιορισμούς που ισχύουν για τις παραμέτρους του μοντέλου. Συνεπώς, πρέπει να λύσουμε τις παρακάτω εξισώσεις με τους ακόλουθους περιορισμούς για τα C_m, π_k^m, a_{fg}^m:

$$\sum_{m=1}^{M} C_m = 1 \quad , \sum_{k=1}^{K} \pi_k^m = 1 \ \forall m \quad , \sum_{j=1}^{K} a_{ij}^m = 1 \ \forall i, m \quad .$$

Για να λάβουμε υπόψη τους περιορισμούς εισάγουμε τη χρήση των πολλαπλασιαστών Lagrance. Έτσι, οι εξισώσεις μηδενισμού των παραγώγων θα είναι:

$$\frac{\partial}{\partial C_m}\{ Q^L - \lambda (\sum_{m=1}^{M} C_m - 1) \} = 0 \quad \textbf{(1)}$$

$$\frac{\partial}{\partial \pi_k^m}\{ Q^L - \sum_{m=1}^{M} \lambda_m (\sum_{k=1}^{K} \pi_k^m - 1) \} = 0 \quad \textbf{(2)}$$

$$\frac{\partial}{\partial a_{fg}^m}\{ Q^L - \sum_{m=1}^{M} \sum_{f=1}^{K} \lambda_{mf} (\sum_{g=1}^{K} a_{fg}^m - 1) \} = 0 \quad \textbf{(3)}$$

Η λύση των (1), (2), (3) δίνει τις εξισώσεις υπολογισμού των τιμών των παραμέτρων. Λύνοντας την (1) παίρνουμε:

$$\frac{\partial}{\partial C_m}\{ Q^L - \lambda(\sum_{m=1}^{M} C_m - 1) \} = 0 \Rightarrow \sum_{i=1}^{N} Z_{im} \frac{1}{C_m} - \lambda = 0 \Rightarrow \lambda C_m = \sum_{i=1}^{N} Z_{im} \Rightarrow \sum_{m=1}^{M} \lambda C_m = \sum_{i=1}^{N} \sum_{m=1}^{M} Z_{im} \Rightarrow$$

$\lambda = N$, αφού $\sum_{m=1}^{M} C_m = 1 \, \kappa\alpha\iota \sum_{m=1}^{M} Z_{im} = 1$.

Επομένως:

$$\sum_{i=1}^{N} Z_{im} \frac{1}{C_m} - N = 0 \Rightarrow \boxed{C_m = \frac{\sum_{i=1}^{N} Z_{im}}{N}}$$

Πρέπει να διευκρινιστεί, ότι η Z_{im} είναι η εκτίμηση $E[Z_{im}]$ που έχουμε υπολογίσει στο E – Step που προηγήθηκε και απλά για λόγους απλότητας τη συμβολίζουμε ως Z_{im}.

Με παρόμοιο τρόπο λύνουμε τις εξισώσεις για τις υπόλοιπες παραμέτρους:

(2) για το π:

$$\frac{\partial}{\partial \pi_k^m} \{ Q^L - \sum_{m=1}^{M} \lambda_m (\sum_{k=1}^{K} \pi_k^m - 1) \} = 0 \Rightarrow \sum_{i=1}^{N} (Z_{im} * \delta(X_{i1}, S_k) * \frac{1}{\pi_k^m}) - \lambda_m = 0 \Rightarrow$$

$$\lambda_m \pi_k^m = \sum_{i=1}^{N} Z_{im} * \delta(X_{i1}, S_k) \Rightarrow \lambda_m \sum_{k=1}^{K} \pi_k^m = \sum_{i=1}^{N} Z_{im} * \sum_{k=1}^{K} \delta(X_{i1}, S_k) \Rightarrow \lambda_m = \sum_{i=1}^{N} Z_{im}$$

Οπότε: $$\sum_{i=1}^{N} Z_{im} * \delta(X_{i1}, S_k) = \pi_k^m * \sum_{i=1}^{N} Z_{im} \Rightarrow \pi_k^m = \frac{\sum_{i=1}^{N} Z_{im} * \delta(X_{i1}, S_k) + \beta_k}{\sum_{i'=1}^{N} Z_{i'm'} + \sum_{k'=1}^{K} \beta_{k'}}$$

(3) για το a:

$$\frac{\partial}{\partial a_{fg}^m} \{ Q^L - \sum_{m=1}^{M} \sum_{f=1}^{K} \lambda_{mf} (\sum_{g=1}^{K} a_{fg}^m - 1) \} = 0 \Rightarrow \sum_{i=1}^{N} Z_{im} * \sum_{t=1}^{Ti-1} \delta(X_{it}, S_f, X_{it+1}, S_g) * \frac{1}{a_{fg}^m} - \lambda_{mf} = 0 \Rightarrow$$

$$\lambda_{mf} * a_{fg}^m = \sum_{i=1}^{N} Z_{im} \sum_{t=1}^{Ti-1} \delta(X_{it}, S_f, X_{it+1}, S_g) \Rightarrow \lambda_{mf} * \sum_{g=1}^{K} a_{fg}^m = \sum_{i=1}^{N} Z_{im} * \sum_{t=1}^{Ti-1} \sum_{g=1}^{K} \delta(X_{it}, S_f, X_{it+1}, S_g) \Rightarrow$$

$$\lambda_{mf} = \sum_{i=1}^{N} Z_{im} * \sum_{t=1}^{Ti-1} \sum_{g=1}^{K} \delta(X_{it}, S_f, X_{it+1}, S_g)$$

Οπότε: $$\sum_{i=1}^{N} Z_{im} * \sum_{t=1}^{Ti-1} \delta(X_{it}, S_f, X_{it+1}, S_g) * \frac{1}{a_{fg}^m} = \sum_{i=1}^{N} Z_{im} * \sum_{t=1}^{Ti-1} \sum_{g=1}^{K} \delta(X_{it}, S_f, X_{it+1}, S_g) \Rightarrow$$

$$a_{fg}^{m} = \frac{\sum_{i=1}^{N} Z_{im} * \sum_{t=1}^{Ti-1} \delta(X_{it}, S_f, X_{it+1}, S_g) + \beta_{fg}}{\sum_{i'=1}^{N} Z_{i'm} * \sum_{t'=1}^{Ti-1} \sum_{g'=1}^{K} \delta(X_{i't'}, S_f, X_{i't'+1}, S_{g'}) + \sum_{g'=1}^{K} \beta_{fg'}}$$

Οι όροι β που προσθέτουμε στις εξισώσεις υπολογισμού των π και a είναι ψευδοτυχαίοι αριθμοί και χρησιμοποιούνται προκειμένου να αποφύγουμε αριθμητικά σφάλματα. Οι τιμές των β είναι ένα μικρό ποσοστό (πχ. 0,01) των τιμών των παραμέτρων π^1 και A^1 το ολικού μοντέλου που μεριγράφει το πρόβλημα. Ουσιαστικά τους χρησιμοποιούμε για να αποφύγουμε να λάβει κάποια παράμετρος τιμή 0, οπότε και υπάρχει κίνδυνος μηδενισμού της πιθανοφάνειας και συνεπώς πρόβλημα στον υπολογισμού του λογαρίθμου.

Τελειώνοντας το M – Step έχουμε υπολογίσει τις νέες τιμές των παραμέτρων C_m, π_k^m, a_{fg}^m με βάση τις τιμές της προηγούμενης επανάληψης. Αυτή η διαδικασία συνεχίζεται μέχρι να θεωρήσουμε ότι η πιθανοφάνεια του μικτού μοντέλου έχει μεγιστοποιηθεί. Αυτό το κάνουμε υπολογίζοντας την πιθανοφάνεια $P(X \mid \Theta)$ μετά από κάθε M – Step και συγκρίνοντας την τιμή της με αυτή στο αμέσως προηγούμενο βήμα. Εάν η διαφορά των δυο τιμών είναι αρκούντως μικρή (πχ. < 10^{-2} ή 10^{-3}) θεωρούμε ότι η πιθανοφάνεια έχει συγκλίνει ικανοποιητικά στο τοπικό μέγιστο που ψάχνουμε και ο αλγόριθμος τερματίζει. Για πρακτικούς λόγους μπορούμε να ορίσουμε ένα μέγιστο πλήθος επαναλήψεων, οπότε να τερματίζει ο αλγόριθμος, ακόμα και εάν δεν έχει συγκλίνει.

3.5. ΕΜ για Κρυμμένα Μοντέλα Markov

Στα κρυμμένα μικτά μοντέλα, εκτός από την κρυμμένη πληροφορία για το μοντέλο που ανήκει κάθε X_i, υπάρχει και κρυμμένη πληροφορία για την κατάσταση από την οποία προκύπτει κάθε στοιχείο X_{it}. Δηλαδή για κάθε ακολουθία παρατηρήσεων $X_i = X_{i1}, \ldots, X_{iT}$ υπάρχει η αντίστοιχη ακολουθία καταστάσεων $Q_i=q_{i1}, \ldots, q_{iT}$. Θεωρούμε επομένως, ότι εκτός από την Z_i υπάρχει και η κρυμμένη μεταβλητή I με

$$I_{it}^{m}(k) = \begin{cases} 1, & \alpha v \quad qt=S_k^m \\ 0, & \delta\iota\alpha\varphi o\rho\varepsilon\tau\iota\kappa\acute{\alpha} \end{cases}.$$

Η πλήρης πιθανοφάνεια για το μικτό μοντέλο θα περιλαμβάνει, εκτός της X και την πληροφορία για τα Z,I και θα είναι:

$$P(X_i, Z_i, I_i \mid \Theta) = \prod_{m=1}^{M} [C_m * P(X_i, I_i^m \mid \Theta)]^{Zim} .$$

Η $P(X_i, I_i^m \mid \theta_m) = \sum_{Q=\{q_{m1}, q_{m2}, \ldots, q_{mT}\}} P(X_i, Q \mid \theta_m)$ υπολογίζει την πιθανότητα της X_i για όλα τα δυνατά μονοπάτια καταστάσεων Q του μοντέλου m και μπορεί να γραφεί ως:

$$P(X_i, I_i^m \mid \theta_m) = \sum_{Q} \pi_{q1}^m * b_{q1}^m(X_{i1}) * a_{q1q2}^m * \ldots * a_{qT-1qT}^m * b_{qT}^m(X_{iT}) =$$

$$= \sum_{Q} \pi_{q1}^m * \prod_{t=1}^{T-1} a_{qtqt+1}^m * \prod_{t=1}^{T} b_{qt}^m(X_{it}) .$$

Όπως και στο απλό μοντέλο εκφράζουμε τους παράγοντες $\pi_{q1}^m, a_{qtqt+1}^m, b_{qt}^m(X_{it})$ με τη μορφή γινομένων:

$$\pi_{q1}^{m} = \prod_{k=1}^{Km} \pi_{k}^{m\, I_{i1}^{m}(k)}$$

$$a_{qtqt+1}^{m} = \prod_{f=1}^{Km} \prod_{g=1}^{Km} a_{fg}^{m\, I_{it}^{m}(f) * I_{it+1}^{m}(g)}$$

$$b_{qt}^{m}(X_{it}) = \prod_{k=1}^{Km} \prod_{l=1}^{L} b_{k}^{m}(l)^{I_{it}^{m}(k) * \delta(X_{it}, V_l)}$$

Επομένως:

$$P(X_i, I_i^m \mid \theta_m) = \prod_{k=1}^{Km} \pi_{k}^{m\, I_{i1}^{m}(k)} * \prod_{f=1}^{Km} \prod_{g=1}^{Km} a_{fg}^{m\, I_{it}^{m}(f) * I_{it+1}^{m}(g)} * \prod_{k=1}^{Km} \prod_{l=1}^{L} b_{k}^{m}(l)^{I_{it}^{m}(k) * \delta(X_{it}, V_l)}$$

Η συνάρτηση Q που θα χρησιμοποιήσουμε στον αλγόριθμο ΕΜ θα είναι:

$Q = \prod_{i=1}^{N} P(X_i, I_i \mid \Theta)$. Για τη λογαριθμική Q, όπως στο απλό μοντέλο θα έχουμε:

$$Q^L = \log(Q) = \sum_{i=1}^{N} \sum_{m=1}^{M} Z_{im} * [\, \log C_m + \log P(X_i, I_i^m \mid \Theta^m)\,] =$$

$$Q^L = \log(Q) = \sum_{i=1}^{N} \sum_{m=1}^{M} Z_{im} * [\, \log C_m + \sum_{k=1}^{Km} I_{i1}^{m}(k) * \log \pi_{k}^{m} + \sum_{t=1}^{Ti-1} \sum_{f=1}^{Km} \sum_{g=1}^{Km} I_{it}^{m}(f) * I_{it+1}^{m}(g) * \log a_{fg}^{m} +$$

$$+ \sum_{t=1}^{Ti} \sum_{k=1}^{Km} \sum_{l=1}^{L} I_{it}^{m}(k) * \delta(X_{it}, V_l) * \log b_{k}^{m}(l)\,]$$

Χρησιμοποιούμε ξανά τον αλγόριθμο ΕΜ για να μεγιστοποιήσουμε τη συνάρτηση Q^L.

3.5.1. E – Step

Στο E – Step γίνεται εκτίμηση των εξής ποσοτήτων, χρησιμοποιώντας τις τρέχουσες τιμές των παραμέτρων:

Της Z_{im} όπως και στο απλό μικτό μοντέλο:

$$E[Z_{im}] = P(m \mid X_i, \Theta) = \frac{P(m) * P(X_i \mid m, \Theta)}{P(X_i \mid \Theta)} \Rightarrow E[Z_{im}] = \frac{C_m * P(X_i, I_i^m \mid \theta_m)}{\sum_{m=1}^{M} C_m * P(X_i, I_i^m \mid \theta_m)}$$

Της πιθανότητας να βρισκόμαστε σε μια δεδομένη κατάσταση μια χρονική στιγμή, δηλαδή:

$$E[I_{it}^m(k)] = P(I_{it}^m(k) = 1 \mid X_i, \theta_m) = P(q_t = S_k^m \mid X_i, \theta_m) = \gamma_{it}^m(k) = \frac{\alpha_{it}^m(k) * \beta_{it}^m(k)}{\sum_{k=1}^{Km} \alpha_{it}^m(k) * \beta_{it}^m(k)}$$

Της πιθανότητας να βρισκόμαστε σε μια δεδομένη κατάσταση f τη στιγμή t και σε μια άλλη g τη στιγμή t+1:

$$E[I_{it}^m(f,g)] = E[I_{it}^m(f) * I_{it+1}^m(g)] = P(I_{it}^m(f) = 1, I_{it+1}^m(g) = 1 \mid X_i, \theta_m) = P(q_{it} = S_f^m, q_{it+1} = S_g^m \mid X_i, \theta_m) =$$

$$= \xi_{it}^m(f,g) = \frac{\alpha_{it}^m(f) * a_{fg}^m * b_g^m(X_{it+1}) * \beta_{it}^m(g)}{\sum_{f=1}^{Km} \sum_{g=1}^{Km} \alpha_{it}^m(f) * a_{fg}^m * b_g^m(X_{it+1}) * \beta_{it}^m(g)}$$

3.5.2. M – Step

Σε αυτό το βήμα γίνεται μεγιστοποίηση της τιμής της συνάρτησης Q^L, ως προς τις παραμέτρους του μικτού μοντέλου $C_m, \pi_k^m, a_{fg}^m, b_g^m(l)$. Η μεγιστοποίηση γίνεται βρίσκοντας τις μερικές παραγώγους και εξισώνοντας με το μηδέν. Κατά την εύρεση των τιμών που μηδενίζουν τις παραγώγους δεν πρέπει να ξεχνάμε τους περιορισμούς

που ισχύουν για τις παραμέτρους του μοντέλου. Συνεπώς, πρέπει να λύσουμε τις παρακάτω εξισώσεις με τους ακόλουθους περιορισμούς για τα $C_m, \pi_k^m, a_{fg}^m, b_g^m(l)$:

$$\sum_{m=1}^{M} C_m = 1 \;,\; \sum_{k=1}^{K} \pi_k^m = 1 \;\forall m \;,\; \sum_{j=1}^{K} a_{ij}^m = 1 \;\forall i,m \;,\; \sum_{l=1}^{L} b_k^m(l) = 1 \;\forall m,k \;.$$

Για να λάβουμε υπόψη τους περιορισμούς εισάγουμε τη χρήση των πολλαπλασιαστών Lagrance. Έτσι, οι εξισώσεις μηδενισμού των παραγώγων θα είναι:

$$\frac{\partial}{\partial C_m}\{ Q^L - \lambda (\sum_{m=1}^{M} C_m - 1) \} = 0 \quad \textbf{(1)}$$

$$\frac{\partial}{\partial \pi_k^m}\{ Q^L - \sum_{m=1}^{M} \lambda_m (\sum_{k=1}^{K} \pi_k^m - 1) \} = 0 \quad \textbf{(2)}$$

$$\frac{\partial}{\partial a_{fg}^m}\{ Q^L - \sum_{m=1}^{M} \sum_{f=1}^{K} \lambda_{mf} (\sum_{g=1}^{K} a_{fg}^m - 1) \} = 0 \quad \textbf{(3)}$$

$$\frac{\partial}{\partial b_k^m(l)}\{ Q^L - \sum_{m=1}^{M} \sum_{f=1}^{K} \lambda_{mk} (\sum_{k=1}^{L} b_k^m(l) - 1) \} = 0 \quad \textbf{(4)}$$

Η λύση των (1), (2), (3) δίνει τις εξισώσεις υπολογισμού των τιμών των παραμέτρων. Λύνοντας την (1) παίρνουμε:

$$\frac{\partial}{\partial C_m}\{ Q^L - \lambda(\sum_{m=1}^{M} C_m - 1) \} = 0 \Rightarrow \sum_{i=1}^{N} Z_{im} \frac{1}{C_m} - \lambda = 0 \Rightarrow \lambda C_m = \sum_{i=1}^{N} Z_{im} \Rightarrow \sum_{m=1}^{M} \lambda C_m = \sum_{i=1}^{N} \sum_{m=1}^{M} Z_{im} \Rightarrow$$

$\lambda = N$, αφού $\sum_{m=1}^{M} C_m = 1 \,\kappa\alpha\iota\, \sum_{m=1}^{M} Z_{im} = 1$.

Επομένως:

$$\sum_{i=1}^{N} Z_{im} \frac{1}{C_m} - N = 0 \Rightarrow \boxed{C_m = \frac{\sum_{i=1}^{N} Z_{im}}{N}}$$

Σε αντιστοιχία με ό,τι κάναμε στο απλό μοντέλο, οι Z_{im}, $I_{it}^{m}(k)$ και $I_{it}^{m}(f,g)$ είναι οι εκτιμήσεις που έχουμε υπολογίσει στο E – Step που προηγήθηκε.

Με παρόμοιο τρόπο λύνουμε τις εξισώσεις για τις υπόλοιπες παραμέτρους:

$$Q^{L} = \log(Q) = \sum_{i=1}^{N}\sum_{m=1}^{M} Z_{im} * [\log C_m + \sum_{k=1}^{Km} I_{i1}^{m}(k) * \log \pi_k^m + \sum_{t=1}^{Ti-1}\sum_{f=1}^{Km}\sum_{g=1}^{Km} I_{it}^{m}(f) * I_{it+1}^{m}(g) * \log a_{fg}^{m} +$$

$$+ \sum_{t=1}^{Ti}\sum_{k=1}^{Km}\sum_{l=1}^{L} I_{it}^{m}(k) * \delta(X_{it}, V_l) * \log b_k^m(l)]$$

(2) για το π:

$$\frac{\partial}{\partial \pi_k^m}\{ Q^{L} - \sum_{m=1}^{M}\lambda_m (\sum_{k=1}^{K}\pi_k^m - 1) \} = 0 \Rightarrow \sum_{i=1}^{N}(Z_{im} * I_{i1}^{m}(k) * \frac{1}{\pi_k^m}) - \lambda_m = 0 \Rightarrow$$

$$\lambda_m \pi_k^m = \sum_{i=1}^{N} Z_{im} * I_{i1}^{m}(k) \Rightarrow \lambda_m \sum_{k=1}^{K}\pi_k^m = \sum_{i=1}^{N} Z_{im} * \sum_{k=1}^{K} I_{i1}^{m}(k) \Rightarrow \lambda_m = \sum_{i=1}^{N} Z_{im}$$

Οπότε: $$\sum_{i=1}^{N} Z_{im} * I_{i1}^{m}(k) = \pi_k^m * \sum_{i=1}^{N} Z_{im} \Rightarrow \pi_k^m = \frac{\sum_{i=1}^{N} Z_{im}] * I_{i1}^{m}(k) + \beta_k}{\sum_{i=1}^{N} Z_{im} + \sum_{k=1}^{K}\beta_k}$$

(3) για το a:

$$\frac{\partial}{\partial a_{fg}^{m}}\{ Q^{L} - \sum_{m=1}^{M}\sum_{f=1}^{K}\lambda_{mf} (\sum_{g=1}^{K} a_{fg}^{m} - 1) \} = 0 \Rightarrow \sum_{i=1}^{N} Z_{im} * \sum_{t=1}^{Ti-1} I_{it}^{m}(f,g) * \frac{1}{a_{fg}^{m}} - \lambda_{mf} = 0 \Rightarrow$$

$$\lambda_{mf} * a_{fg}^{m} = \sum_{i=1}^{N} Z_{im} \sum_{t=1}^{Ti-1} I_{it}^{m}(f,g) \Rightarrow \lambda_{mf} * \sum_{g=1}^{K} a_{fg}^{m} = \sum_{i=1}^{N} Z_{im} * \sum_{t=1}^{Ti-1}\sum_{g=1}^{K} I_{it}^{m}(f,g) \Rightarrow$$

$$\lambda_{mf} = \sum_{i=1}^{N} Z_{im} * \sum_{t=1}^{Ti-1}\sum_{g=1}^{K} I_{it}^{m}(f,g)$$

Οπότε:

$$\sum_{i=1}^{N} Z_{im} * \sum_{t=1}^{Ti-1} I_{it}^{m}(f,g) * \frac{1}{a_{fg}^{m}} = \sum_{i=1}^{N} Z_{im} * \sum_{t=1}^{Ti-1} \sum_{g=1}^{K} I_{it}^{m}(f,g) \Rightarrow$$

$$a_{fg}^{m} = \frac{\sum_{i=1}^{N} Z_{im} * \sum_{t=1}^{Ti-1} I_{it}^{m}(f,g) + \beta_{fg}}{\sum_{i=1}^{N} Z_{im} * \sum_{t=1}^{Ti-1} \sum_{g=1}^{K} I_{it}^{m}(f,g) + \sum_{g=1}^{K} \beta_{fg}}$$

(4) για το b:

$$\frac{\partial}{\partial\, b_{k}^{m}(l)} \{ Q^{L} - \sum_{m=1}^{M} \sum_{f=1}^{K} \lambda_{mf} (\sum_{l=1}^{L} b_{k}^{m}(l) - 1) \} = 0 \Rightarrow \sum_{i=1}^{N} Z_{im} * \sum_{t=1}^{Ti} I_{it}^{m}(k) * \delta(X_{it}, l) * \frac{1}{b_{k}^{m}(l)} - \lambda_{mk} = 0 \Rightarrow$$

$$\lambda_{mf} * b_{k}^{m}(l) = \sum_{i=1}^{N} Z_{im} \sum_{t=1}^{Ti} I_{it}^{m}(k) * \delta(X_{it}, l) \Rightarrow \lambda_{mf} * \sum_{l=1}^{L} b_{k}^{m}(l) = \sum_{i=1}^{N} Z_{im} * \sum_{t=1}^{Ti} I_{it}^{m}(k) \sum_{l=1}^{L} \delta(X_{it}, l) \Rightarrow$$

$$\lambda_{mf} = \sum_{i=1}^{N} Z_{im} * \sum_{t=1}^{Ti} I_{it}^{m}(k)$$

Οπότε:

$$\sum_{i=1}^{N} Z_{im} * \sum_{t=1}^{Ti} I_{it}^{m}(k) * \delta(X_{t}, l) * \frac{1}{b_{k}^{m}(l)} = \sum_{i=1}^{N} Z_{im} * \sum_{t=1}^{Ti} I_{it}^{m}(k) \Rightarrow$$

$$b_{k}^{m}(l) = \frac{\sum_{i=1}^{N} Z_{im} * \sum_{t=1}^{Ti} I_{it}^{m}(k) * \delta(X_{it}, l) + \beta_{kl}}{\sum_{i=1}^{N} Z_{im} * \sum_{t=1}^{Ti-1} I_{it}^{m}(k) + \sum_{l'=1}^{L} \beta_{kl'}}$$

Χρησιμοποιούμε τους ψευδοτυχαίους αριθμούς για αποφυγή αριθμητικών σφαλμάτων όπως προηγουμένως.

Τελειώνοντας το M – Step έχουμε υπολογίσει τις νέες τιμές των παραμέτρων $C_m, \pi_k^m, a_{fg}^m, b_k^m(l)$ με βάση τις τιμές της προηγούμενης επανάληψης. Αυτή η διαδικασία συνεχίζεται μέχρι να θεωρήσουμε ότι η πιθανοφάνεια του μικτού μοντέλου έχει μεγιστοποιηθεί. Αυτό το κάνουμε υπολογίζοντας την πιθανοφάνεια $P(X \mid \Theta)$ μετά από κάθε M – Step και συγκρίνοντας την τιμή της με αυτή στο αμέσως προηγούμενο βήμα. Εάν η διαφορά των δυο τιμών είναι αρκούντως μικρή (πχ. $< 10^{-2}$ ή 10^{-3}) θεωρούμε ότι η πιθανοφάνεια έχει συγκλίνει ικανοποιητικά στο τοπικό μέγιστο που ψάχνουμε και ο αλγόριθμος τερματίζει. Για πρακτικούς λόγους μπορούμε να ορίσουμε ένα μέγιστο πλήθος επαναλήψεων, οπότε να τερματίζει ο αλγόριθμος, ακόμα και εάν δεν έχει συγκλίνει.

ΚΕΦΑΛΑΙΟ 4. ΑΥΞΗΤΙΚΟ ΜΟΝΤΕΛΟ ΓΙΑ ΑΠΛΑ ΜΟΝΤΕΛΑ MARKOV

4.1. Αυξητική Μέθοδος

Δυο είναι τα βασικότερα ζητήματα που πρέπει να σκεφτούμε κατά τη χρήση του μικτού μοντέλου που περιγράψαμε παραπάνω:

Η επιλογή του πλήθους Μ των μοντέλων που θα χρησιμοποιήσουμε.
Η επιλογή των αρχικών τιμών των παραμέτρων.

Το πρόβλημα που θέλουμε να επιλύσουμε είναι η εύρεση των κατηγοριών που υπάρχουν στα δεδομένα. Αμέσως προκύπτει το ερώτημα πόσες κατηγορίες υπάρχουν πραγματικά στα δεδομένα μας. Αυτό είναι δύσκολο να το γνωρίζουμε. Συνήθως αυτή η πληροφορία δεν είναι γνωστή. Το μόνο που έχουμε είναι το σύνολο των δεδομένων, δηλαδή οι ακολουθίες παρατηρήσεων X_i. Επίσης, οι αρχικές τιμές των παραμέτρων έχουν σημαντική επίδραση στο αποτέλεσμα της εκπαίδευσης, αφού όπως είπαμε ο ΕΜ μεγιστοποιεί μόνο τοπικά την πιθανοφάνεια των δεδομένων. Το τοπικό μέγιστο που βρίσκει εξαρτάται από τις αρχικές τιμές των παραμέτρων. Επίσης, οι αρχικές

τιμές επηρεάζουν και των αριθμό των επαναλήψεων που απαιτούνται για την εύρεση του μεγίστου.

Τέλεια λύση δεν υπάρχει για να αντιμετωπίσουμε τα παραπάνω προβλήματα, αφού το μόνο που έχουμε στη διάθεσή μας είναι οι ακολουθίες συμβόλων X_i. Η τυχαία επιλογή του αριθμού των μοντέλων και των αρχικών τιμών των παραμέτρων δεν μπορεί να θεωρηθεί καλή λύση.

Για να αντιμετωπίσουμε το πρόβλημα προτείνουμε τη δημιουργία ενός μικτού μοντέλου με αυξητικό τρόπο. Η κεντρική ιδέα είναι να θεωρήσουμε αρχικά ότι όλα τα δεδομένα αναπαρίστανται από ένα μοντέλο και να αυξάνουμε κατά ένα τον αριθμό των μοντέλων σε κάθε βήμα μέχρι κάποιο σημείο, όπου θεωρούμε ότι έχουμε προσεγγίσει τον πραγματικό αριθμό των κατηγοριών. Ταυτόχρονα, η αρχικοποίηση των παραμέτρων των μοντέλων που προστίθενται δε γίνεται με τυχαίο τρόπο, αλλά προσπαθούμε το νέο μοντέλο να ταιριάξει πιο καλά σε κάποια κατηγορία ακολουθιών. Επομένως, σε δυο σημεία πρέπει να εστιάσουμε:

Στον τρόπο με τον οποίο δημιουργούμε το αυξητικό μικτό μοντέλο.
Στον αριθμό των μοντέλων που θα περιλαμβάνει, δηλαδή πότε πρέπει να σταματήσουμε την αύξηση του μικτού μοντέλου.

4.2. Αυξητική δημιουργία του μικτού μοντέλου Markov

Όπως είπαμε πρωτύτερα, προσπαθούμε να δημιουργήσουμε ένα μικτό μοντέλο με M συνιστώσες, ξεκινώντας από M=1 και προσθέτοντας ένα μοντέλο κάθε φορά μέχρι τον επιθυμητό αριθμό. Προς το παρόν δεν θα ασχοληθούμε με τον τελικό αριθμό των μοντέλων, αλλά μόνο με την περιγραφή του αυξητικού αλγορίθμου. Έστω λοιπόν ότι θέλουμε να δημιουργήσουμε ένα μικτό μοντέλο με M συνιστώσες. Η γενική ιδέα του αλγορίθμου είναι η ακόλουθη:

t=1: Ορισμός των παραμέτρων ενός μοναδικού μοντέλου που αναπαριστά όλο το σύνολο των δεδομένων X (Θ^1).

Στο βήμα t=m+1 έχουμε ήδη m μοντέλα και προσθέτουμε ένα ακόμα (με αρχικές παραμέτρους g^*, θ^*). Το νέο μοντέλο επιλέγεται από ένα σύνολο μοντέλων που έχουμε κατασκευάσει κατά την προεπεξεργασία με κριτήριο τη συνολική αύξηση της πιθανοφάνειας του μικτού μοντέλου με m+1 συνιστώσες.

Partial EM: Χρησιμοποιούμε τον ΕΜ αλγόριθμο και ρυθμίζουμε μόνο τις παραμέτρους του νέου μοντέλου (g^*, θ^*).

General EM: Εφαρμόζουμε τον ΕΜ και στα m+1 μοντέλα.

Η διαδικασία των βημάτων 2 έως 4 επαναλαμβάνεται μέχρι t=M.

Αυτό που προσπαθεί να επιτύχει η αυξητική μέθοδος είναι το εξής: Θεωρούμε ότι έχουμε αρχικά ένα μοντέλο που περιγράφει το σύνολο των δεδομένων. Κάθε φορά που προσθέτουμε ένα νέο μοντέλο προσπαθούμε να ταιριάξει καλύτερα στα δεδομένα μιας κατηγορίας και ταυτόχρονα τα υπόλοιπα εξειδικεύονται στις υπόλοιπες. Τελικά σταματάμε όταν φτάσουμε στον επιθυμητό αριθμό μοντέλων. Τότε προσδοκούμε ότι το αρχικό μοντέλο θα έχει διασπαστεί και κάθε ένα από αυτά που προστέθηκαν θα έχει προσαρμοστεί σε μια κατηγορία, ενώ και στο αρχικό θα αντιστοιχεί επίσης μια κατηγορία, την οποία δεν θα έχει αναπαραστήσει κανένα άλλο μοντέλο.

4.2.1. Αρχικό Μοντέλο

Το πρώτο βήμα είναι ο ορισμός των παραμέτρων του αρχικού μοντέλου. Αυτό είναι εύκολο και μπορεί να γίνει με στατιστικό τρόπο. Προφανώς, αφού υπάρχει μόνο ένα μοντέλο C_1=1. Θυμίζουμε, ότι οι καταστάσεις στο απλό μοντέλο αντιστοιχούν στα διαφορετικά σύμβολα που υπάρχουν, οπότε οι παρατηρήσεις X_{i1} αντιστοιχούν και στις καταστάσεις που συναντάμε.

Οι πιθανότητες αρχικών καταστάσεων βρίσκονται με καταμέτρηση των εμφανίσεων των συμβόλων στην πρώτη θέση κάθε ακολουθίας, δηλαδή: $\pi_k = \frac{\sum_{i=1}^{N} \delta(X_{i1}, k)}{N}$.

Οι πιθανότητες μεταβάσεων a_{fg} βρίσκονται με καταμέτρηση των μεταβάσεων από την f στην g. προς τον αριθμό των συνολικών μεταβάσεων από την f σε άλλη κατάσταση,

$$a_{fg} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{Ti-1}\delta(X_{it}, f, X_{it+1}, g)}{\sum_{i=1}^{N}\sum_{t=1}^{Ti-1}\delta(X_{it}, f)}$$

δηλαδή: .

4.2.2. Προσθήκη Μοντέλου

Το κρίσιμο βήμα του αλγορίθμου είναι η προσθήκη του m+1 μοντέλου. Έχουμε ήδη m μοντέλα. Θεωρούμε ότι οι κατηγορίες που υπάρχουν στα δεδομένα μας είναι περισσότερες από m, οπότε ένα ή περισσότερα μοντέλα αναπαριστούν περισσότερες από μια κατηγορίες. Γι' αυτό το λόγο θέλουμε να προσθέσουμε ένα ακόμα, το οποίο θα βοηθήσει στην καλύτερη κατηγοριοποίηση. Το ζήτημα είναι ο καθορισμός των αρχικών παραμέτρων του νέου μοντέλου. Σκοπός είναι το νέο μοντέλο να μην προστεθεί τυχαία, αλλά με κάποια λογικά κριτήρια, ώστε να είναι πιο εύκολο να αναπαραστήσει κάποια κατηγορία. Αυτό που κάνουμε είναι να επιλέξουμε ένα μοντέλο, το οποίο θα μεγιστοποιεί την πιθανοφάνεια του μικτού μοντέλου.

Θεωρούμε ότι έχουμε ένα σύνολο έτοιμων μοντέλων, από τα οποία επιλέγουμε ένα, το οποίο προβλέπουμε ότι θα μεγιστοποιεί την πιθανοφάνεια του νέου m+1 μικτού

$$P(X \mid \Theta^{m}) = \sum_{m'=1}^{m} C_{m'} * P(X \mid \theta^{m'})$$

μοντέλου. Έστω λοιπόν η συνάρτηση για το μικτό μοντέλο με m συνιστώσες που υπάρχει ήδη. Αν θεωρήσουμε ότι προσθέτουμε ένα μοντέλο ακόμα το θ^{*} η συνάρτηση πιθανότητας για το μικτό μοντέλο με m+1 συνιστώσες θα είναι:

$$P(X \mid \Theta^{m+1}) = (1 - g^{*}) * \sum_{m'=1}^{m} C_{m'} * P(X \mid \theta^{m'}) + g^{*} * P(X \mid \theta^{*})$$

. Η παράμετρος g^{*} λαμβάνει μια αρχική τιμή ($g^{*} = 1 / (m+1)$), η οποία προσαρμόζεται κατάλληλα κατά την

εκπαίδευση που ακολουθεί. Θυμίζουμε ότι ισχύει $\sum_{m'=1}^{m} C_{m'} = 1$ για το m μικτό μοντέλο. Επομένως, ισχύει $(1-g)*[\sum_{m'=1}^{m} C_{m'}] + g = 1$. Πολλαπλασιάζοντας τα στοιχεία C_m με 1-g^* προκειμένου να διατηρήσουμε το άθροισμα $\sum_{m'=1}^{m+1} C_{m'} = 1$ για το m+1 μικτό μοντέλο.

Ουσιαστικά, θεωρούμε ότι το νέο μοντέλο που προκύπτει είναι ένα μικτό μοντέλο με δυο συνιστώσες. Η μια είναι το μοντέλο θ^* που προσθέτουμε, με βάρος g^* και η άλλη το μικτό μοντέλο με m συνιστώσες που προϋπήρχε με βάρος 1-g^*. Με αυτό τον τρόπο επιχειρούμε να βελτιστοποιήσουμε την πιθανοφάνεια του γενικού μικτού μοντέλου με m+1 συνιστώσες βελτιστοποιώντας την πιθανοφάνεια του παραπάνω μικτού μοντέλου με 2 συνιστώσες **[3, 12]**. Αυτό το πετυχαίνουμε εφαρμόζοντας τον ΕΜ πρώτα για τη νέα συνιστώσα (Partial EM) και ακολούθως σε ολόκληρο το μικτό μοντέλο (General EM).

4.2.3. Partial EM

Θεωρώντας τις αρχικές τιμές $\{g^* \ \theta^*\}$ για το νέο μοντέλο εφαρμόζουμε μερικό ΕΜ, προκειμένου να ρυθμίσουμε τις παραμέτρους μόνο για το νέο μοντέλο. μεγιστοποιώντας, την πιθανοφάνεια για ολόκληρο το μικτό μοντέλο. Δηλαδή κριτήριο για τον τερματισμό του Partial EM είναι η μείωση της συνολικής πιθανοφάνειας του νέου μικτού μοντέλου.

Οι εξισώσεις του ΕΜ είναι οι ίδιες όπως στο απλό μικτό μοντέλο της παραγράφου 3.2. Η διαφορά είναι ότι εφαρμόζονται μόνο για το νέο μοντέλο που προσθέτουμε.

E- Step:

$$E[Z_i]^t = \frac{g^{*t-1} * P(X_i \mid \theta^{*t-1})}{(1-g^*) * \sum_{m'=1}^{m} C_{m'}^{t-1} * P(X_i \mid \theta_{m'}^{t-1}) + g * P(X_i \mid \theta^{*t-1})}$$

M – Step:

$$g^{*} = \frac{\sum_{i=1}^{N} E[Z_i]}{N} \qquad \pi_k^{*} = \frac{\sum_{i=1}^{N} E[Z_{im+1}] * \delta_1(X_{i1}, S_k) + \beta_k}{\sum_{i'=1}^{N} E[Z_{i'm'}] + \sum_{k'=1}^{K} \beta_{k'}}$$

$$a_{fg}^{*} = \frac{\sum_{i=1}^{N} E[Z_i] * \sum_{t=1}^{Ti-1} \delta_2(X_{it}, S_f, X_{it+1}, S_g) + \beta_{fg}}{\sum_{i'=1}^{N} E[Z_{i'}] * \sum_{t'=1}^{Ti-1} \sum_{g'=1}^{K} \delta_2(X_{i't'}, S_f, X_{i't'+1}, S_{g'}) + \sum_{g'=1}^{K} \beta_{fg'}}$$

Εφαρμόζοντας τον ΕΜ μόνο στο νέο μοντέλο το αποτέλεσμα είναι να ρυθμιστούν κατάλληλα οι τιμές των παραμέτρων του, ώστε να αναπαραστήσει καλύτερα τα δεδομένα της κατηγορίας που του ταιριάζουν. Ταυτόχρονα, ρυθμίζονται κατάλληλα οι τιμές των παραμέτρων C για το m+1 μικτό μοντέλο που προκύπτει.

4.2.4. General EM

Σε αυτό το βήμα εφαρμόζουμε τον κανονικό ΕΜ προκειμένου να εκπαιδεύσουμε και τις m+1 συνιστώσες του μικτού μοντέλου με συνάρτηση πιθανότητας:

$$P(X \mid \Theta^{m+1}) = \sum_{m'=1}^{m+1} C_{m'} P(X \mid \theta^{m'})$$

Οι αρχικές τιμές των παραμέτρων για την εφαρμογή του γενικού ΕΜ είναι:

$C_{m+1} = g^{*}$ και $Cj = (1-g^{*})Cj \quad j=1,\ldots,m$

$\theta_{m+1} = \theta^{*}$ και $\theta j = \theta j$

όπου g^{*}, θ^{*} αυτές που υπολογίστηκαν από τον Partial EM του προηγήθηκε. Οι εξισώσεις του Ε – Step και του Μ – Step είναι ξανά αυτές που δώσαμε στην παράγραφο 3.2.

Η εφαρμογή του ΕΜ στο μοντέλο με m+1 συνιστώσες τροποποιεί κατάλληλα τις παραμέτρους όλων των μοντέλων, ώστε να ταιριάξουν καλύτερα στις κατηγορίες των δεδομένων του προβλήματος.

4.2.5. Επιλογή μοντέλου

Έχοντας υπολογίσει την πιθανοφάνεια του μικτού μοντέλου με m+1 συνιστώσες επιλέγουμε να προσθέσουμε το μοντέλο που μεγιστοποιεί την πιθανοφάνεια, δηλαδή:

$$\theta^* = \underset{\Theta^*}{\operatorname{argmax}}\{ \log(P(X \mid \Theta^{m+1})) \} =$$

$$\underset{\Theta^*}{\operatorname{argmax}}(\sum_{i=1}^{N} \log((1 - g^*) * \sum_{m'=1}^{m} C_{m'} * P(X_i \mid \theta^{m'}) + g^* * P(X_i \mid \theta^*))$$

Ιδιαίτερα σημαντικό ζήτημα είναι ποια είναι τα έτοιμα μοντέλα από τα οποία επιλέγουμε κάθε φορά για να προσθέσουμε. Αυτό που κάνουμε είναι να κατασκευάζουμε ένα σύνολο από μοντέλα χρησιμοποιώντας το σύνολο δεδομένων. Η διαδικασία κατασκευής των μοντέλων αποτελείται από δυο στάδια:

Ομαδοποίηση των ακολουθιών X_i σε R ομάδες με τη χρήση του αλγορίθμου K – Means.
Κατασκευή ενός μοντέλου από κάθε ομάδα που σχηματίστηκε.

4.2.6. K – Means

Ο K – Means είναι ένας απλός αλγόριθμος ομαδοποίησης ενός συνόλου δεδομένων σε K_m ομάδες, όπου ο αριθμός K_m είναι προκαθορισμένος **[7]**. Στο πρόβλημα που εξετάζουμε έχουμε N ακολουθίες δεδομένων τις οποίες θέλουμε να χωρίσουμε σε K_m ομάδες. Ο αλγόριθμος λειτουργεί ως εξής:

- Επιλέγονται K_m ακολουθίες

- αρχικά κέντρα των ομάδων - Means(1:K_m)
- Αντιστοιχίζουμε κάθε X_i στο κέντρο που απέχει λιγότερο
  - Min{ distance(X_i, Means(j)) } j=1:K_m
- Για κάθε ομάδα βρίσκουμε το νέο κέντρο
  - Min{ sum(distance(X_i, X_j)) } j∈Group(i)
- Επανάληψη 2 και 3 μέχρι τη σταθεροποίηση των κέντρων

Ο αλγόριθμος απαιτεί τον ορισμό της απόστασης μεταξύ των ακολουθιών. Η απόσταση είναι ο βαθμός ομοιότητας μεταξύ δυο ακολουθιών. Την απόσταση μπορούμε να την ορίσουμε με δυο διαφορετικούς τρόπους.

Ο πρώτος τρόπος είναι χρησιμοποιώντας την πιθανοφάνεια. Κατασκευάζουμε ένα μοντέλο Markov M_i για κάθε ακολουθία X_i του συνόλου δεδομένων. Για κάθε ακολουθία υπολογίζουμε την λογαριθμική πιθανοφάνεια $L(i, j) = \log(P(X_i \mid M_j))$ με τον γνωστό τρόπο. Η απόσταση D μεταξύ δυο ακολουθιών X_i και X_j μπορεί να οριστεί τότε ως **D(i, j) = 1/2 * (L(i,j)+L(j,i))**. Υψηλή τιμή πιθανοφάνειας σημαίνει μεγάλο βαθμό ομοιότητας, συνεπώς μικρή απόσταση μεταξύ των ακολουθιών. Ο υπολογισμός της απόστασης χρειάζεται να γίνει μόνο μια φορά.

Ένας δεύτερος τρόπος είναι χρησιμοποιώντας το score ευθυγράμμισης που μπορούν να μας δώσουν δυο γνωστοί αλγόριθμοι: Ο αλγόριθμος τοπικής ευθυγράμμισης των Smith – Waterman και ο αλγόριθμος ολικής ευθυγράμμισης των Needleman – Wunsch **[9]**. Υψηλό score ευθυγράμμισης σημαίνει υψηλό βαθμό ομοιότητας, οπότε και μικρή απόσταση μεταξύ των ακολουθιών. Ο υπολογισμός της απόστασης μπορεί να γίνει μια φορά για όλες τις ακολουθίες, αφού δε μεταβάλλεται.

Μπορούμε να χρησιμοποιήσουμε και τους δυο τρόπους (score ευθυγράμμισης – Πιθανοφάνεια) για να υπολογίσουμε μια απόσταση μεταξύ των ακολουθιών. Προτιμούμε τη λύση της πιθανοφάνειας. Η απόσταση με την πιθανοφάνεια δεν εξαρτάται από παραμέτρους που δεν έχουν άμεση σχέση με το πρόβλημα, όπως

πίνακες score για ταίριασμα μεταξύ συμβόλων που πρέπει να ορίσουμε για την ευθυγράμμιση ακολουθιών.

Συνεπώς, στο βήμα 2 του K – Means για κάθε ακολουθία βρίσκουμε σε ποια ομάδα ανήκει υπολογίζοντας την απόσταση από κάθε κέντρο και επιλέγοντας αυτήν, από την οποία έχει τη μικρότερη. Στο βήμα 3 έχοντας υπολογίσει τα μέλη για κάθε ομάδα επαναπροσδιορίζουμε το κέντρο της ομάδας. Το νέο κέντρο είναι αυτό, το οποίο έχει συνολικά τη μικρότερη απόσταση από όλα τα μέλη της ομάδας, δηλαδή αυτό με το μικρότερο άθροισμα αποστάσεων από τα υπόλοιπα μέλη. Όταν όλα τα κέντρα πάψουν να μεταβάλλονται θεωρούμε ότι έχει γίνει η κατηγοριοποίηση.

Είναι φανερό ότι το αποτέλεσμα του K – Means εξαρτάται από την επιλογή του αριθμού των ομάδων καθώς και από την αρχική επιλογή των κέντρων. Ειδικά η επιλογή των κέντρων είναι καθοριστικής σημασίας. Η τυχαία αρχικοποίηση δεν κρίνεται ικανοποιητική λύση. Εάν κατά την αρχικοποίηση δεν επιλεγούν κέντρα που να αντιπροσωπεύουν όλες τις ομάδες ο K – Means δεν θα μπορέσει να βρει όλες τις ομάδες. Για να αντιμετωπίσουμε το πρόβλημα δεν επιλέγουμε τα αρχικά κέντρα εντελώς τυχαία. Αυτό που κάνουμε είναι να επιλέξουμε κέντρα που απέχουν όσο το δυνατό περισσότερο μεταξύ τους. Ο αλγόριθμος αρχικοποίησης των κέντρων είναι ο ακόλουθος:

Για k=1: Επιλέγουμε το πρώτο κέντρο τυχαία.
Για t=k+1: Επιλέγουμε κάθε επόμενο κέντρο αυτό που έχει τη μεγαλύτερη απόσταση από όλα τα υπάρχοντα k κέντρα. Για κάθε X_i υπολογίζουμε το άθροισμα των αποστάσεων από όλα τα υπάρχοντα κέντρα και επιλέγουμε αυτό με τη μεγαλύτερη συνολική απόσταση. Δηλαδή $\text{new} = \underset{i}{\text{argmax}} \sum_{j=1}^{N} D(i, j)$. Υπενθυμίζουμε ότι η απόσταση είναι αντιστρόφως ανάλογη με την τιμή του D, γι' αυτό το λόγο παίρνουμε το μέγιστο κι όχι το ελάχιστο.
Για t= K_m: Έχουμε βρει K_m κέντρα που απέχουν τη μέγιστη απόσταση μεταξύ τους.

Ένα γενικό πρόβλημα αυτής της τεχνικής είναι ότι μπορεί να αρχικοποιεί τα κέντρα σε ακολουθίες που ίσως να μην είναι αντιπροσωπευτικές κάποιας ομάδας, αλλά να αποτελούν εξαιρέσεις στο σύνολο των δεδομένων (outliers). Συνήθως τέτοια κέντρα δεν είναι καλές επιλογές. Στη δική μας περίπτωση δεν είναι σημαντικό το πρόβλημα διότι:

Η επιλογή των κέντρων γίνεται από τα δεδομένα και όχι με τυχαίες αρχικές τιμές. Επομένως ο κίνδυνος επιλογής outlier είναι μικρός.
Επιλέγουμε μεγάλο αριθμό κέντρων, οπότε είναι σχεδόν σίγουρο ότι θα επιλέξουμε κέντρα από όλες τις κατηγορίες, αφού η επιλογή γίνεται με βάση την απόσταση μεταξύ των κέντρων κι όχι τυχαία.
Ακόμα κι αν σε κάποια ακραία περίπτωση επιλέξουμε πολλά outliers, κατά την αναπροσαρμογή των κέντρων αναμένουμε να γίνει ανακατανομή σε πιο σωστά σημεία.

4.2.7. Κατασκευή Αρχικών Μοντέλων

Έχοντας δημιουργήσει R ομάδες ακολουθιών από τα δεδομένα μπορούμε να κατασκευάσουμε από αυτές R μοντέλα για τον αυξητικό αλγόριθμο. Η κατασκευή γίνεται με παρόμοιο τρόπο με το αρχικό μοντέλο του αυξητικού αλγορίθμου που είδαμε πρωτύτερα. Δηλαδή υπολογίζουμε τις παραμέτρους του μοντέλου με στατιστικό τρόπο. Η μόνη διαφορά είναι ότι κάθε μοντέλο κατασκευάζεται από τις ακολουθίες της ομάδας που ανήκει και όχι λαμβάνοντας υπόψη όλο το σύνολο των ακολουθιών.

4.3. Τερματισμός Αυξητικού Αλγορίθμου – Πλήθος Συνιστωσών

Σημαντικό ζήτημα είναι η επιλογή του πλήθους των συνιστωσών του μικτού μοντέλου, αφού αυτό είναι και το τελικό ζητούμενο: Η εύρεση των κατηγοριών στα δεδομένα. Γενικά δεν μπορούμε να θεωρήσουμε ότι έχουμε κάποια γνώση για τον

αριθμό των κατηγοριών. Μόνο οι ακολουθίες παρατηρήσεων X_i είναι διαθέσιμες. Χωρίς την αυξητική μέθοδο κατασκευής του μικτού μοντέλου θα έπρεπε να εκπαιδεύσουμε έναν αρκετά μεγάλο αριθμό από διαφορετικά μικτά μοντέλα για διάφορες τιμές του M (M=1, 2, …, M*) και να επιλέγαμε κάποιο που θεωρούσαμε καλύτερο. Αντίστοιχα, στο αυξητικό μοντέλο πρέπει να αποφασίσουμε πότε θα σταματήσουμε να προσθέτουμε νέα μοντέλα.

Η πιθανοφάνεια του μικτού μοντέλου στο σύνολο των δεδομένων εκπαίδευσης δεν μπορεί να είναι κριτήριο. Αυξάνοντας το πλήθος των μοντέλων και εκπαιδεύοντας ολοένα και πιο πολύπλοκα μικτά μοντέλα, η πιθανοφάνεια αυξάνει συνεχώς, αφού το μικτό μοντέλο προσαρμόζεται περισσότερο στα ιδιαίτερα χαρακτηριστικά των δεδομένων που χρησιμοποιούνται στην εκπαίδευση. Αυτό όμως δε σημαίνει ότι βελτιώνεται συνολικά η ικανότητα ορθής κατηγοριοποίησης του συστήματος. Αντίθετα, η γενικευτική ικανότητα του συνολικού συστήματος θα μειώνεται, αφού τα επιπλέον μοντέλα απλά χωρίζουν τις πραγματικές κατηγορίες σε περισσότερες, δίνοντας βάση περισσότερο στο θόρυβο που μπορεί να εμφανίζεται παρά στα πραγματικά πρότυπα που υπάρχουν. Είναι προφανές άλλωστε, ότι αν υπάρχουν M κατηγορίες στην πραγματικότητα και θεωρήσουμε ότι υπάρχουν περισσότερες από M το μοντέλο μας δεν ανταποκρίνεται ακριβώς στις απαιτήσεις του προβλήματος.

Πρέπει λοιπόν να ορίσουμε κάποιο κριτήριο με βάση το οποίο θα προσεγγίζουμε ικανοποιητικά το πλήθος των μοντέλων M. Αυτό που μπορεί να γίνει είναι να παρατηρήσουμε τη συμπεριφορά του συστήματος σε ένα άλλο σύνολο δεδομένων, τα οποία δεν έχουν χρησιμοποιηθεί κατά την εκπαίδευση. Το σύνολο αυτό ονομάζεται σύνολο ελέγχου (ή επικύρωσης) και βοηθά στην αξιολόγηση της γενικευτικής ικανότητας του μοντέλου.

Το μέτρο για την αξιολόγηση του μικτού μοντέλου και την επιλογή του M βασίζεται στην πιθανοφάνεια για το σύνολο ελέγχου. Επιλέγουμε λοιπόν το μοντέλο με M συνιστώσες που ελαχιστοποιεί την παρακάτω ποσότητα στο σύνολο ελέγχου **[4]**:

$$\text{Score}(\Theta^M, X^{test}) = -\frac{\sum_{i=1}^{N} \log_2 P(X_i^{test} \mid \Theta^M)}{\sum_{i=1}^{N} \text{length}(X_i^{test})}.$$

Η παραπάνω ποσότητα είναι ανάλογη της πιθανοφάνειας στο σύνολο ελέγχου. Μεγιστοποιώντας την πιθανοφάνεια βελτιώνουμε την πρόβλεψη του score. Ο δυαδικός λογάριθμος της πιθανοφάνειας με την κανονικοποίηση από το μήκος των ακολουθιών αντιστοιχεί στο μέσο αριθμό bits που απαιτούνται για την κωδικοποίηση των κατηγοριών.

Χρησιμοποιώντας αυτή την τεχνική μπορούμε να εκτιμήσουμε πότε πρέπει να τερματίσουμε τον αυξητικό αλγόριθμο, δηλαδή πότε έχουμε φτάσει στο κατάλληλο πλήθος συνιστωσών και δε χρειάζεται να προσθέσουμε νέα μοντέλα.

Κάθε φορά που προσθέτουμε ένα νέο μοντέλο εφαρμόζουμε πρώτα μερικό ΕΜ για το νέο μοντέλο και ακολούθως γενικό ΕΜ και στις Μ+1 συνιστώσες. Μετά το πέρας του γενικού ΕΜ υπολογίζουμε το score του νέου μικτού μοντέλου. Αν δούμε ότι η προσθήκη του νέου μοντέλου δεν επιφέρει σημαντική μεταβολή στην τιμή του score (δηλαδή το score έχει συγκλίνει στην βέλτιστη τιμή) τότε καταλαβαίνουμε ότι η προσθήκη του επιπλέον μοντέλου δεν είχε όφελος και το μικτό μοντέλο με Μ συνιστώσες είναι αρκετό για την επίλυση του προβλήματος.

ΚΕΦΑΛΑΙΟ 5. ΠΕΙΡΑΜΑΤΑ – ΣΥΜΠΕΡΑΣΜΑΤΑ

5.1. Εισαγωγή

Σε αυτό το κεφάλαιο παρουσιάζουμε τα αποτελέσματα της κατηγοριοποίησης των μικτών μοντέλων σε πειραματικά αλλά και σε πραγματικά δεδομένα. Τα μικτά μοντέλα που εξετάζουμε είναι:

Το απλό μοντέλο Markov, όπου οι παράμετροι λαμβάνονται προσθέτοντας θόρυβο στο στατιστικό μοντέλο αντιστοιχεί σε όλο το σύνολο δεδομένων (Random Markov Model).
Το απλό μοντέλο Markov, όπου οι παράμετροι των αρχικών μοντέλων προέρχονται από εφαρμογή του αλγορίθμου K – Means (K – Means Markov Model).
Το απλό μοντέλο Markov που κατασκευάζει ο αυξητικός αλγόριθμος που ορίσαμε (Incremental Markov Model).
Επιπλέον, πραγματοποιούμε κάποια πειράματα και για το κρυμμένο μοντέλο Markov, του οποίου το πλήθος των καταστάσεων και οι τιμές των παραμέτρων αρχικοποιούνται σε τυχαίες τιμές (Hidden Markov Model).

5.2. Πειραματικά Δεδομένα

Τα πειραματικά δεδομένα κατασκευάζονται με σκοπό να μελετήσουμε τη συμπεριφορά των διαφορετικών μικτών μοντέλων. Τα πειραματικά δεδομένα κατασκευάζονται με τρεις τρόπους:

Από ένα σύνολο προκατασκευασμένων ακολουθιών παράγουμε τις ακολουθίες του συνόλου δεδομένων προσθέτοντας θόρυβο. Κάνουμε δηλαδή sampling από αυτές τις ακολουθίες. Με αυτό τον τρόπο μπορούμε να κατασκευάσουμε ακολουθίες με ξεκάθαρα ή αλλοιωμένα πρότυπα για όσες κατηγορίες επιθυμούμε. Ένα παράδειγμα είναι η παρακάτω ακολουθία: [1 2 3 4 5 5 5 5 5 5 5 5 2 2 2 2 2 6 6 6 6 6 6 7 6 7 6 7 4 3 4 3 4 3 4 3 4 3 4 3 9 2 8 4 8 4 8 4 8 4 8 4 8 4 4 4 4 4 4 6 4 6 4 6 4 6 4 6 6 6 6 6 6 6 6 6 6]. Αυτή η μέθοδος παράγει ακολουθίες όπου τα χαρακτηριστικά των ομάδων είναι ευδιάκριτα για χαμηλό ποσοστό θορύβου ή δυσδιάκριτα για υψηλό ποσοστό θορύβου.
Χρησιμοποιούμε ένα σύνολο προκατασκευασμένων προτύπων και τα εισάγουμε αυτούσια ή αλλοιωμένα με θόρυβο μέσα σε ακολουθίες που παράγονται τυχαία. Αυτή η μέθοδος δίνει πρότυπα που είναι λιγότερο ευδιάκριτα λόγω της ομοιομορφίας των ακολουθιών στις οποίες εισάγουμε τα πρότυπα. Παράδειγμα ατόφιου προτύπου που εισάγουμε: [1 1 1 1 5 6 5 6 5 6].
Τυχαία παραγόμενα πρότυπα που εμφυτεύονται μέσα σε ακολουθίες όπως αυτά της δεύτερης μεθόδου. Τα πρότυπα που παράγονται παρουσιάζουν μεγαλύτερη ομοιομορφία από αυτά των δυο προηγούμενων μεθόδων και είναι πιο δύσκολο να ανιχνευθούν λόγω του ομοιόμορφα τυχαίου τρόπου παραγωγής των προτύπων.

Η προσθήκη θορύβου για την αλλοίωση των προτύπων επιτυγχάνεται μεταλλάσσοντας τα σύμβολα των προτύπων με μια πιθανότητα. Πχ. Θεωρώντας 40% θόρυβο, για κάθε σύμβολο του προτύπου υπάρχει πιθανότητα 0.4 να μεταλλαχθεί σε κάποιο άλλο σύμβολο. Το νέο σύμβολο επιλέγεται τυχαία τρόπο. Επομένως, η προσθήκη θορύβου καθιστά περισσότερο ομοιόμορφες τις ακολουθίες, δυσχεραίνοντας την ομαδοποίησή τους.

Για κάθε σύνολο δεδομένων κατασκευάζουμε δυο υποσύνολα. Ένα για να χρησιμοποιήσουμε στην εκπαίδευση του μικτού μοντέλου και ένα μικρότερο για σύνολο ελέγχου, το οποίο εκτός από τη μέτρηση της γενικευτικής ικανότητας του μοντέλου το χρησιμοποιούμε και για τον υπολογισμού του score στον τερματισμό του αυξητικού μοντέλου. Το μέγεθος των συνόλων είναι 100* (πλήθος κατηγοριών), αλλά ο αριθμός των ακολουθιών που ανήκουν σε κάθε κατηγορία έχει διακυμάνσεις.

Τα πραγματικά δεδομένα προέρχονται από το site MSNBC.COM και πρόκειται για ακολουθίες που έχουν δημιουργηθεί από την επεξεργασία των αρχείων καταγραφής του αντίστοιχου server (log files). Οι ακολουθίες αναπαριστούν τις κινήσεις των χρηστών του δικτυακού τόπου με τη χρονολογική σειρά που πραγματοποιήθηκαν. Το σύνολο δεδομένων είναι ιδιαίτερα μεγάλο σε μέγεθος. Επίσης, τα μήκη των ακολουθιών παρουσίαζαν σημαντικές διαφορές. Γι' αυτό το λόγο, το αρχικό σύνολο υπέστη κατάλληλη προεπεξεργασία και από αυτό δημιουργήθηκαν μικρότερα σύνολα με τυχαία δειγματοληψία.

5.3. Μετρήσεις

Στα πειράματα μετράμε την πιθανοφάνεια των μικτών μοντέλων όπως προκύπτει από την εκπαίδευσή τους. Επίσης, γνωρίζοντας σε ποια κατηγορία ανήκει κάθε ακολουθία μπορούμε να μετρήσουμε τα ποσοστά επιτυχούς κατηγοριοποίησης τόσο στα σύνολα εκπαίδευσης, όσο και στα σύνολα ελέγχου. Υπολογίζουμε τους μέσους όρους και τις αποκλίσεις των παραπάνω μετρήσεων και συγκρίνουμε τις επιδόσεις των μικτών μοντέλων. Τα πειράματα εστιάζονται στις επιδόσεις του αυξητικού μοντέλου σε σύγκριση με το τυχαίο μοντέλο και το μοντέλο που κατασκευάζεται με αρχικοποίηση των παραμέτρων από τον K – Means, τα οποία είναι άμεσα συγκρίσιμα. Σε κάθε πείραμα κατασκευάζουμε ένα μικτό μοντέλο με τον αυξητικό αλγόριθμο και 20 μοντέλα με τυχαία αρχικοποίηση και αρχικοποίηση με τον K – means. Για αυτά τα μοντέλα μετράμε τη μέση τιμή, την τυπική απόκλιση καθώς και πόσες φορές βρέθηκε η βέλτιστη λύση για την πιθανοφάνεια και τα ποσοστά επιτυχίας.

Στη συνέχεια παρουσιάζονται αναλυτικοί πίνακες με τα αποτελέσματα των παραπάνω μικτών μοντέλων στα προηγούμενα σύνολα δεδομένων καθώς και τα συμπεράσματα που προκύπτουν από τη μελέτη των αποτελεσμάτων. Στους πίνακες απεικονίζονται:

Για την πιθανοφάνεια:
Στις στήλες *Μέγιστη* και *Ελάχιστη* η μέγιστη και η ελάχιστη τιμή της πιθανοφάνειας (για το αυξητικό μοντέλο υπάρχει μόνο μια τιμή).
Στη στήλη *Επιτυχίες* το ποσοστό των μοντέλων που συνέκλιναν κοντά στην καλύτερη τιμή πιθανοφάνειας της μεθόδου. (πχ για το τυχαίο μοντέλο, εάν από τα 20 μοντέλα που εκπαιδεύτηκαν τα 10 είχαν πιθανοφάνεια -20000 και τα άλλα 10 -22000 τότε λέμε ότι η βέλτιστη λύση ήταν το -20000 και σε αυτή συνέκλινε το 50% των μοντέλων).
Για το ποσοστό επιτυχούς κατηγοριοποίησης:
Στις στήλες *Μέση τιμή* και *Απόκλιση* η μέση τιμή και η τυπική απόκλιση του ποσοστού των ακολουθιών που ομαδοποιήθηκαν σωστά.

Επίσης, για ορισμένα παραδείγματα υπάρχει και οπτική παρουσίαση των αποτελεσμάτων του αυξητικού αλγορίθμου με bitmaps. Κάθε γραμμή της εικόνας απεικονίζει μια από τις ακολουθίες που ανήκουν στην ομάδα. Κάθε σύμβολο κωδικοποιείται με διαφορετικό χρώμα οπότε φαίνεται ο βαθμός ομοιότητας μεταξύ των ακολουθιών της ίδιας ομάδας, αλλά και η διαφοροποίηση μεταξύ των ομάδων.

5.3.1. Τεχνητό σύνολο δεδομένων 1

Εδώ τα σύνολα δεδομένων κατασκευάζονται με τον πρώτο τρόπο που αναφέραμε (δειγματοληψία από έτοιμες συμβολοσειρές). Για κάθε αριθμό κατηγοριών M = { 4, 6, 8, 10 } έχουμε σύνολα ακολουθιών που κατασκευάστηκαν με διαφορετικό ποσοστό θορύβου P_m={40%, 50%, 60%, 70%}.

Για M=4

Πίνακας 5.1 M=4 Θόρυβος 40%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-48756**			**100,000%**	
K - Means EM	**-48756**	**-48756**	**100,00%**	**100,000%**	**0,000**
Random EM	**-48756**	**-50121**	**80,00%**	**99,238%**	**0,024**

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-7527**			**100,000%**	
K - Means EM	**-7527**	**-7527**	**100,00%**	**100,000%**	**0,000**
Random EM	**-7527**	**-8352**	**80,00%**	**99,550%**	**1,605**

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-52216**			**99,250%**	
K - Means EM	**-52216**	**-52216**	**100,00%**	**99,488%**	**0,001**
Random EM	**--52216**	**-53073**	**85,00%**	**98,413%**	**0,027**

Πίνακας 5.2 M=4 Θόρυβος 50%

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-8710**			**100,000%**	
K - Means EM	**-8710**	**-8711**	**90,00%**	**100,000%**	**0,000**
Random EM	**-8710**	**-9526**	**60,00%**	**99,600%**	**1,569**

Πίνακας 5.3 M=4 Θόρυβος 60%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-54356**			**97,750%**	
K - Means EM	**-54356**	**-54676**	**95,00%**	**97,425%**	**0,023**
Random EM	**-54356**	**-54681**	**95,00%**	**9733,750%**	**0,026**

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-9967**			**100,000%**	
K - Means EM	**-9967**	**-10596**	**15,00%**	**99,950%**	**0,224**
Random EM	**-9967**	**-10528**	**35,00%**	**100,000%**	**0,000**

Πίνακας 5.4 M=4 Θόρυβος 70%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-55794**			**77,500%**	
K - Means EM	**-55772**	**-55836**	**55,00%**	**77,338%**	**0,025**
Random EM	**-55772**	**-55880**	**60,00%**	**76,363%**	**0,050**

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-11693**			**100,000%**	
K - Means EM	**-11592**	**-12056**	**15,00%**	**99,350%**	**2,159**
Random EM	**-11577**	**-12041**	**5,00%**	**99,500%**	**2,236**

Για M=6

Πίνακας 5.5 M=6 Θόρυβος 40%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-75544			100,000%	
K - Means EM	-75544	-76769	50,00%	97,158%	0,033
Random EM	-75544	-76920	60,00%	98,300%	0,028

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-12085			100,000%	
K - Means EM	-12085	-12932	50,00%	99,100%	1,864
Random EM	-12085	-12884	60,00%	99,767%	0,583

Πίνακας 5.6 M=6 Θόρυβος 50%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-80675			98,667%	
K - Means EM	-80675	-81281	65,00%	96,933%	0,029
Random EM	-80675	-81277	40,00%	96,125%	0,025

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-14134			100,000%	
K - Means EM	-14134	-14812	10,00%	98,800%	2,741
Random EM	-14134	-14831	30,00%	99,633%	0,823

Πίνακας 5.7 M=6 Θόρυβος 60%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέση Τιμή	Απόκλιση	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-83401			95,000%	
K - Means EM	-83401	-83632	50,00%	9313,333%	0,023
Random EM	-83401	-83657	50,00%	92,875%	0,024

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-15724			100,000%	
K - Means EM	-15724	-16307	35,00%	99,767%	0,726
Random EM	-15724	-16300	35,00%	99,000	1,640

Πίνακας 5.8 M=6 Θόρυβος 70%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-85730			73,000%	
K - Means EM	-85678	-85746	5,00%	68,892%	0,032
Random EM	-85679	-85745	5,00%	67,350%	0,042

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-18425			97,333%	
K - Means EM	-17944	-18490	10,00%	98,867%	1,772
Random EM	-18018	-18538	5,00%	97,800%	2,109

Για M=8

Πίνακας 5.9 M=8 Θόρυβος 50%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-102548**			**100,000%**	
K - Means EM	**-103378**	**-104544**	**5,00%**	**93,994%**	**0,029**
Random EM	**-102548**	**-103946**	**50,00%**	**98,325%**	**0,021**

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-15478**			**100,000%**	
K - Means EM	**-16114**	**-17116**	**5,00%**	**98,425%**	**1,340**
Random EM	**-15476**	**-16525**	**15,00%**	**99,525%**	**0,835**

Πίνακας 5.10 M=8 Θόρυβος 50%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-110168**			**99,500%**	
K - Means EM	**-110546**	**-111549**	**10,00%**	**95,450%**	**0,024**
Random EM	**-110168**	**-111527**	**45,00%**	**97,088%**	**0,026**

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-17873**			**100,000%**	
K - Means EM	**-18485**	**-19493**	**10,00%**	**98,800%**	**1,642**
Random EM	**-17874**	**-19411**	**45,00%**	**99,100%**	**1,492**

Πίνακας 5.11 M=8 Θόρυβος 60%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-115549**			**95,375%**	
K - Means EM	**-115556**	**-116049**	**10,00%**	**91,581%**	**0,023**
Random EM	**-115370**	**-115612**	**60,00%**	**94,200%**	**0,016**

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-21177**			**96,500%**	
K - Means EM	**-21027**	**-21989**	**5,00%**	**96,825%**	**2,184**
Random EM	**-20549**	**-21202**	**10,00%**	**99,825**	**0,591**

Πίνακας 5.12 M=8 Θόρυβος 70%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-117816**			**66,375%**	
K - Means EM	**-117814**	**-117899**	**5**	**62,219%**	**0,031**
Random EM	**-117796**	**-117864**	**5,00%**	**61,713%**	**0,029**

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-24781**			**100,000%**	
K - Means EM	**-24596**	**-25228**	**5,00%**	**96,675%**	**3,246**
Random EM	**-24480**	**-25243**	**5,00%**	**97,450%**	**1,959**

Για M=10

Πίνακας 5.13 M=10 Θόρυβος 40%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-133319**			**100,000%**	
K - Means EM	**-133319**	**-136287**	**10%**	**95,810%**	**0,030**
Random EM	**-133319**	**-135444**	**25%**	**98,180%**	**0,017**

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-20867**			**100,000%**	
K - Means EM	**-20867**	**-23082**	**10%**	**9916,000%**	**0,889**
Random EM	**-20867**	**-22476**	**25%**	**99,220%**	**1,120**

Πίνακας 5.14 M=10 Θόρυβος 50%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-143092**			**98,500%**	
K - Means EM	**-142638**	**-144222**	**5,00%**	**95,705%**	**0,020**
Random EM	**-142638**	**-143726**	**50,00%**	**97,270%**	**0,021**

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-24626**			**100,000%**	
K - Means EM	**-23941**	**-25764**	**5,00%**	**98,140%**	**2,126**
Random EM	**-23940**	**-25415**	**20,00%**	**99,600%**	**0,779**

Πίνακας 5.15 M=10 Θόρυβος 60%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-149511**			**89,700%**	
K - Means EM	**-149176**	**-149794**	**5,00%**	**87,740%**	**0,020**
Random EM	**-149176**	**-149520**	**30,00%**	**88,750%**	**0,022**

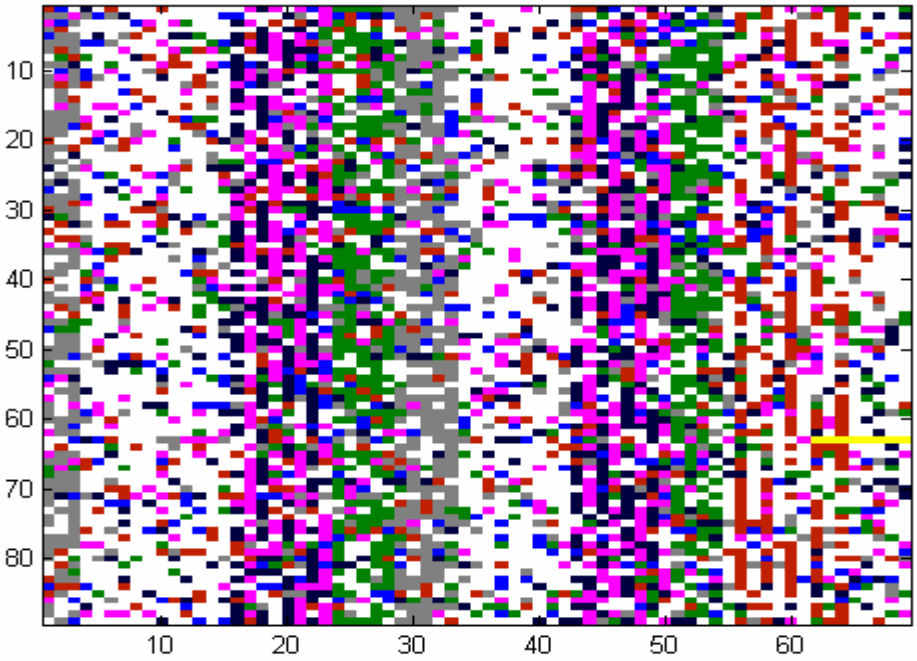
Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-28714**			**98,400%**	
K - Means EM	**-27639**	**-29293**	**5,00%**	**99,160%**	**1,682**
Random EM	**-27632**	**-28759**	**5,00%**	**99,760**	**0,704**

Πίνακας 5.16 M=10 Θόρυβος 70%

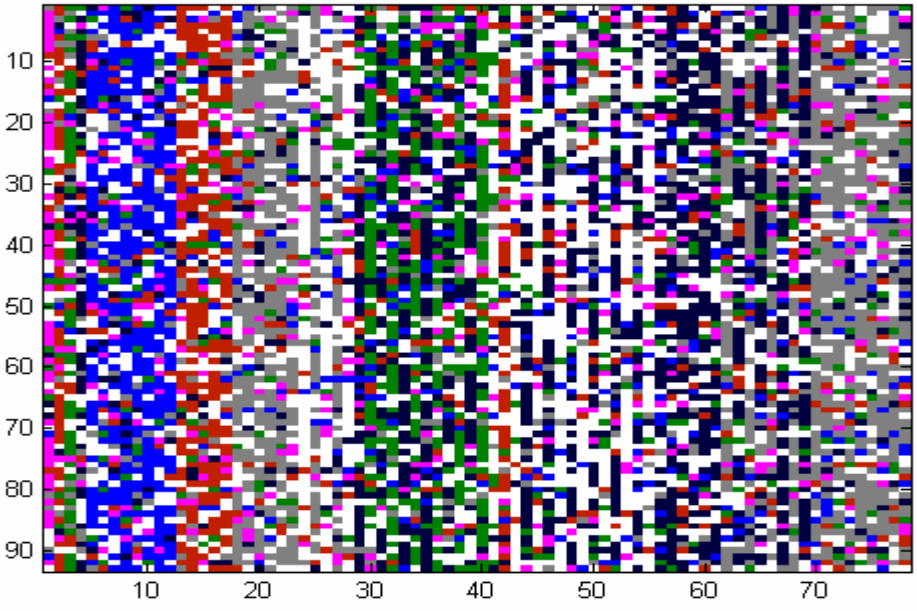
Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-152283**			**58,300%**	
K - Means EM	**-152233**	**-152330**	**5,00%**	**48,250%**	**0,031**
Random EM	**-152250**	**-152303**	**5,00%**	**48,500%**	**0,031**

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-33654**			**91,200%**	
K - Means EM	**-32710**	**-33622**	**5,00%**	**95,480%**	**3,049**
Random EM	**-32573**	**-33516**	**5,00%**	**95,220%**	**2,944**

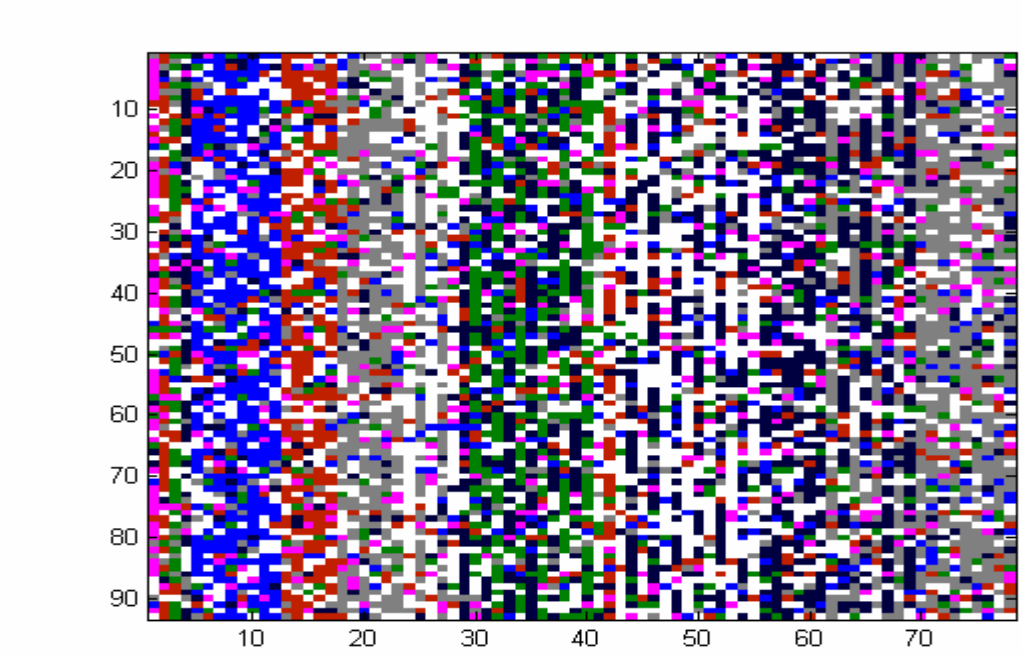
Οπτικοποίηση για M=10 και θόρυβο 50%



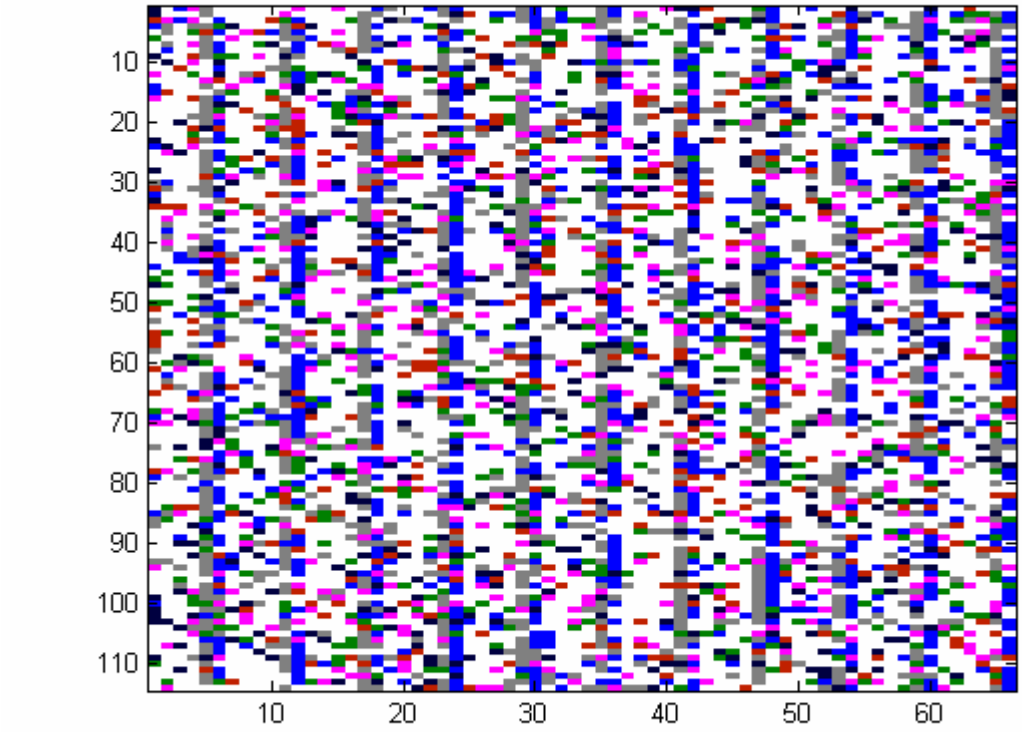
Σχήμα 5.1 Ομάδα 1



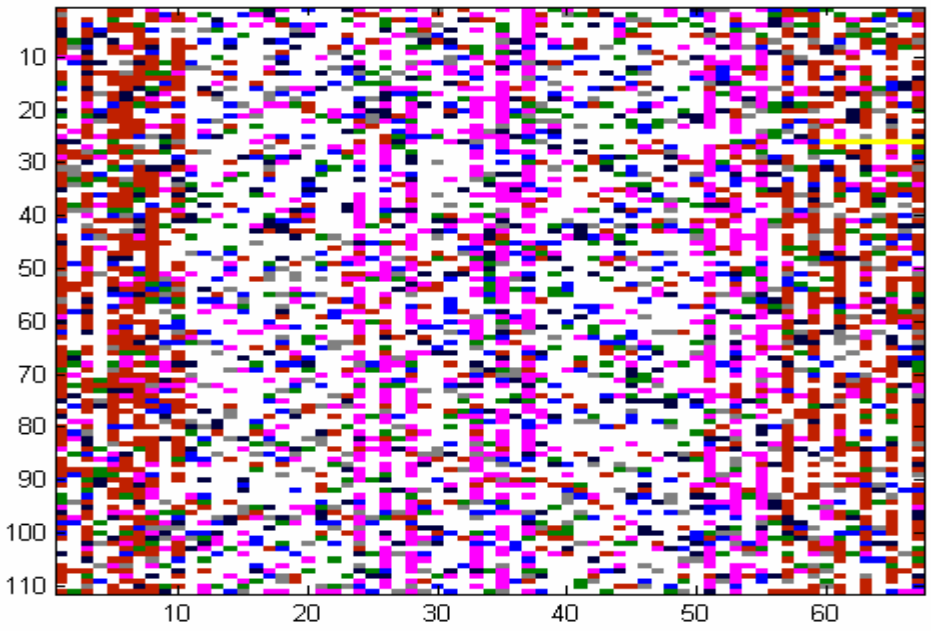
Σχήμα 5.2 Ομάδα 2



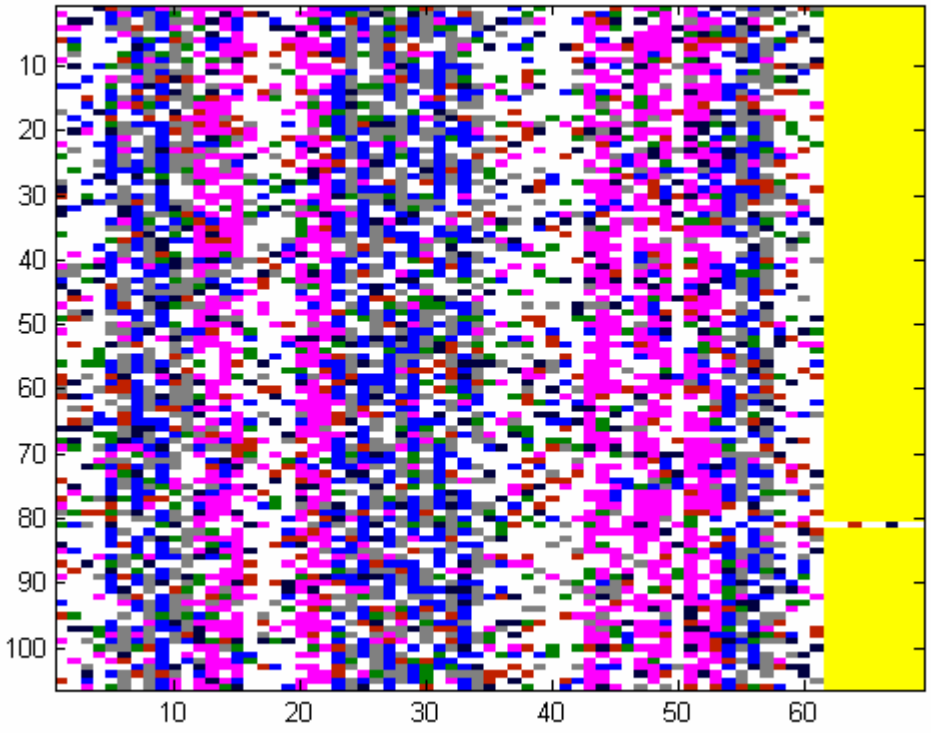
Σχήμα 5.3 Ομάδα 3



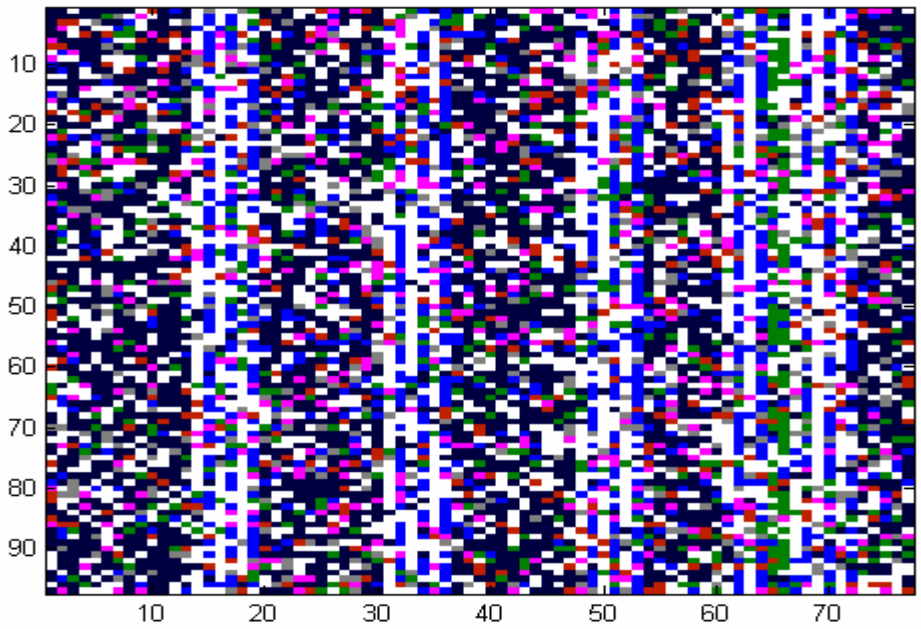
Σχήμα 5.4 Ομάδα 4



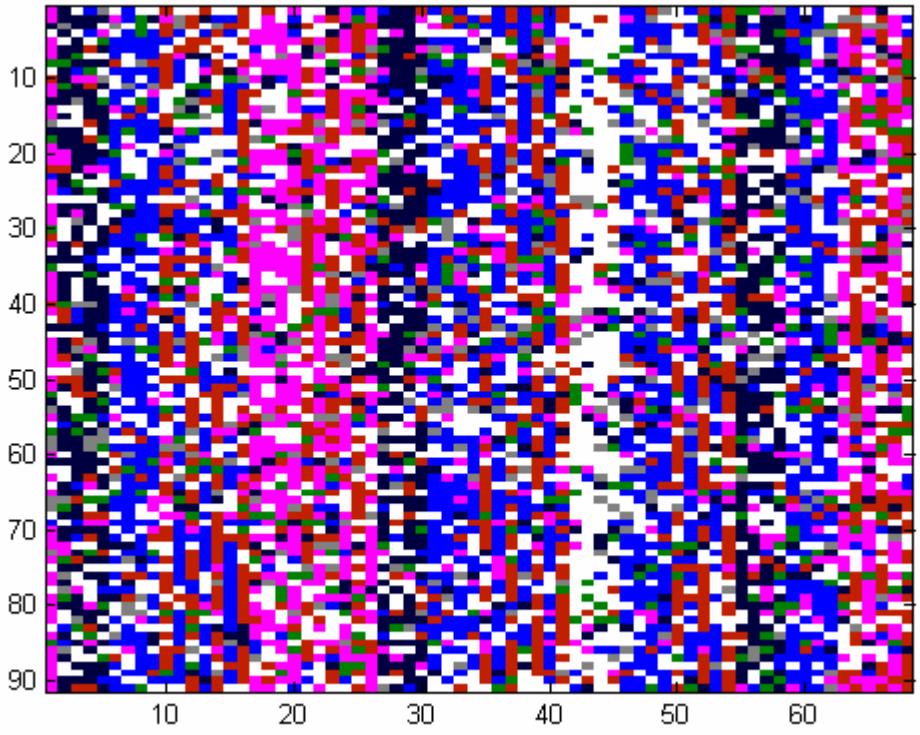
Σχήμα 5.5 Ομάδα 5



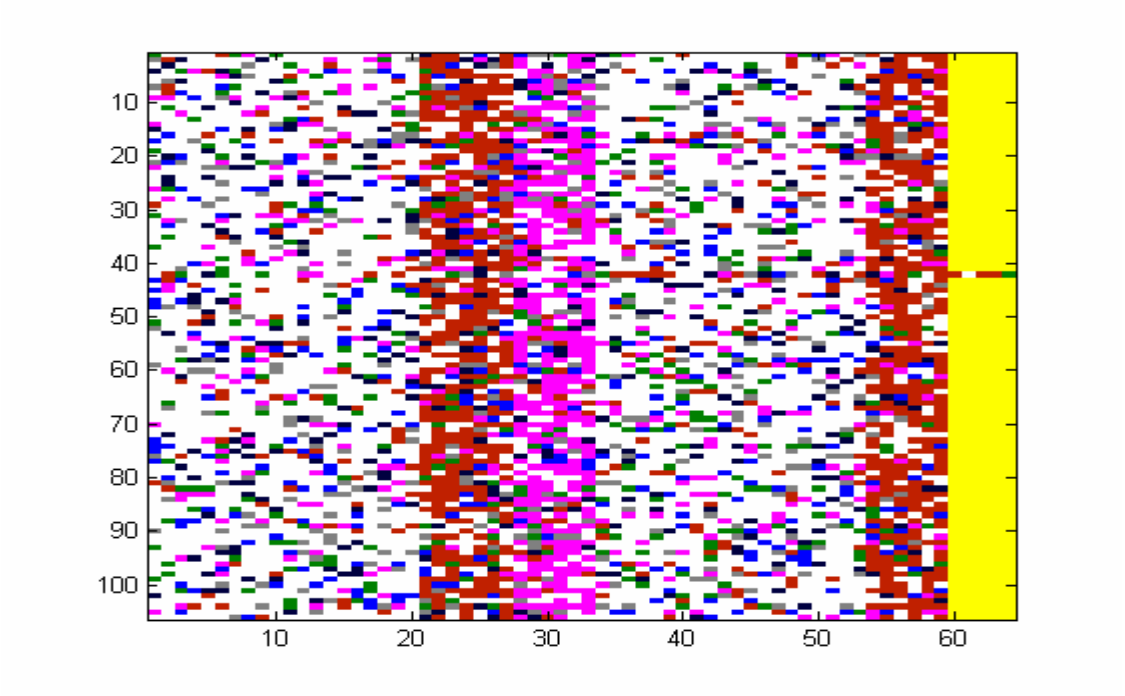
Σχήμα 5.6 Ομάδα 6



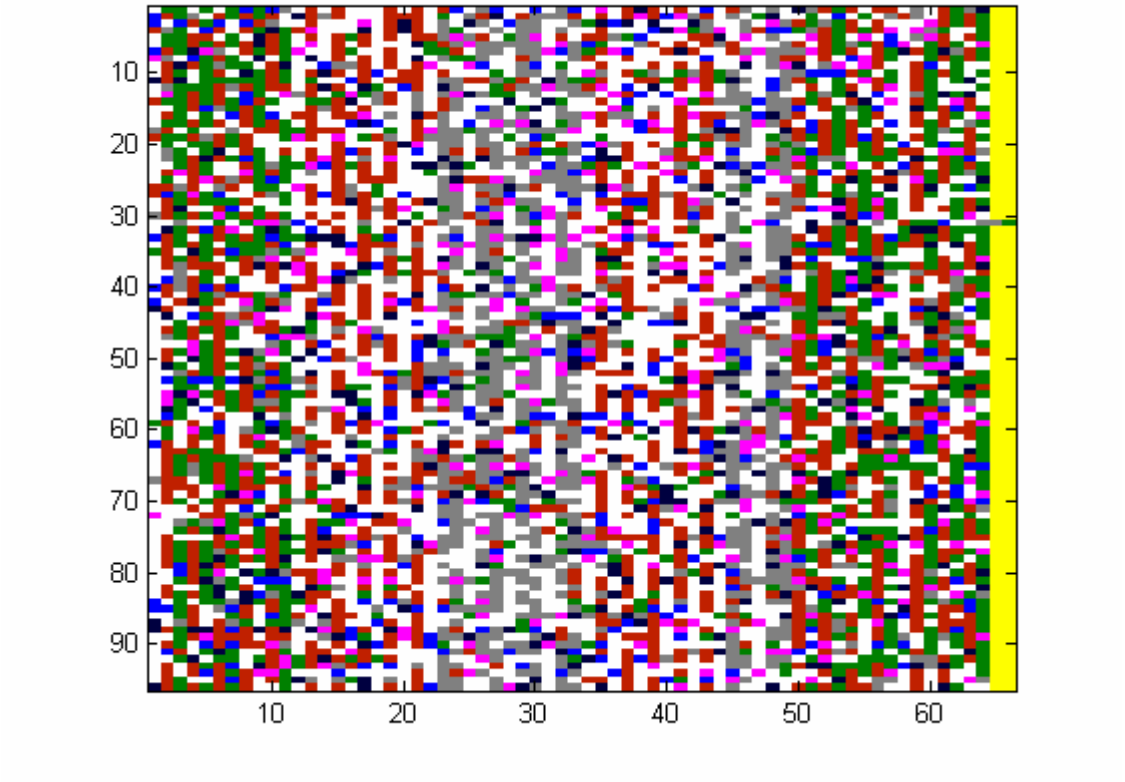
Σχήμα 5.7 Ομάδα 7



Σχήμα 5.8 Ομάδα 8



Σχήμα 5.9 Ομάδα 9



Σχήμα 5.10 Ομάδα 10

5.3.2. Τεχνητό σύνολο δεδομένων 2

Εδώ τα δεδομένα κατασκευάζονται από τυχαία πρότυπα μήκους 30 συμβόλων που εμφυτεύονται σε ακολουθίες τυχαίου μήκους από 0 έως 20 συμβόλων. Για κάθε αριθμό κατηγοριών M = { 4, 6, 8, 10} έχουμε σύνολα ακολουθιών που κατασκευάστηκαν με διαφορετικό ποσοστό θορύβου P_m={20%, 40%, 60%, 80%}.

Για M=4

Πίνακας 5.17 M=4 Θόρυβος 20%

Σύνολο Εκπαίδευσης	**Πιθανοφάνεια**			**Ποσοστό Κατηγοριοποίησης**	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-64428**			**99,750%**	
K - Means EM	**-64428**	**-65228**	**95,00%**	**99,350%**	**0,018**
Random EM	**-64428**	**-65252**	**95,00%**	**99,338%**	**0,018**

Σύνολο Ελέγχου	**Πιθανοφάνεια**			**Ποσοστό Κατηγοριοποίησης**	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-14594**			**100,000%**	
K - Means EM	**-14594**	**-15029**	**95,00%**	**99,950%**	**0,224**
Random EM	**-14594**	**-15177**	**95,00%**	**99,750%**	**1,118**

Πίνακας 5.18 M=4 Θόρυβος 40%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-67370			92,500%	
K - Means EM	-67370	-67467	65,00%	89,988%	0,042
Random EM	-67370	-67471	75,00%	89,988%	0,033

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-15566			100,000%	
K - Means EM	-15556	-15859	5,00%	99,050%	3,364
Random EM	-15555	-15865	5,00%	99,600%	1,789

Πίνακας 5.19 M=4 Θόρυβος 60%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-68723			84,250%	
K - Means EM	-68669	-68710	5,00%	45,770%	0,060
Random EM	-68680	-68719	5,00%	42,938%	0,056

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-16913			93,000%	
K - Means EM	-16645	-16866	5,00%	85,700%	6,457
Random EM	-16693	-16872	5,00%	82,150%	7,631

Πίνακας 5.20 M=4 Θόρυβος 80%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-68852			87,000%	
K - Means EM	-68819	-68866	5,00%	39,300%	0,062
Random EM	-68823	-68861	10,00%	42,400%	0,093

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-17288			100,000%	
K - Means EM	-17312	-17380	5,00%	73,300%	11,649
Random EM	-17318	-17394	5,00%	74,850%	12,193

Για M=6

Πίνακας 5.21 M=6 Θόρυβος 20%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-97073			100,000%	
K - Means EM	-97073	-97945	85,00%	9903,333%	0,024415
Random EM	-97073	-97991	45,00%	9708,333%	0,032248

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-22174			100,000%	
K - Means EM	-22174	-22789	85,00%	99,267%	2,348572
Random EM	-22174	-22914	45,00%	98,833%	1,87161

Πίνακας 5.22 M=4 Θόρυβος 40%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-102081**			**80,333%**	
K - Means EM	**-102081**	**-102176**	**65,00%**	**80,092%**	**0,022001**
Random EM	**-102081**	**-102222**	**50,00%**	**78,350%**	**0,032421**

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-23792**			**100,000%**	
K - Means EM	**-23780**	**-24023**	**10,00%**	**99,63%**	**0,929692**
Random EM	**-23780**	**-24147**	**5,00%**	**98,76%**	**2,57995**

Πίνακας 5.23 M=4 Θόρυβος 60%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-103046**			**76,000%**	
K - Means EM	**-103029**	**-103069**	**10,00%**	**39,367%**	**0,04514**
Random EM	**-103021**	**-103070**	**5,00%**	**36,133%**	**0,041915**

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-25630**			**96,66%**	
K - Means EM	**-25394**	**-25672**	**10,00%**	**78,76%**	**6,202339**
Random EM	**-25503**	**-25652**	**10,00%**	**78,43%**	**7,226575**

Πίνακας 5.24 M=4 Θόρυβος 80%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-102926			77,667%	
K - Means EM	-102915	-102963	0,00%	34,808%	0,066386
Random EM	-102904	-102964	0,00%	32,692%	0,064228

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-25833			98,000%	
K - Means EM	-25777	-25882	0,00%	75,100%	5,338923
Random EM	-25771	-25904	0,00%	69,200%	11,43423

Για M=8

Πίνακας 5.25 M=8 Θόρυβος 20%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-128839			99,875%	
K - Means EM	-128839	-130525	60,00%	97,981%	0,028288
Random EM	-128839	-129843	50,00%	98,013%	0,023999

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-29471			100,000%	
K - Means EM	-29471	-30269	60,00%	98,650%	2,373649
Random EM	-29471	-30179	50,00%	98,900%	1,909808

Πίνακας 5.26 M=8 Θόρυβος 40%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-136329			77,250%	
K - Means EM	-136253	-136422	25,00%	77,319%	0,018662
Random EM	-136253	-136349	20,00%	76,956%	0,018989

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-31824			100,000%	
K - Means EM	-31518	-32002	5,00%	99,375%	1,571749
Random EM	-31520	-31826	20,00%	99,425%	1,680186

Πίνακας 5.27 M=8 Θόρυβος 60%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-137266			77,250%	
K - Means EM	-137261	-137328	5,00%	34,188%	0,070239
Random EM	-137257	-137321	5,00%	32,525%	0,075052

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-34199			91,000%	
K - Means EM	-33891	-34184	5,00%	68,600%	6,556636
Random EM	-33990	-34200	5,00%	67,650%	6,343459

Πίνακας 5.28 M=8 Θόρυβος 80%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-137505			69,375%	
K - Means EM	-137492	-137569	5,00%	33,331%	0,065367
Random EM	-137477	-137569	5,00%	26,894%	0,061583

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-34452			91,500%	
K - Means EM	-34454	-34566	5,00%	68,350%	7,744438
Random EM	-34460	-34588	5,00%	62,600%	7,107446

Για M=10

Πίνακας 5.29 M=10 Θόρυβος 20%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-162709			99,400%	
K - Means EM	-162709	-163500	30,00%	97,855%	0,016388
Random EM	-162709	-164338	40,00%	97,585%	0,018667

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-37333			100,000%	
K - Means EM	-37333	-37782	30,00%	99,280%	1,500035
Random EM	-37333	-38220	40,00%	98,820%	1,926983

Πίνακας 5.30 M=10 Θόρυβος 40%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-170144**			**73,900%**	
K - Means EM	**-169796**	**-170089**	**25,00%**	**77,230%**	**0,021912**
Random EM	**-169796**	**-170080**	**5,00%**	**76,550%**	**0,023596**

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-40132**			**100,000%**	
K - Means EM	**-39163**	**-39908**	**5,00%**	**99,360%**	**1,418079**
Random EM	**-39160**	**-39877**	**5,00%**	**99,360%**	**1,049511**

Πίνακας 5.31 M=10 Θόρυβος 60%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-171720**			**75,000%**	
K - Means EM	**-171692**	**-171772**	**5,00%**	**33,245%**	**0,065869**
Random EM	**-171686**	**-171761**	**5,00%**	**28,275%**	**0,058564**

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-42736**			**95,200%**	
K - Means EM	**-42577**	**-42714**	**5,00%**	**69,080%**	**7,327641**
Random EM	**-42601**	**-42795**	**5,00%**	**67,720%**	**6,221322**

Πίνακας 5.32 M=10 Θόρυβος 80%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-171679,259**			**76,500%**	
K - Means EM	**-171641**	**-171704**	**5,00%**	**33,685%**	**0,078641**
Random EM	**-171631**	**-171705**	**5,00%**	**26,420%**	**0,063044**

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-43439**			**96,800%**	
K - Means EM	**-43397**	**-43560**	**5,00%**	**69,960%**	**9,395094**
Random EM	**-43460**	**-43617**	**5,00%**	**62,800%**	**8,772205**

5.3.3. Τεχνητό σύνολο δεδομένων 3

Εδώ τα δεδομένα κατασκευάζονται από έτοιμα πρότυπα που εμφυτεύονται σε ακολουθίες τυχαίου μήκους από 0 έως 20 συμβόλων. Για κάθε αριθμό κατηγοριών M = { 4, 6, 8, 10 } έχουμε σύνολα ακολουθιών που κατασκευάστηκαν με διαφορετικό ποσοστό θορύβου P_m={20%, 40%, 60%, 80%}.

Για M=4

Πίνακας 5.33 M=4 Θόρυβος 20%

Σύνολο Εκπαίδευσης	**Πιθανοφάνεια**			**Ποσοστό Κατηγοριοποίησης**	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-64580**			**97,500%**	
K - Means EM	**-64633**	**-64633**	**100,00%**	**100,000%**	**0**
Random EM	**-64633**	**-65642**	**80,00%**	**98,350%**	**0,034796**

Σύνολο Ελέγχου	**Πιθανοφάνεια**			**Ποσοστό Κατηγοριοποίησης**	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-14700**			**100,000%**	
K - Means EM	**-14700**	**-14700**	**100,00%**	**100,000%**	**0**
Random EM	**-14700**	**-15263**	**80,00%**	**99,100%**	**2,789076**

Πίνακας 5.34 M=4 Θόρυβος 40%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-28862			93,000%	
K - Means EM	-28938	-29128	95,00%	95,188%	0,026937
Random EM	-28938	-29146	85,00%	94,738%	0,027356

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-6182			100,000%	
K - Means EM	-6182	-6358	95,00%	99,450%	2,459675
Random EM	-6182	-6350	40,00%	99,400%	1,957442

Πίνακας 5.35 M=4 Θόρυβος 60%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-30663			76,500%	
K - Means EM	-30701	-30733	20,00%	56,100%	0,050272
Random EM	-30694	-30731	5,00%	54,575%	0,053077

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-7266			73,000%	
K - Means EM	-7016	-7201	5,00%	86,050%	8,432238
Random EM	-7003	-7222	5,00%	84,350%	5,479867

Πίνακας 5.36 M=4 Θόρυβος 80%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-30681**			**80,250%**	
K - Means EM	**-30659**	**-30675**	**10,00%**	**38,163%**	**0,03305**
Random EM	**-30657**	**-30693**	**10,00%**	**40,163%**	**0,059357**

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-7617**			**65,000%**	
K - Means EM	**-7550**	**-7677**	**5,00%**	**74,900%**	**7,503683**
Random EM	**-7509**	**-7734**	**5,00%**	**72,550%**	**8,976431**

Για M=6

Πίνακας 5.37 M=6 Θόρυβος 20%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-40347**			**97,000%**	
K - Means EM	**-40424**	**-40869**	**80,00%**	**97,692%**	**0,026282**
Random EM	**-40424**	**-40852**	**70,00%**	**97,783%**	**0,0277**

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-8727**			**100,000%**	
K - Means EM	**-8727**	**-8954**	**60,00%**	**99,300%**	**1,324091**
Random EM	**-8727**	**-8966**	**50,00%**	**99,133%**	**1,886858**

Πίνακας 5.38 M=6 Θόρυβος 40%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-44133**			**90,833%**	
K - Means EM	**-44221**	**-44221**	**85,00%**	**92,900%**	**0,021209**
Random EM	**-44221**	**-44408**	**85,00%**	**92,917%**	**0,021285**

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-9212**			**100,000%**	
K - Means EM	**-9212**	**-9446**	**20,00%**	**99,867%**	**0,463902**
Random EM	**-9213**	**-9449**	**5,00%**	**99,767%**	**0,658725**

Πίνακας 5.39 M=6 Θόρυβος 60%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-45848**			**55,333%**	
K - Means EM	**-45873**	**-45938**	**5,00%**	**44,758%**	**0,021687**
Random EM	**-45868**	**-45941**	**5,00%**	**43,233%**	**0,047361**

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-10812**			**70,000%**	
K - Means EM	**-10565**	**-10806**	**5,00%**	**80,600%**	**7,117724**
Random EM	**-10571**	**-10866**	**5,00%**	**79,900%**	**9,827221**

Πίνακας 5.40 M=6 Θόρυβος 80%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-46805			87,333%	
K - Means EM	-46530	-46566	5,00%	28,950%	0,032647
Random EM	-46533	-46576	5,00%	30,142%	0,026365

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-11674			38,000%	
K - Means EM	-11514	-11702	5,00%	62,467%	6,209218
Random EM	-11443	-11641	5,00%	64,800%	5,448751

Για M=8

Πίνακας 5.41 M=8 Θόρυβος 20%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-54804			97,500%	
K - Means EM	-54883	-55841	5,00%	95,775%	0,026472
Random EM	-54883	-55488	40,00%	97,794%	0,021587

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-11738,938			100,000%	
K - Means EM	-11739	-12236	5,00%	98,175%	1,94175
Random EM	-11739	-12034	35,00%	99,075%	2,085508

Πίνακας 5.42 M=8 Θόρυβος 40%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-60491**			**84,625%**	
K - Means EM	**-60458**	**-60766**	**40,00%**	**85,769%**	**0,030523**
Random EM	**-60458**	**-60637**	**60,00%**	**86,744%**	**0,021481**

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-12818**			**97,000%**	
K - Means EM	**-12695**	**-13060**	**20,00%**	**97,800%**	**2,773939**
Random EM	**-12694**	**-12968**	**5,00%**	**98,700%**	**2,154555**

Πίνακας 5.43 M=8 Θόρυβος 60%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-63151**			**62,750%**	
K - Means EM	**-63104**	**-63172**	**5,00%**	**41,744%**	**0,0301**
Random EM	**-63099**	**-63161**	**10,00%**	**40,194%**	**0,034251**

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-14854**			**67,500%**	
K - Means EM	**-14683**	**-14866**	**5,00%**	**73,325%**	**5,664467**
Random EM	**-14679**	**-14929**	**5,00%**	**74,975%**	**4,848969**

Πίνακας 5.44 M=8 Θόρυβος 80%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-63408			99,875%	
K - Means EM	--63355	-63442	25,00%	96,888%	0,023612
Random EM	-63361	-63415	35,00%	97,356%	0,02114

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-15804			100,000%	
K - Means EM	-15662	-15823	15,00%	98,450%	1,669384
Random EM	-15594	-15893	30,00%	99,000%	1,450953

Για M=10

Πίνακας 5.45 M=10 Θόρυβος 20%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-72892			98,900%	
K - Means EM	-72893	-74220	5,00%	94,655%	0,025792
Random EM	-72893	-73775	20,00%	96,165%	0,022915

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-15706			100,000%	
K - Means EM	-15707	-16419	5,00%	97,780%	2,234561
Random EM	-15706	-15995	15,00%	98,420%	1,970199

Πίνακας 5.46 M=10 Θόρυβος 40%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-80671			83,800%	
K - Means EM	-80605	-80961	10,00%	80,990%	0,028923
Random EM	-80605	-80853	15,00%	81,890%	0,019241

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-17131			100,000%	
K - Means EM	-16981	-17535	5,00%	97,880%	2,469519
Random EM	-16978	-17350	5,00%	98,580%	1,227578

Πίνακας 5.47 M=10 Θόρυβος 60%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-82578			58,200%	
K - Means EM	-82535	-82739	5,00%	34,605%	0,035654
Random EM	-82521	-82589	10,00%	31,705%	0,033284

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-20092			72,400%	
K - Means EM	-19867	-20136	5,00%	72,720%	5,685957
Random EM	-19863	-20372	5,00%	69,920%	6,092066

Πίνακας 5.48 M=10 Θόρυβος 80%

Σύνολο Εκπαίδευσης	**Πιθανοφάνεια**			**Ποσοστό Κατηγοριοποίησης**	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-83004**			**73,000%**	
K - Means EM	**-82938**	**-83043**	**5,00%**	**24,145%**	**0,056225**
Random EM	**-82929**	**-82991**	**5,00%**	**23,950%**	**0,044938**

Σύνολο Ελέγχου	**Πιθανοφάνεια**			**Ποσοστό Κατηγοριοποίησης**	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-20757**			**70,400%**	
K - Means EM	**-20596**	**-20791**	**5,00%**	**56,200%**	**6,517991**
Random EM	**-20636**	**-20803**	**5,00%**	**56,260%**	**6,961276**

5.3.4. Τεχνητό σύνολο δεδομένων 4

Εδώ τα δεδομένα κατασκευάζονται από έτοιμα πρότυπα που εμφυτεύονται σε ακολουθίες τυχαίου μήκους από 0 έως 40 συμβόλων, ενώ προηγουμένως το μέγιστο μήκος της πρόσθετης ακολουθίας ήταν 20. Αυτό περιμένουμε να προσδώσει μεγαλύτερη ομοιομορφία στις ακολουθίες και να δυσχεράνει την εύρεση των προτύπων. Για κάθε αριθμό κατηγοριών M = { 4, 6, 8, 10 } έχουμε σύνολα ακολουθιών που κατασκευάστηκαν με διαφορετικό ποσοστό θορύβου P_m={20%, 40%, 60%, 80%}.

Για M=4

Πίνακας 5.49 M=4 Θόρυβος 20%

Σύνολο Εκπαίδευσης	**Πιθανοφάνεια**			**Ποσοστό Κατηγοριοποίησης**	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-40669**			**99,750%**	
K - Means EM	**-40669**	**-41193**	**95,00%**	**99,488%**	**0,016493**
Random EM	**-40669**	**-41109**	**75,00%**	**98,200%**	**0,031608**

Σύνολο Ελέγχου	**Πιθανοφάνεια**			**Ποσοστό Κατηγοριοποίησης**	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-9414**			**100,000%**	
K - Means EM	**-9414**	**-9629**	**95,00%**	**99,750%**	**1,118034**
Random EM	**-9414**	**-9611**	**60,00%**	**99,150%**	**2,345769**

Πίνακας 5.50 M=4 Θόρυβος 40%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-43585**			**95,000%**	
K - Means EM	**-43585**	**-43768**	**85,00%**	**93,913%**	**0,024147**
Random EM	**-43585**	**-43757**	**90,00%**	**94,275%**	**0,01938**

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-9835**			**100,000%**	
K - Means EM	**-9834**	**-10034**	**10,00%**	**99,900%**	**0,307794**
Random EM	**-9834**	**-10049**	**40,00%**	**99,800%**	**0,894427**

Πίνακας 5.51 M=4 Θόρυβος 60%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-44115**			**80,500%**	
K - Means EM	**-44070**	**-44174**	**5,00%**	**54,313%**	**0,071676**
Random EM	**-44072**	**-44106**	**5,00%**	**50,888%**	**0,035051**

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-10657**			**74,000%**	
K - Means EM	**-10510**	**-10653**	**5,00%**	**86,400%**	**9,810843**
Random EM	**-10486**	**-10673**	**5,00%**	**86,200%**	**8,420526**

Πίνακας 5.52 M=4 Θόρυβος 80%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-44813**			**84,000%**	
K - Means EM	**-44806**	**-44836**	**10,00%**	**39,300%**	**0,050677**
Random EM	**-44794**	**-44845**	**5,00%**	**43,138%**	**0,086532**

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-11309**			**87,000%**	
K - Means EM	**-11227**	**-11306**	**5,00%**	**70,300%**	**6,751998**
Random EM	**-11222**	**-11299**	**5,00%**	**69,700%**	**8,360937**

Για M=6

Πίνακας 5.53 M=6 Θόρυβος 20%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-61818**			**99,500%**	
K - Means EM	**-61819**	**-62297**	**70,00%**	**98,167%**	**0,025809**
Random EM	**-61819**	**-62271**	**45,00%**	**96,167%**	**0,035343**

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-14445**			**100,000%**	
K - Means EM	**-14445**	**-14772**	**75,00%**	**99,167%**	**2,050531**
Random EM	**-14445**	**-14651**	**45,00%**	**97,167%**	**3,631409**

Πίνακας 5.54 M=6 Θόρυβος 40%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-66206			86,833%	
K - Means EM	-66145	-66277	65,00%	87,517%	0,021743
Random EM	-66145	-66324	60,00%	86,700%	0,029278

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-15255			89,333%	
K - Means EM	-15087	-15295	30,00%	98,867%	2,217422
Random EM	-15086	-15297	5,00%	99,400%	1,100771

Πίνακας 5.55 M=6 Θόρυβος 60%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-67468			71,500%	
K - Means EM	-67443	-67492	10,00%	41,033%	0,03852
Random EM	-67429	-67496	5,00%	39,275%	0,038137

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-16264			77,333%	
K - Means EM	-15988	-16229	5,00%	74,967%	7,236926
Random EM	-16031	-16257	5,00%	73,033%	6,142461

Πίνακας 5.56 M=6 Θόρυβος 80%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-66596**			**64,833%**	
K - Means EM	**-66570**	**-66616**	**5,00%**	**31,533%**	**0,054202**
Random EM	**-66568**	**-66630**	**10,00%**	**29,000%**	**0,035326**

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-16877**			**65,333%**	
K - Means EM	**-16808**	**-16897**	**5,00%**	**58,733%**	**6,470446**
Random EM	**-16757**	**-16890**	**5,00%**	**61,167%**	**7,238219**

Για M=8

Πίνακας 5.57 M=8 Θόρυβος 20%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-84678**			**99,500%**	
K - Means EM	**-84678**	**-85474**	**40,00%**	**96,431%**	**0,029743**
Random EM	**-84678**	**-85523**	**35,00%**	**96,363%**	**0,028814**

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-19387**			**100,000%**	
K - Means EM	**-19387**	**-19822**	**40,00%**	**97,575%**	**2,369072**
Random EM	**-19387**	**-19824**	**35,00%**	**98,375%**	**1,911702**

Πίνακας 5.58 M=8 Θόρυβος 40%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-89156			79,375%	
K - Means EM	-89049	-89210	45,00%	81,869%	0,041576
Random EM	-89049	-89309	10,00%	80,306%	0,035614

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-20713			96,000%	
K - Means EM	-20410	-20848	5,00%	98,275%	2,672841
Random EM	-20412	-20903	5,00%	97,375%	2,443439

Πίνακας 5.59 M=8 Θόρυβος 60%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-90579			68,375%	
K - Means EM	-90521	-90584	10,00%	33,025%	0,033143
Random EM	-90518	-90574	5,00%	32,519%	0,02565

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-22467			78,500%	
K - Means EM	-22323	-22578	5,00%	63,675%	5,32935
Random EM	-22377	-22569	5,00%	65,775%	5,378992

Πίνακας 5.60 M=8 Θόρυβος 80%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-90265			74,125%	
K - Means EM	-90220	-90294	5,00%	28,063%	0,057458
Random EM	-90227	-90293	5,00%	27,300%	0,050701

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-22769			81,500%	
K - Means EM	-22670	-22756	5,00%	54,500%	5,498804
Random EM	-22693	-22786	5,00%	53,225%	5,420417

Για M=10

Πίνακας 5.61 M=10 Θόρυβος 20%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-110098			98,600%	
K - Means EM	-110099	-111345	5,00%	92,695%	0,027676
Random EM	-110099	-111112	30,00%	95,525%	0,027737

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-25231			100,000%	
K - Means EM	-25231	-25887	5,00%	96,760%	2,828129
Random EM	-25231	-25620	30,00%	98,260%	2,214878

Πίνακας 5.62 M=10 Θόρυβος 40%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-116870**			**71,600%**	
K - Means EM	**-116660**	**-116856**	**5,00%**	**70,980%**	**0,034291**
Random EM	**-116639**	**-116779**	**10,00%**	**72,060%**	**0,025257**

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-27024**			**92,000%**	
K - Means EM	**-26606**	**-27184**	**5,00%**	**95,940%**	**3,208607**
Random EM	**-26495**	**-26837**	**5,00%**	**97,000%**	**2,708272**

Πίνακας 5.63 M=10 Θόρυβος 60%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-118438**			**69,500%**	
K - Means EM	**-118392**	**-118444**	**5,00%**	**26,885%**	**0,02619**
Random EM	**-118388**	**-118472**	**5,00%**	**27,535%**	**0,030595**

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-28655**			**69,600%**	
K - Means EM	**-28424**	**-28679**	**5,00%**	**62,600%**	**5,876626**
Random EM	**-28486**	**-28707**	**5,00%**	**58,640%**	**5,561371**

Πίνακας 5.64 M=10 Θόρυβος 80%

Σύνολο Εκπαίδευσης	**Πιθανοφάνεια**			**Ποσοστό Κατηγοριοποίησης**	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-118260**			**68,100%**	
K - Means EM	**-118166**	**-118246**	**5,00%**	**22,850%**	**0,063197**
Random EM	**-118155**	**-118237**	**5,00%**	**25,070%**	**0,045088**

Σύνολο Ελέγχου	**Πιθανοφάνεια**			**Ποσοστό Κατηγοριοποίησης**	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-29700**			**75,600%**	
K - Means EM	**-29487**	**-29718**	**5,00%**	**48,400%**	**5,11921**
Random EM	**-29581**	**-29693**	**5,00%**	**47,120%**	**5,085542**

5.3.5. Τεχνητό σύνολο δεδομένων 5

Εδώ τα σύνολα δεδομένων κατασκευάζονται με τον πρώτο τρόπο που αναφέραμε (δειγματοληψία από έτοιμες συμβολοσειρές). Για κάθε αριθμό κατηγοριών M = { 4, 6, 8 } έχουμε σύνολα ακολουθιών που κατασκευάστηκαν με διαφορετικό ποσοστό θορύβου P_m={ 40%, 60%, 80%}. Η διαφορά με τα προηγούμενα σύνολα είναι ότι έχουμε σημαντικές διαφορές στον αριθμό ακολουθιών που ανήκουν σε κάθε κατηγορία:

Για M=4 το πλήθος των ακολουθιών για κάθε κατηγορία είναι: [25 50 75 100].
Για M=6 το πλήθος των ακολουθιών για κάθε κατηγορία είναι: [25 100 200 50 50 200].
Για M=8 το πλήθος των ακολουθιών για κάθε κατηγορία είναι: [25 50 75 100 200 150 125 100].

Για M=4

Πίνακας 5.65 M=4 Θόρυβος 40%

Σύνολο Εκπαίδευσης	**Πιθανοφάνεια**			**Ποσοστό Κατηγοριοποίησης**	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-30962**			**100,000%**	
K - Means EM	**-30962**	**-30962**	**100,00%**	**100,000%**	**0**
Random EM	**-30962**	**-31362**	**75,00%**	**98,340%**	**0,036205**

Σύνολο Ελέγχου	**Πιθανοφάνεια**			**Ποσοστό Κατηγοριοποίησης**	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-4854**			**100,000%**	
K - Means EM	**-4852**	**-4852**	**20,00%**	**100,000%**	**0**
Random EM	**-4852**	**-5127**	**55,00%**	**99,426%**	**2,210637**

Πίνακας 5.66 M=4 Θόρυβος 60%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-34605**			**92,000%**	
K - Means EM	**-34561**	**-34639**	**20,00%**	**91,260%**	**0,059489**
Random EM	**-34560**	**-34626**	**15,00%**	**93,060%**	**0,0508**

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-6374**			**100,000%**	
K - Means EM	**-6329**	**-6558**	**10,00%**	**99,590%**	**1,289176**
Random EM	**-6332**	**-6532**	**5,00%**	**99,836%**	**0,50458**

Για ποσοστό θορύβου 80%:

Πίνακας 5.67 M=4 Θόρυβος 80%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-35729**			**81,600%**	
K - Means EM	**-35689**	**-35731**	**5,00%**	**41,260%**	**0,046126**
Random EM	**-35696**	**-35729**	**5,00%**	**41,520%**	**0,050037**

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-8560**			**96,721%**	
K - Means EM	**-8147**	**-8520**	**5,00%**	**90,000%**	**8,866509**
Random EM	**-8182**	**-8480**	**5,00%**	**86,066%**	**8,162136**

Για M=6

Πίνακας 5.68 M=6 Θόρυβος 40%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-76894			100,000%	
K - Means EM	-76894	-78018	50,00%	98,024%	0,037488
Random EM	-76894	-78129	25,00%	94,744%	0,044969

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-11930			100,000%	
K - Means EM	-11929	-13005	5,00%	99,419%	0,956929
Random EM	-11930	-12624	25,00%	98,710%	2,386592

Πίνακας 5.69 M=6 Θόρυβος 60%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-85624			92,480%	
K - Means EM	-85700	-85956	30,00%	90,144%	0,048504
Random EM	-85700	-85938	15,00%	86,704%	0,063261

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-15619			100,000%	
K - Means EM	-15605	-16416	5,00%	98,903%	2,648513
Random EM	-15627	-16084	5,00%	98,419%	2,507233

Πίνακας 5.70 M=6 Θόρυβος 80%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-85624			92,480%	
K - Means EM	-88434	-88498	30,00%	90,144%	0,048504
Random EM	-88438	-88485	15,00%	86,704%	0,063261

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-15620			100,000%	
K - Means EM	-20586	-21237	5,00%	98,903%	2,648513
Random EM	-20720	-21236	5,00%	98,419%	2,507233

Για M=8

Πίνακας 5.71 M=8 Θόρυβος 40%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-107575			100,000%	
K - Means EM	-108292	-109083	0,00%	98,024%	0,037488
Random EM	-107644	-109366	0,00%	94,744%	0,044969

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	Ελάχιστη	Επιτυχίες	Μέση Τιμή	Απόκλιση
Incremental EM	-16150			100,000%	
K - Means EM	-16701	-17135	5,00%	99,419%	0,956929
Random EM	-16150	-17470	5,00%	98,710%	2,386592

Πίνακας 5.72 M=8 Θόρυβος 60%

Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-85625**			**92,480%**	
K - Means EM	**-108292**	**-109083**	**30,00%**	**90,144%**	**0,048504**
Random EM	**-107644**	**-109366**	**15,00%**	**86,704%**	**0,063261**

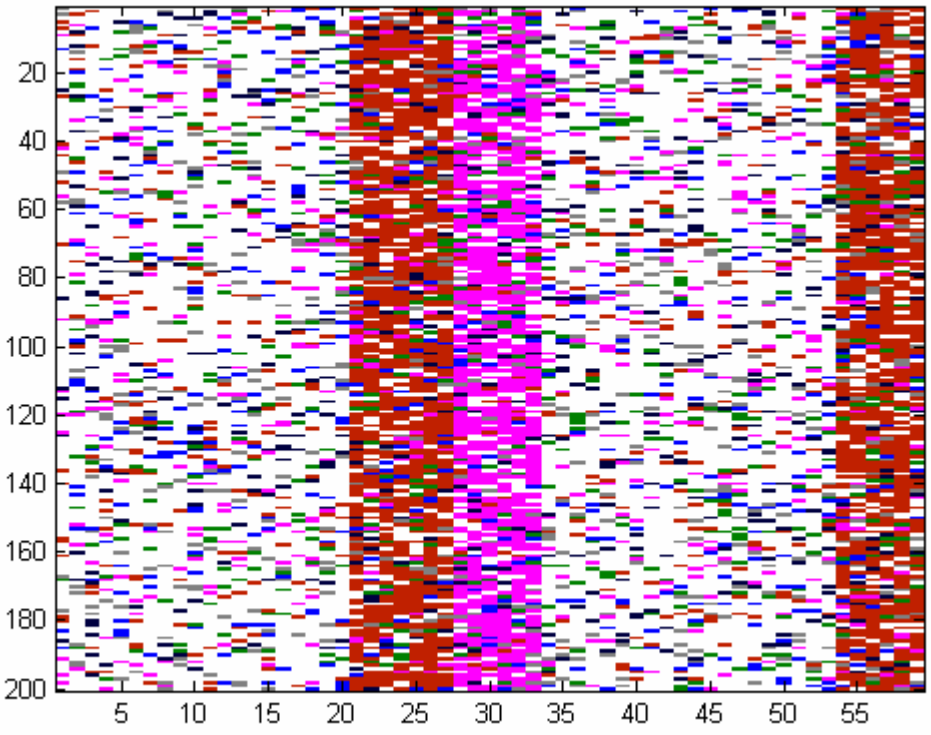
Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-15619**			**100,000%**	
K - Means EM	**-16701**	**-17135**	**5,00%**	**98,903%**	**2,648513**
Random EM	**-16150**	**-17470**	**5,00%**	**98,419%**	**2,507233**

Πίνακας 5.73 M=8 Θόρυβος 80%

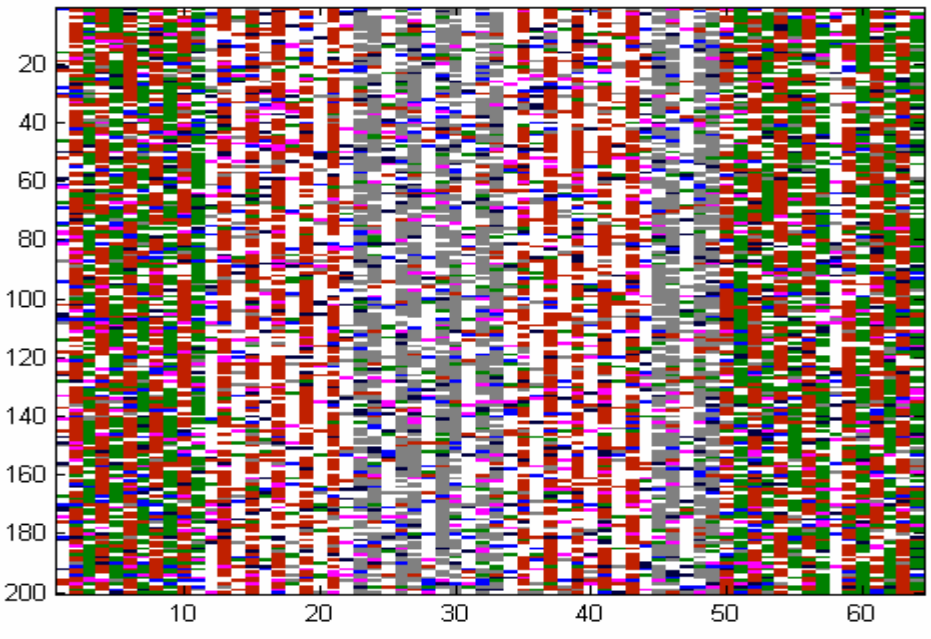
Σύνολο Εκπαίδευσης	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-88441**			**77,120%**	
K - Means EM	**-108292**	**-109083**	**5,00%**	**37,896%**	**0,066278**
Random EM	**-107644**	**-109366**	**5,00%**	**33,848%**	**0,064923**

Σύνολο Ελέγχου	Πιθανοφάνεια			Ποσοστό Κατηγοριοποίησης	
	Μέγιστη	**Ελάχιστη**	**Επιτυχίες**	**Μέση Τιμή**	**Απόκλιση**
Incremental EM	**-21306**			**61,935%**	
K - Means EM	**-16701**	**-17135**	**5,00%**	**89,387%**	**6,822851**
Random EM			**5,00%**	**82,839%**	**8,007335**

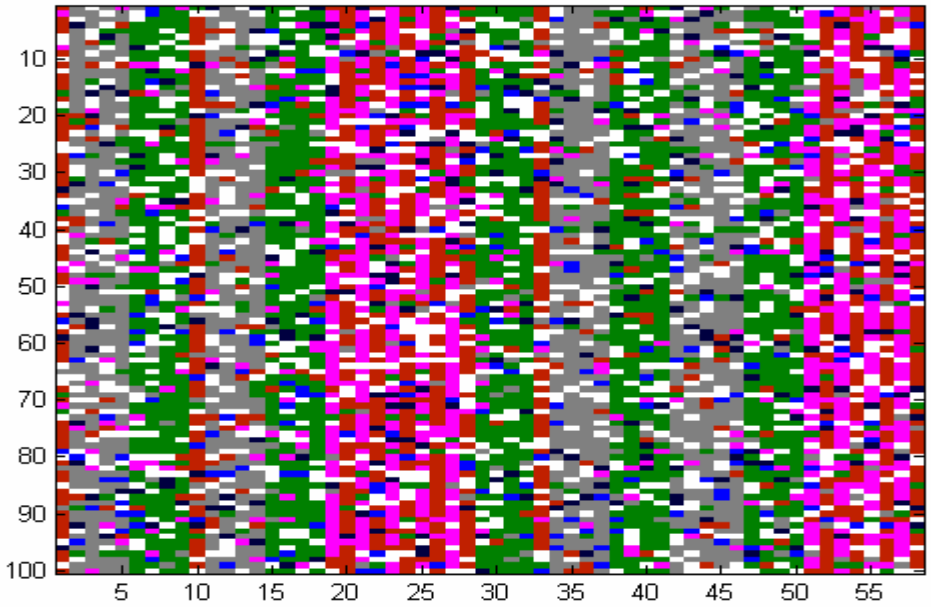
Οπτικοποίηση για M=6 και θόρυβο 40%



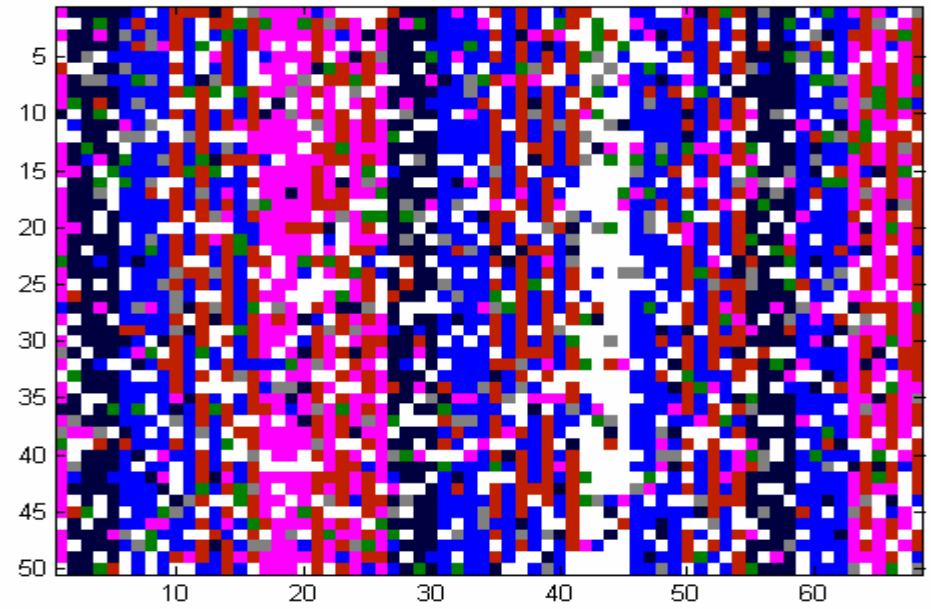
Σχήμα 5.11 Ομάδα 1



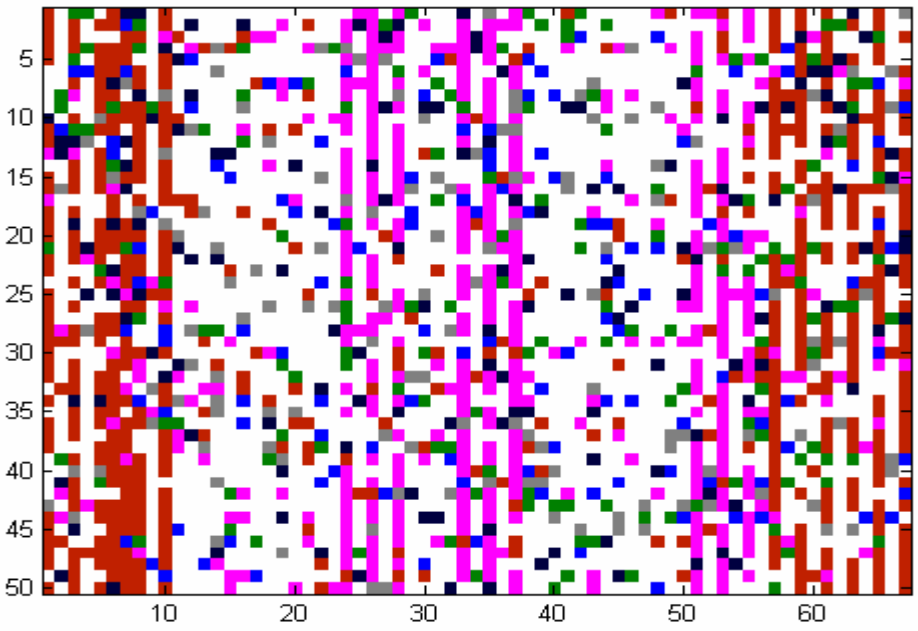
Σχήμα 5.12 Ομάδα 2



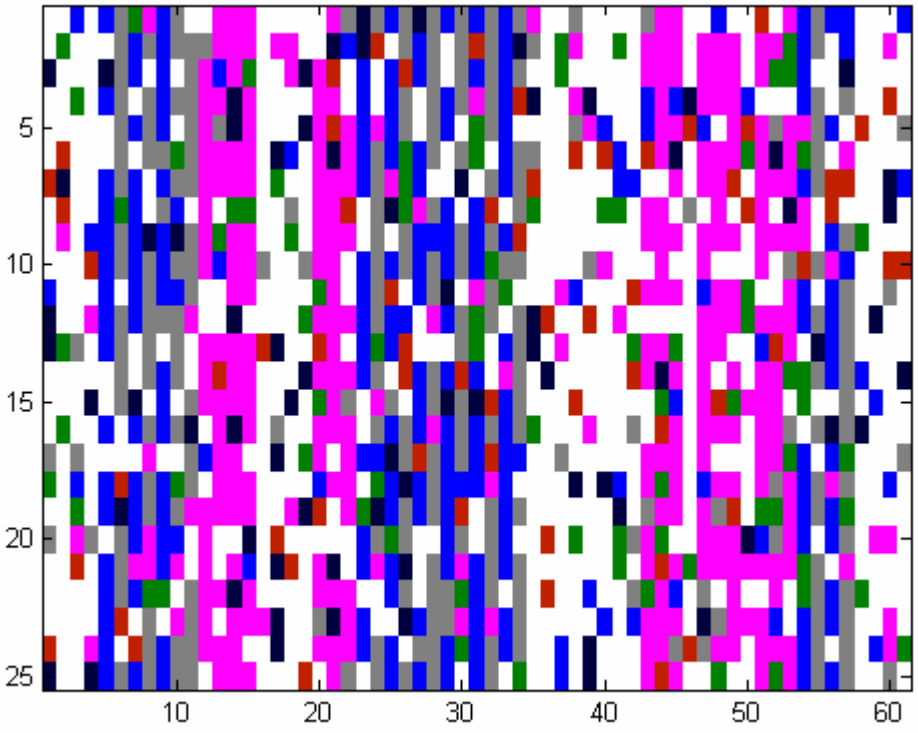
Σχήμα 5.13 Ομάδα 3



Σχήμα 5.14 Ομάδα 4



Σχήμα 5.15 Ομάδα 5



Σχήμα 5.16 Ομάδα 6

5.3.6. Πραγματικό σύνολο δεδομένων

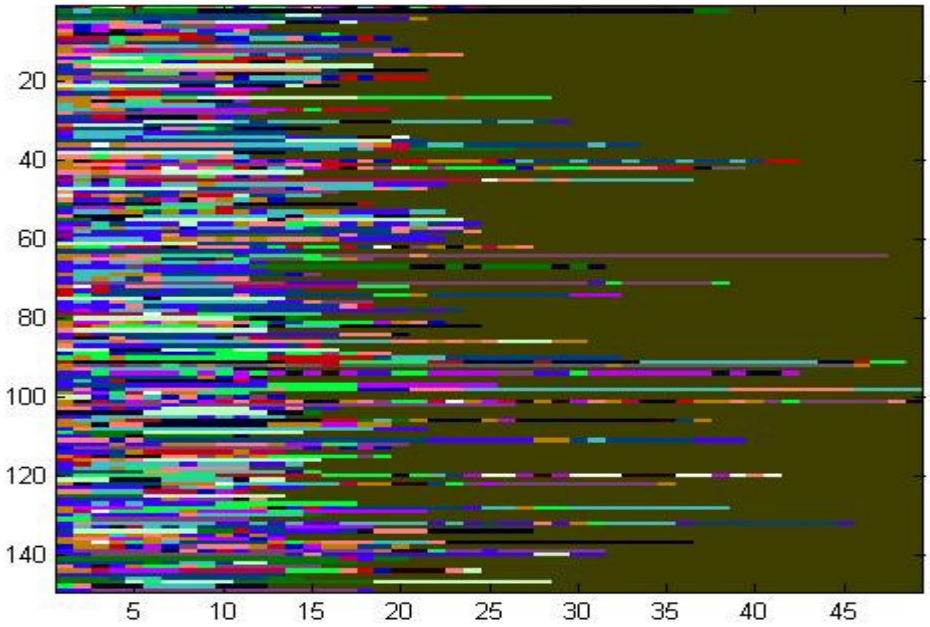
Αυτό το σύνολο προέκυψε από το πραγματικό σύνολο, το οποίο είχε φιλτραριστεί, ώστε να περιέχει ακολουθίες μήκους από 10 έως 50. Από αυτό το υποσύνολο προκύπτει με τυχαία δειγματοληψία 2000 ακολουθιών το σύνολο εκπαίδευσης που χρησιμοποιούμε. Σε αυτό μετρήσαμε την πιθανοφάνεια για το αυξητικό μοντέλο καθώς και μέση τιμή και τυπική απόκλιση για τυχαία μοντέλα και μοντέλα με αρχικοποίηση από τον K – means για 10 ομάδες. Τέλος, παρουσιάζουμε τις ακολουθίες για κάθε ομάδα που προέκυψαν από τον αυξητικό αλγόριθμο με bitmaps.

Πίνακας 5.74 Πραγματικό Σύνολο Δεδομένων

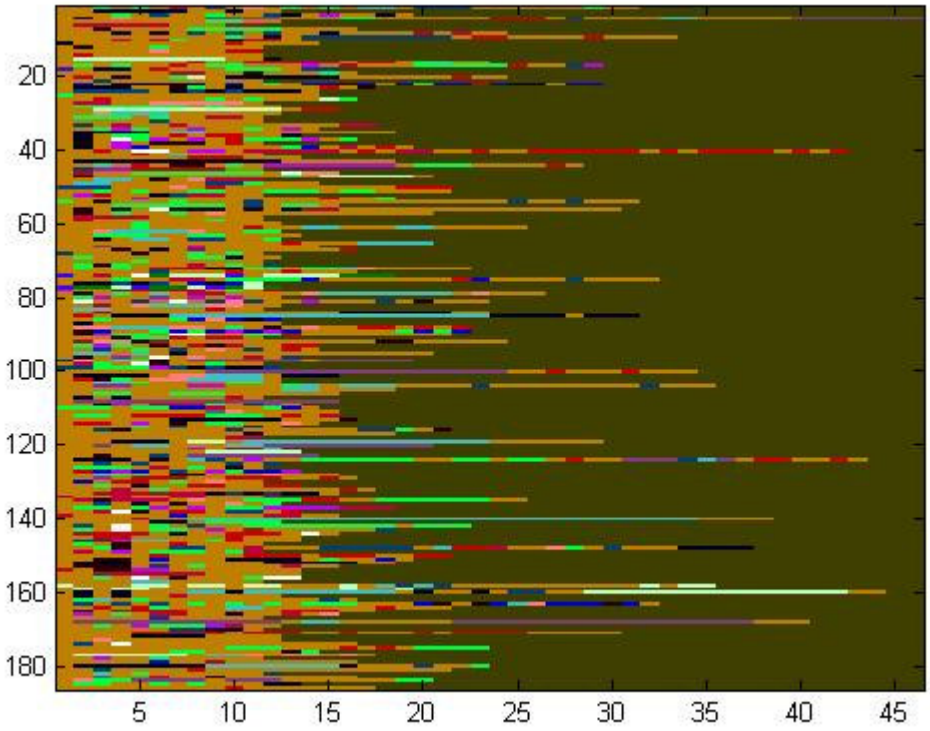
Σύνολο Εκπαίδευσης	**Πιθανοφάνεια**		
	Μέση Τιμή	**Απόκλιση**	**Επιτυχίες**
Incremental EM	**-21567,913**		
K - Means EM	**-21828,112**	**108,618**	**5,00%**
Random EM	**-21637,740**	**41,969**	**5,00%**

Παρατηρούμε ότι το μικτό μοντέλο που κατασκευάζεται αυξητικά δίνει σημαντικά καλύτερη λύση από τα άλλα μοντέλα.

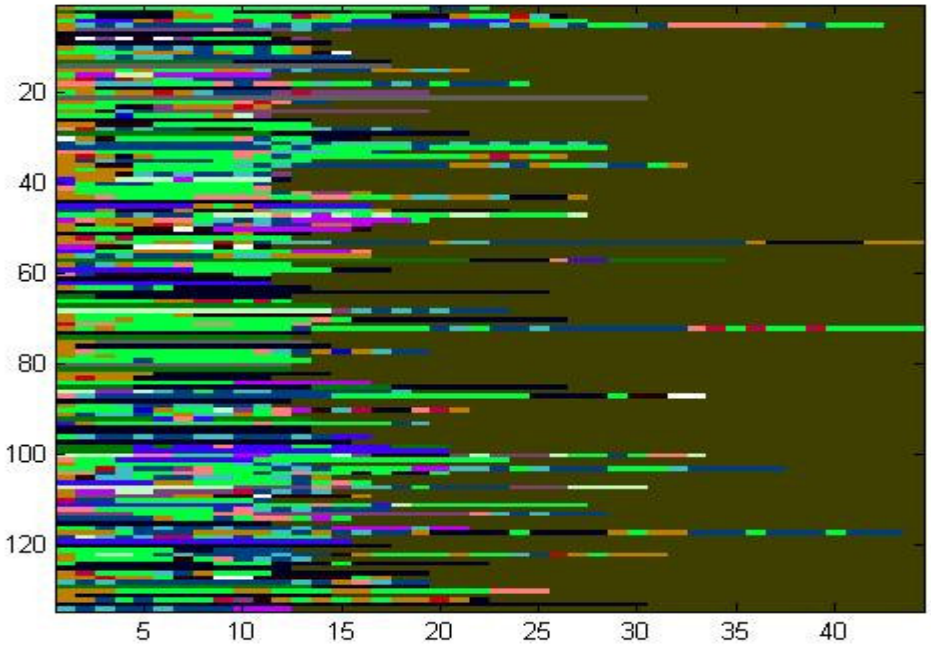
Οπτικοποίηση



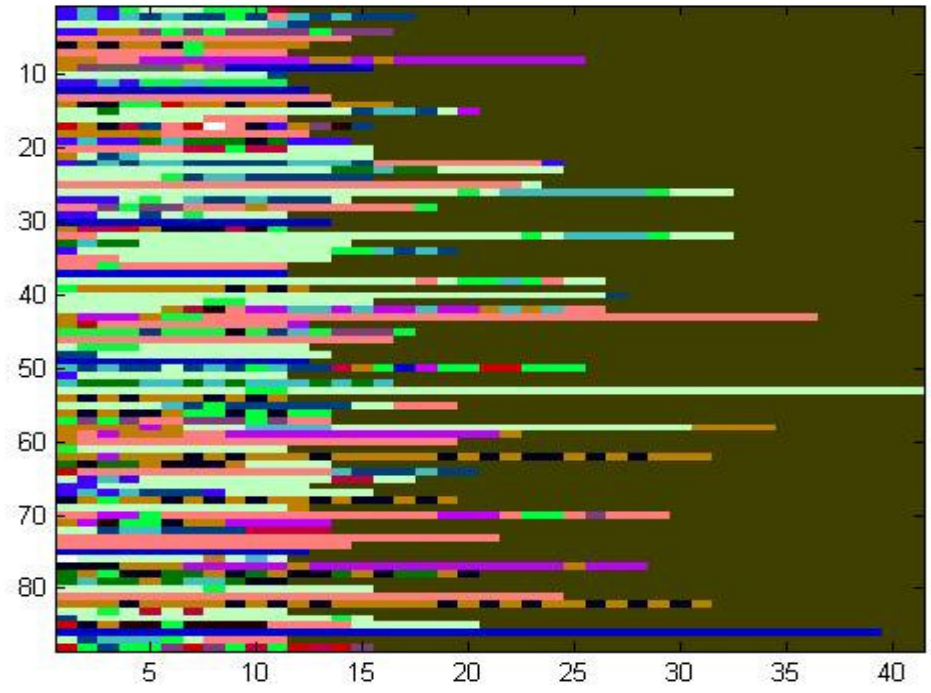
Σχήμα 5.17 Ομάδα 1



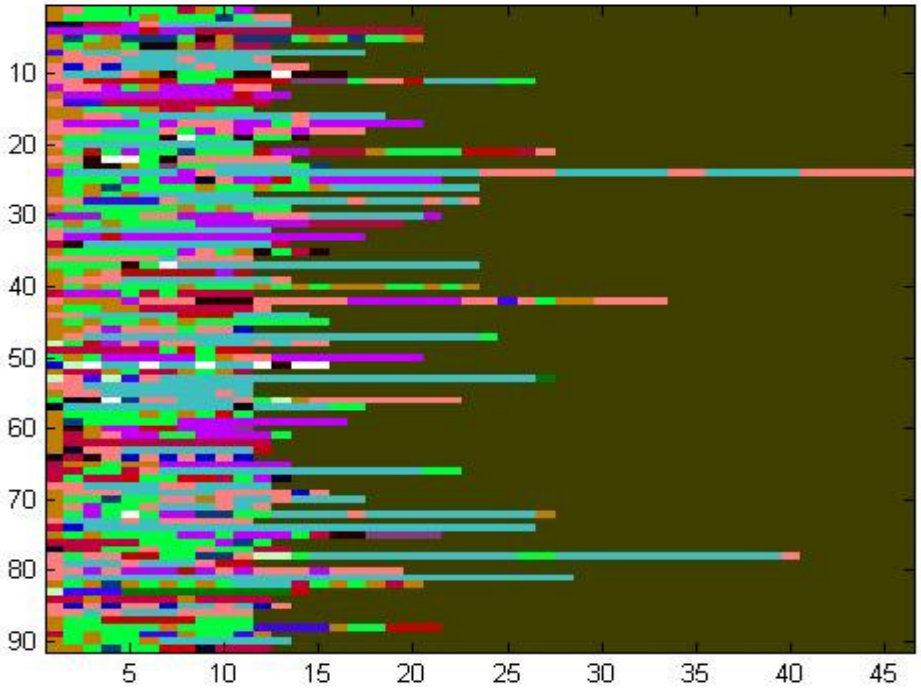
Σχήμα 5.18 Ομάδα 2



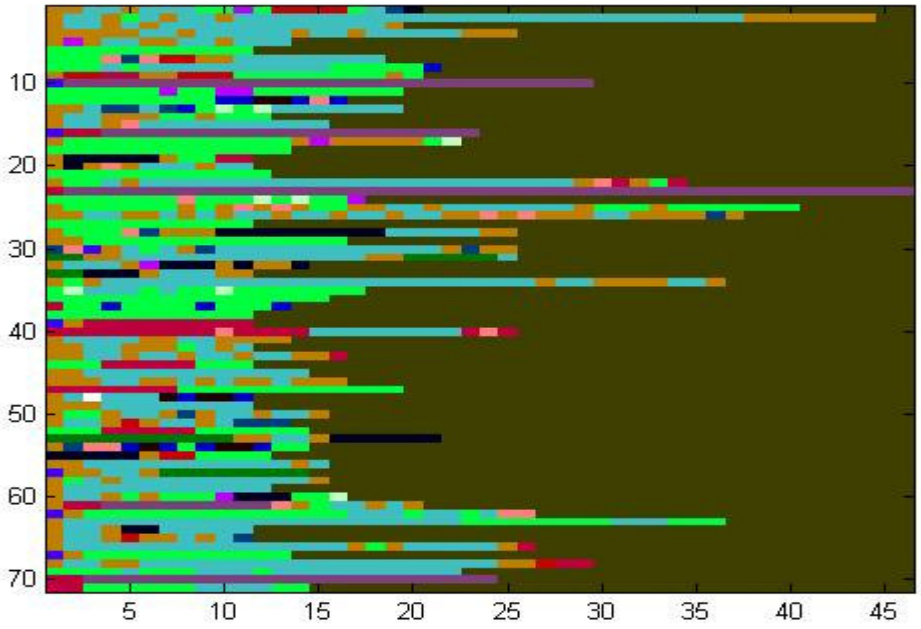
Σχήμα 5.19 Ομάδα 3



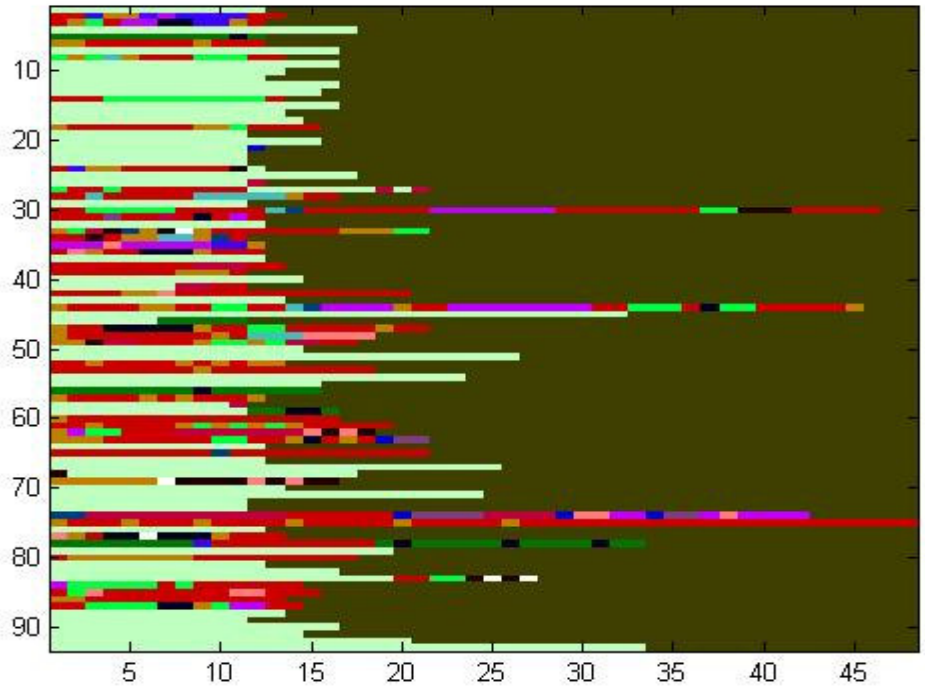
Σχήμα 5.20 Ομάδα 4



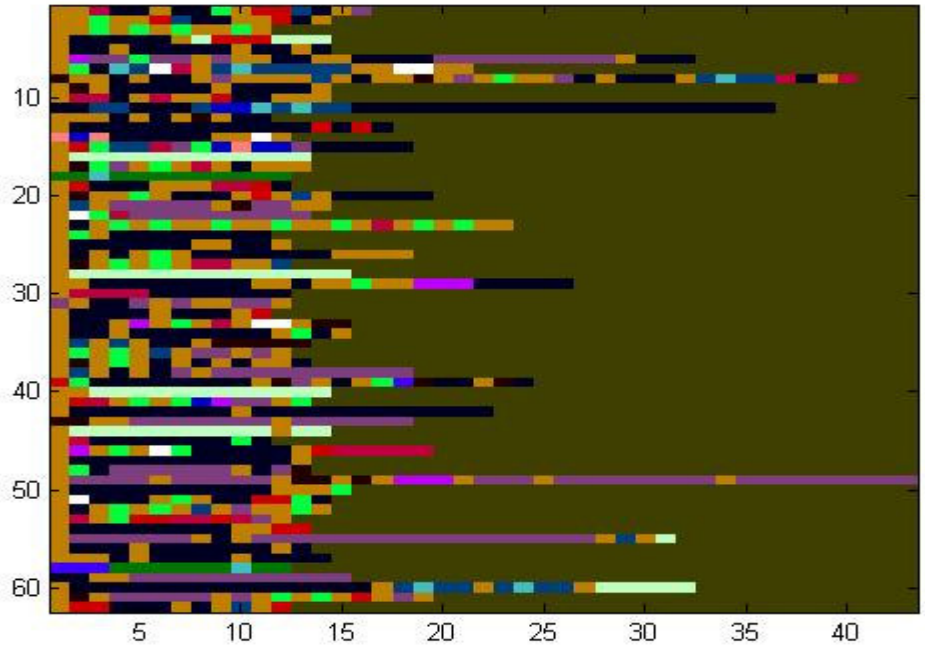
Σχήμα 5.21 Ομάδα 5



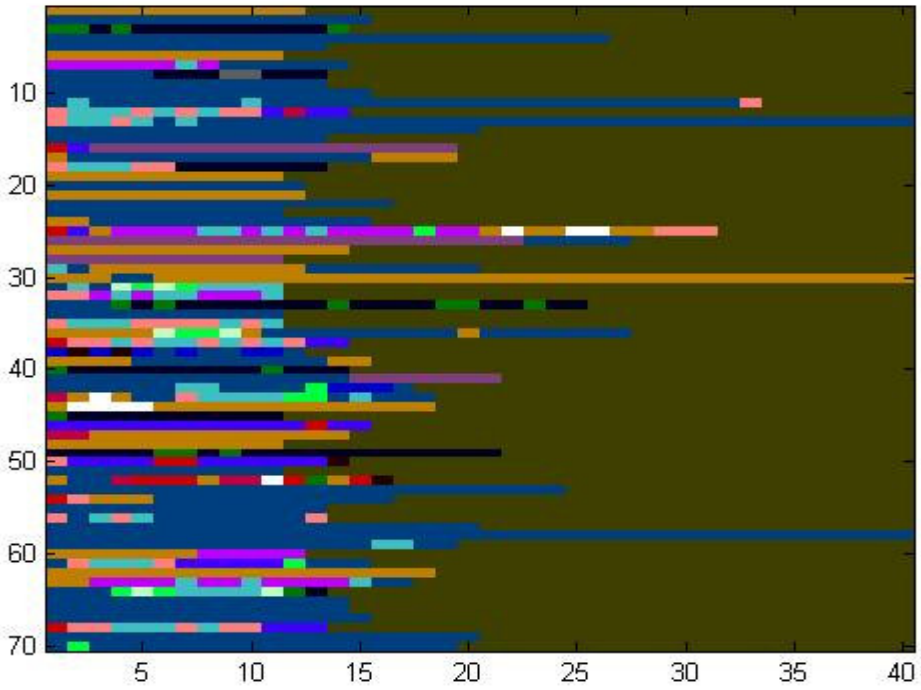
Σχήμα 5.22 Ομάδα 6



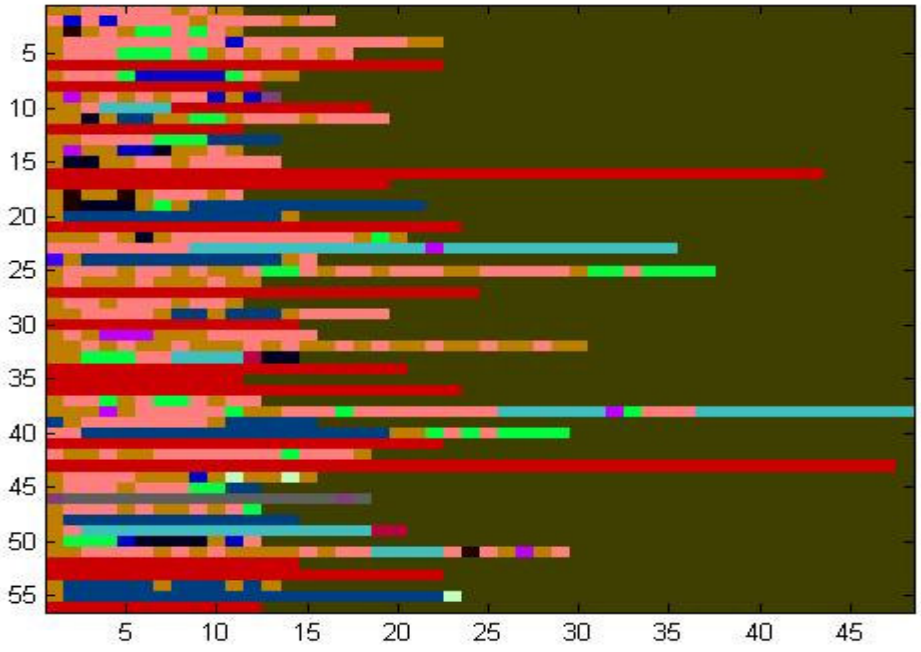
Σχήμα 5.23 Ομάδα 7



Σχήμα 5.24 Ομάδα 8



Σχήμα 5.25 Ομάδα 9



Σχήμα 5.26 Ομάδα 10

Τα συμπεράσματα που προκύπτουν από την ανάλυση των αποτελεσμάτων είναι ενθαρρυντικά για την αυξητική μέθοδο. Όσον αφορά στα τεχνητά δεδομένα, παρατηρούμε ότι το αυξητικό μοντέλο υπολείπεται σε επιδόσεις του τυχαίου και του K – means μικτού μοντέλου μόνο σε περιπτώσεις, όπου το ποσοστό θορύβου είναι πολύ υψηλό. Τα σύνολα δεδομένων που κατασκευάστηκαν περιείχαν σημαντικά ποσοστά θορύβου για να τονιστεί αυτό ακριβώς το χαρακτηριστικό. Για κανονικά ποσοστά θορύβου ο αυξητικός αλγόριθμος έβρισκε τις βέλτιστες λύσεις που τα άλλα μοντέλα έβρισκαν σε ορισμένες περιπτώσεις. Υψηλό ποσοστό θορύβου σημαίνει ότι τα κοινά πρότυπα που βρίσκονται στις ακολουθίες αλλοιώνονται σε τέτοιο βαθμό που είναι πολύ δύσκολο να τα διαχωρίσουμε. Αυτό έχει επίπτωση και στη λειτουργία του K – means, αφού είναι η ομαδοποίηση που προκύπτει δε μπορεί να είναι καλή. Συνεπώς, τα αρχικά μοντέλα με τα οποία προμηθεύουμε τον αυξητικό αλγόριθμο δεν είναι αρκετά καλά με αρνητική επίπτωση στην αυξητική διαδικασία.

Οι αρχικές παράμετροι του τυχαίου και του K – means μικτού μοντέλου παρουσιάζουν περισσότερη ομοιομορφία λόγω του τρόπου κατασκευής τους. Οι αρχικοποίηση δίνει περισσότερο ομοιόμορφες τιμές στις παραμέτρους, επομένως ταιριάζουν καλύτερα σε δεδομένα υψηλού θορύβου (μεγαλύτερης ομοιομορφίας), χωρίς αυτό να σημαίνει ότι επιτυγχάνουν καλή ομαδοποίηση. Απλά βρίσκουν καλύτερη λύση πιθανοφάνειας από το αυξητικό μοντέλο, του οποίου οι αρχικές παράμετροι προσπαθούν να επικεντρωθούν σε διαφορετικές ομάδες, παρουσιάζοντας συνεπώς, λιγότερη ομοιομορφία και μεγαλύτερη πόλωση στις τιμές των αρχικών παραμέτρων.

Στα σύνολα δεδομένων όπου τα πρότυπα των ακολουθιών είναι περισσότερο ευδιάκριτα, ο αυξητικός αλγόριθμος κατασκευάζει μικτά μοντέλα που ταιριάζουν στα δεδομένα, μεγιστοποιούν την πιθανοφάνεια και επιτυγχάνουν σωστή ομαδοποίηση. Αντίθετα, τα άλλα δυο μικτά μοντέλα (Random, K – means) υποφέρουν από το πρόβλημα αρχικοποίησης. Οι λύσεις που βρίσκουν δεν είναι συχνά καλές, ενώ παρουσιάζουν και μεγάλες διακυμάνσεις στην απόδοσή τους.

Συγχρόνως, όσο αυξάνει το πλήθος των ομάδων τόσο περισσότερο δύσκολο είναι το πρόβλημα της ομαδοποίησης. Μεγάλο πλήθος των ομάδων σε συνδυασμό με το

θόρυβο καθιστούν ιδιαίτερη δύσκολη την ομαδοποίηση. Το αυξητικό μοντέλο παρουσιάζει γενικά καλύτερη συμπεριφορά, εκτός από τις περιπτώσεις ιδιαίτερα υψηλού θορύβου και πολλών ομάδων, όπως είπαμε παραπάνω.

Στα πραγματικά δεδομένα που μελετήσαμε, το αυξητικό μοντέλο έδωσε σημαντικά καλύτερη λύση πιθανοφάνειας. Ωστόσο, από τη στιγμή που δε γνωρίζουμε την πραγματική ομάδα κάθε δεδομένου δεν μπορούμε να μετρήσουμε την ικανότητα κατηγοριοποίησης. Πάντως, από την οπτικοποίηση των αποτελεσμάτων φαίνεται να κατασκευάζονται ομάδες με συνάφεια μεταξύ των ακολουθιών τους.

ΚΕΦΑΛΑΙΟ 6. ΣΥΝΟΨΗ ΚΑΙ ΜΕΛΛΟΝΤΙΚΗ ΕΡΓΑΣΙΑ

Στη παρούσα εργασία μελετήσαμε τις ικανότητες κατηγοριοποίησης των μικτών μοντέλων Markov σε ακολουθιακά δεδομένα. Η έρευνα επικεντρώθηκε στα απλά μοντέλα Markov πρώτης τάξης. Προτείναμε μια αυξητική μέθοδο κατασκευής ενός μικτού μοντέλου που θα ταιριάζει βέλτιστα στο σύνολο των δεδομένων. Η σύγκριση με τα άλλα μικτά μοντέλα που αρχικοποιούνται τυχαία ή με τον K - means έδειξε ότι το αυξητικό μοντέλο ανταποκρίνεται στις προσδοκίες μας. Οι επιδόσεις του αποδεικνύονται ισάξιες ή καλύτερες από τα άλλα μικτά μοντέλα, ενώ διαθέτει το πλεονέκτημα ότι ο αριθμός των μοντέλων που χρειάζεται να προστεθούν μπορεί να βρεθεί κατά την εκπαίδευση, σε αντίθεση με τα άλλα μικτά μοντέλα, των οποίων ο αριθμός των συνιστωσών είναι προκαθορισμένος.

Περαιτέρω έρευνα πάνω στον αυξητικό αλγόριθμο θα μπορούσε να γίνει στον τρόπο κατασκευής των αρχικών μοντέλων. Ο K – means είναι απλός στην υλοποίηση και γρήγορος στην εφαρμογή του. Ωστόσο, η ομαδοποίηση που κάνει δεν είναι η καλύτερη δυνατή, ιδιαίτερα για υψηλά ποσοστά θορύβου στα δεδομένα. Μια διαφορετική προσέγγιση που θα δημιουργούσε πιο αμιγείς ομάδες θα επέτρεπε την κατασκευή καλύτερων αρχικών μοντέλων. Τροφοδοτώντας τον αυξητικό αλγόριθμο με καλύτερα αρχικά μοντέλα, αναμένουμε να λάβουμε καλύτερα τελικά αποτελέσματα για το αυξητικό μοντέλο.

Ενδιαφέρον παρουσιάζει το κρυφό μικτό μοντέλο. Τα κρυφά μοντέλα Markov διαθέτουν περισσότερες δυνατότητες απεικόνισης. Ωστόσο, παρουσιάζουν αυξημένη πολυπλοκότητα. Το πλήθος των καταστάσεων και οι τιμές των παραμέτρων τους δεν

είναι εύκολο να υπολογιστούν για το πρόβλημα των ακολουθιών. Τα πειράματα έδειξαν ότι η τυχαία αρχικοποίηση των παραμέτρων του κρυφού μοντέλου δεν δίνει ικανοποιητική λύση και το απλό μοντέλο παρουσιάζεται καταλληλότερο για το ίδιο πρόβλημα.

Η χρήση ενός αυξητικού αλγορίθμου για την κατασκευή ενός κατάλληλου μικτού μοντέλου από κρυφά μοντέλα Markov αξίζει να μελετηθεί. Το σημαντικότερο βήμα της αυξητικής μεθόδου είναι η κατασκευή των μοντέλων από τα οποία επιλέγουμε σε κάθε βήμα. Εδώ, για κάθε μοντέλο πρέπει να ορίσουμε το πλήθος των καταστάσεων, τις πιθανότητες αρχικών καταστάσεων, τις μεταξύ τους μεταβάσεις και τις κατανομές των συμβόλων κάθε κατάστασης.

ΑΝΑΦΟΡΕΣ

[1] Jeff Bilmes, "A gentle turorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden markov models", international computer science institute, TR-97-021, 1998

[2] Christopher Bishop, "Neural networks for pattern recognition", Oxford, 1995

[3] Konstantinos Blekas, D.I. Fotiadis, A. Likas, "Greedy mixture Learning for multiple motif discovering in biological sequences", Bioinformatics, 19(5), 607-617, 2003

[4] Igor Cadez, David Heckerman, Christopher Meek, Padhraic Smith, Steven White, "Model based clustering and visualization of navigation patterns on a web site", Technical Report MSR-TR-00-18, Microsoft Research, 2000

[5] Igor Cadez, Scott Gaffney, Padhraic Smyth, "A general probabilistic framework for clustering individuals", Knowledge Discovery and Data Mining, pages 140--149, 2000

[6] A.P. Dempster, N.M. Laird, D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", J.Roy. Statist. Soc. B, 39, 1-38, 1977

[7] Charles Elkan, "Clustering with k-means: faster, smarter, cheaper", Workshop on Clustering High Dimensional Data and its Application, 2004

[8] Yarip Ephraim, Neri Merhav, "Hidden Markov Processes", IEEE transactions on information theory, vol. 48, no. 6, 2002

[9] Warren Ewens, Gregory Grant, "Statistical methods in bioinformatics: An introduction", ISBN 0-387-95229-2, 2002

[10] E. Manavoglou, D. Pavlov, C.L. Giles, "Probabilistic user behaviour models", in IEEE International Conference on Data Mining (ICDM 03), 2003

[11] Lawrence Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition", IEEE, vol. 77, NO. 2 1989

[12] Nikos Vlassis, Aristidis Likas, “A Greedy EM Algorithm for Gaussian Mixture Learning**”,** Neural Processing Letters 15, 77-87, 2002

[13] Alexander Ypma, Tom Heskes, “Categorization of web pages and user clustering with mixtures of hidden markov models”, Proceedings of the International Workshop on Web Knowledge Discovery and Data Mining, WEBKDD'02, July 23 2002

[14] Καρακώστας Κ.Ξ., “Πιθανότητες κι εφαρμογές”, 1998

ΣΥΝΤΟΜΟ ΒΙΟΓΡΑΦΙΚΟ

Ανδρέας Κακολύρης

Γεννηθείς στον Πειραιά, στις 10 Ιουνίου 1981 με καταγωγή από τη Ζάκυνθο. Τελείωσε το σχολείο στη Ζάκυνθο και ακολούθως φοίτησε στη σχολή Πληροφορικής του Πανεπιστημίου Ιωαννίνων από το 1999 έως το 2003, οπότε και έλαβε το πτυχίο Πληροφορικής. Μέχρι σήμερα συνεχίζει τι σπουδές στη σχολή Πληροφορικής του Πανεπιστημίου Ιωαννίνων και εργάζεται για τη λήψη του μεταπτυχιακού τίτλου σπουδών ως μέλος της ομάδας επεξεργασίας και ανάλυσης πληροφορίας.